DCABES 2008 Proceedings

2008年国际电子商务、工程及科学领域的分布式计算和应用学术研讨会论文集

2008 International Symposium on Distributed Computing and Applications for Business Engineering and Science

July 27~31, 2008, Dalian, China

Volume I

◎主　编：须文波

Editor in Chief: Wenbo Xu

◎副主编：刘　丹

Associate Editor: Dan Liu

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

DCABES 2008 Proceedings

2008 年国际电子商务、工程及科学领域的分布式计算和应用学术研讨会论文集

2008 International Symposium on Distributed Computing and Applications for Business Engineering and Science

Volume I

July 27~31, 2008, Dalian, China

主　编：须文波
Editor in Chief: Wenbo Xu

副主编：刘　丹
Associate Editor: Dan Liu

電子工業出版社

Publishing House of Electronics Industry

北京·BEIJING

内容简要

随着计算机技术的不断发展，分布式并行以及高性能计算对科学、工程技术、经济管理等领域的重要性日益突出。一年一度的 DCABES 国际会议已经成为该领域有影响的学术会议。2008 年 DCABES 国际会议论文集共收录近 300 篇学术论文，内容涉及：分布式并行计算、网格计算、数值计算、网络技术与信息安全、信息处理、信息管理系统、电子商务、图像处理、Web 技术、无线传感技术、智能计算等。对相关研究领域中的本科高中级学生、研究生、教学及科研人员均有较大的帮助。

图书在版编目（CIP）数据

2008 年国际电子商务、工程及科学领域的分布式计算和应用学术研讨会论文集. 上册：英文 / 须文波主编.
—北京：电子工业出版社，2008. 7
ISBN 978-7-121-07018-1

I. 2… II. 须… 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（2008）第 096214 号

责任编辑：秦绪军　　潘　娅
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：93.5　字数：3000 千字
印　　次：2008 年 7 月第 1 次印刷
定　　价：398. 00 元（上、下册）

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，联系电话：（010）68279077；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。

2008年国际电子商务、工程及科学领域的分布式计算和应用学术研讨会论文集

2008 International Symposium on Distributed Computing and Applications for Business Engineering and Science

Volume I

责任编辑：秦绪军
潘　娅
封面设计：喻　晓

ISBN 978-7-121-07018-1

9 787121 070181 >

定价：398.00 元

（上、下册）

PREFACE

The series of meetings, International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES), is now becoming an important international event on various applications and the related computing environments of distributed and grid computing. The first meeting was held at Wuhan University of Technology, Wuhan, and the second meeting was held at Southern Yangtze University, Wuxi, the third meeting was held at Wuhan University of Technology, Wuhan, the fourth meeting was held at Greenwich University, Greenwich, the fifth was organized by Southern Yangtze University and Zhejiang GongShang University and held at Hangzhou, and the sixth was organized by Wuhan University of Technology and held at Yichang. The seventh meeting will be organized by Jiangnan University and held at Dalian. The conference themes include not only its traditional theme such as parallel and distributed computing, but also intelligent computing and other topics that will be described as follows.

It was my pleasure that the DCABES2008 conference had received a great number of papers submitted cover a wide range of topics, such as Parallel/Distributed Computing, Grid Infrastructure and Applications, Image Processing, Network Technology and Information Security, E-Commerce and E-Business, Intelligent Computing, Information Processing, Information Management System, and so forth.

Papers submitting to the conference come from over 15 countries and regions. All papers contained in this Proceeding are peer-reviewed and carefully chosen by members of Scientific Committee, proceeding editorial board and external reviewers. Papers accepted or rejected are based on majority opinions of the referee’s. All papers contained in this proceeding give us a glimpse of what future technology and applications are being researched in the distributed computing area in the world.

I would like to thank all members of the Scientific Committee, the local organizer committee, the proceedings editorial board and external reviewers for selecting the papers. Special thanks are due to Dr. Choi-Hong LAI, Prof. Qingping Guo and Prof. Dan Liu, who sponsored and organized the mini-symposium on Imaging Processing at Shenyang. It is indeed a pleasure to work with them and obtain their suggestions. I am also grateful to Professor Franck Cappello, Professor Kako Takashi, and Professor Padmanabhan Seshaiyer for their contributions of keynote speeches in the conference.

Sincerely thanks should be forwarded to the China Ministry of Science and Technology (MOST), the China Ministry of Education (MOE), National Nature Science Foundation of China (NSFC), Jiangnan University and China Criminal Police University.

Finally I should also thank Dr. Wei Fang, Dr. Jun Sun, Miss Na Tian, Miss Wenjuan Ji for their efforts in conference organizing activities, my postgraduate students, such as Miss Jing Zhao, Miss Hui Li, Miss Yan Kang, Mr Dong Wang Mr. Wei, Chen, Mr. Runian Geng, and Mr. Zhiguo Chen, Miss Ji Zhao, Miss Di Zhou, for their time and help. Without their time and efforts this conference cannot be organized smoothly.

Enjoy your stay in Dalian. Hope to meet you again at the DCABES 2008.

Professor Wenbo Xu,
Chair of the DCABES2008
School of Information Technology
Jiangnan University
Jiangsu, China

COMMITTEES

Steering Committee

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

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

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

Xu, Prof. W. B., Jiangnan University, Wuxi, China

Scientific Committee

Chair : Xu, Prof. W. B., Jiangnan University, China

Co-Chair : Lai, Prof. C.-H., University of Greenwich, U.K.

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

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

Cai, Prof. Jiamei, Zhejiang Industry University, Hangzhou, China

Cao, Prof. J.W., Research and Development Centre for Parallel Algorithms and Software, Beijing, China

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

Feng, Prof. Bin, Jiangnan University, Wuxi, China

Geiser, Dr. J.H. University at Berlin, Germany

He, Dr. H.W. Hohai University, China

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

Jesshope, Prof. C., University of Amsterdam, the Netherlands

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

Keyes, Prof. D.E., Columbia University, USA

Khaddaj, Dr. S. Kingston University, UK.

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

Liddell, Prof. 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

Liu, Prof. Yuan, Jiangnan University, Wuxi, China

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

Ng, Prof. Michael, Baptist?University, Hong Kong, China

Ng, Mr Frank C.k, Chinese University of Hong Kong, Hong Kong, China

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

Sun, Prof. J., Academia Sinica, Beijing, China

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

Wang, Prof. Meiqing Fuzhou University, China

Wang, Prof. Shitong, Jiangnan University, Wuxi, China

Wang, Prof. WeiMing, Zhejiang Gongshang University, Hangzhou, China

Wang, Prof. Zhijian, Hohai University, Nanjing,China

Wang, Prof. G.M., Zhejiang Gongshang University, Hangzhou, China

Zhuang Prof. Yueting, Zhejiang University,Hangzhou, China

Zhang, Prof. J., University of Kentucky, USA

Zhang, Prof. Wenyuan, Jiangnan Computer Research Institute

Zou, Prof. Jun, Chinese University of Hong Kong, Hong Kong, China

Local Organizer Committee

Xu, Prof. W.B, Jiangnan University, Wuxi, China

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

Liu, Prof. Dan, China Criminal Police University, China

Chen, Prof. Jian, Jiangnan University, Wuxi, China

Wang, Prof. Shitong, Jiangnan University, Wuxi, China

Liu, Prof. Yuan, Jiangnan University, Wuxi, China

Wu, Prof. Xiaojun, Jiangnan University, Wuxi, China

Liu, Prof. Lixia, Jiangnan University, Wuxi, China

Liu, Dr. Li, Jiangnan University, Wuxi, China

Sun, Dr. Jun, Jiangnan University, Wuxi, China

Chai, Dr. Zhilei, Jiangnan University, Wuxi, China

CONTENTS

Grid'5000: A Large Scale Reconfigurable and Controllable Platform for Research in Computer Science

Franck Cappello

INRIA
www.grid5000.fr
Email: fci@lri.fr

Abstract

The Computer Science discipline, especially in large scale distributed systems like Grids and P2P systems, tends to address issues related to increasingly complex systems, gathering thousands to millions of non trivial components. Theoretical analysis, simulation and even emulation are reaching their limits. In this paper, we describe the motivations, design and results of Grid'5000. Grid'5000 is a large scale systems designed as scientific instrument for researchers in the domains of Grid, P2P and networking. Computer scientists use this platform to address issues in the different software layers between the hardware and the users: networking protocols, OS, middleware, parallel and distributed application runtimes, and applications.

Keywords: Computer Science Grid, large scale experiments, reconfiguration, controlable experimental conditions

1 Introduction

Grid and P2P systems are very popular as production platforms (EGEE, TeraGrid, SETI@home, Edonkey, Skype) and inspire a wide spectrum of research. These distributed systems are still difficult to design, operate and optimize due to their software complexity, heterogeneity, the volatility of their components and their large scale. As a consequence, many institutes and international programs develop significant funding efforts to foster Grid and P2P research initiatives.

As a matter of fact, the research in Grid and P2P systems span over all the layers of the software stack between the user and the hardware. Applications, programming environments, runtime systems, middleware, OS and networking layers are subject to extensive studies seeking to improve their performance, security, robustness and quality of service.

Like other scientific domains, research in Grid and P2P computing is based on a variety of methodologies and tools. When Grid'5000 was designed, most of the research conducted in Grids and P2P systems was performed using simulators[9][2][3], emulators[7] or production platforms. However, all these tools present limitations making the study of new algorithms and optimizations difficult. Simulators focus on a specific behavior or mechanism of the distributed system and abstract the rest of the system. A main restriction of simulators is the difficulty of their validation. Indeed very few studies have been conducted to validate the existing simulators. When it becomes difficult to capture and extract the factors influencing the distributed systems, emulators can help by executing the actual software of the distributed system, in its whole complexity, on a fully controlled platform. As a consequence, there is still a gap between emulators and the reality: they cannot capture all the dynamic, variety and complexity of real life conditions.

Production platforms may be considered as good candidate for experimentation because they expose the software to experiment to realistic conditions. However, there are several reasons why computer scientists require their own infrastructure and cannot use existing production machines. Foremost, many computer science projects require experiments with the operating system and communication protocols, that are hard to do on production machines. Secondly reproducing the experimental conditions several times is almost impossible on production platforms. Thirdly, there is a clear difference between how computer scientists and application scientists use the resources. Most application scientists want to run large experiments that take much compute time for high job throughput and many applications consist of a large number of sequential jobs. Computer scientists, on the other hand, want to run large-scale distributed experiments that use many sites at the same time, in a more interactive way.

Thus, the complexity of Grid and P2P systems raise the need for real-scale experimental platforms where computer scientists can run experiments, observe the distributed systems at large scale, stress the systems using experimental conditions injectors and make precise measurements.

2 Grid'5000 Design

Grid'5000 is the result of 1) the past experiences on testbeds for Grid research such as eToile in France, and 2) the description by the computer scientists of their needs in experimentation.

These two elements led to propose a large scale experimental platform, with deep reconfiguration capabilities and a strong control and monitoring infrastructure.

During the preparation of the project in 2003, we conducted an analysis on the need of a computer science Grid and the diversity of potential experiments. The researchers of the Grid computing community in France, involved in many French ACI Grid projects and European Grid projects, proposed a set of about 100 experiments. A first conclusion of the analysis was the need for a large scale (several thousands of CPUs), distributed (10 sites) system. A second conclusion was that the experiment diversity nearly covers all layers of the software stack used in Grid computing, from the user interface to the networking protocols. A third conclusion was that most of the researchers need a specific experiment setting, different from the other researchers. Researchers involved in networking protocols, OS and Grid middleware research often require a specific OS for their experiments. Some research on virtual machines, process checkpointing and migration need the installation of specific OS versions or OS patches that may not be compatible. Researchers needs are quite diverse in Grid Middleware: some require Globus, while others need Unicore, Desktop Grid or P2P middleware. Some other researchers need to test applications and mechanisms in a multi-site, multi-cluster environment, without any Grid middleware.

As a consequence, we concluded that Grid'5000 should provide a deep reconfiguration mechanism allowing researchers to deploy, install, boot and run their specific software images, possibly including all the layers of the software stack. This reconfiguration capability led to the experiment workflow followed by Grid'5000 users: 1) reserve a partition of Grid'5000, deploy a software image on the reserved nodes, reboot all the machines of the partition using the software image, run the experiment, collect results and relieve the machines.

Because researchers are able to boot and run their specific software stack on Grid'5000 sites and machines, we decided 1) to isolate Grid'5000 from the rest of the Internet and 2) to let packets fly inside Grid'5000 without limitation. The first choice ensures that Grid'5000 will resist to hacker attacks and will not be used for Internet attacks. The second choice guarantees that communication performance does not suffer from the overhead of an imposed security system. Thus, Grid'5000 is built as a large scale confined cluster of clusters. Strong authentication and authorization checks are done when users log in Grid'5000.

Grid'5000 is composed of heterogeneous resources. However, we decided to keep at least 2/3 of the machine homogeneous in Grid'5000 for two main reasons: 1) speedup evaluation is difficult to evaluate with heterogeneous hardware, 2) hardware diversity increases the complexity of the deployment, reboot and control subsystems and the every day management and maintenance cost.

The capability to reproduce experimental conditions is fundamental in experimental tools, especially when performance comparisons are conducted. To fulfill this strong requirement, we decided to use dedicated network links between sites, to allow users reserving the same set of resources across successive experiments, to allow users running their experiments in dedicated nodes (obtained by reservation) and to let users install and run their proper experimental condition injectors and measurements software. Thus every user has full control of the reserved experimental resources.

3 Grid'5000 Organization and Status

Based on the design decisions presented in the previous section, we decided to build a platform of 5000 CPU-cores distributed over 9 sites in France. Figure 1 presents an overview of Grid'5000. Every site hosts a cluster and most of the sites are connected between each others by high speed network links (RENATER 4: 10 Gbps Dark fibre links).

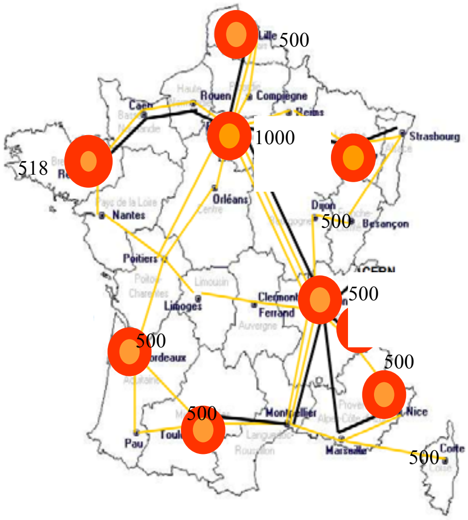


Figure 1 Overview of Grid'5000.

Numbers in Figure 1 give the target number of CPUs for every cluster. 2/3 of the nodes are dual CPU 1U racks featuring 2 AMD Opteron running at 2 Ghz, 2 GB of memory and two 1Gbps Ethernet adapters. Clusters are also equipped with high speed networks (Myrinet, Infiniband, etc.). Disk space varies across the 9 sites. Most of the CPUs racks provide at least 80 Gbytes of local non archived storage. In addition, all Grid'5000 sites provide more than 1 Tbybtes of storage (replicated or archived) for the user experimental data.

Every user has a single account on Grid'5000. Every Grid'5000 site manages its own user accounts and runs an LDAP server containing the same tree: under a common root, a branch is defined for each site. On a given site, the local administrator has read-write access to the branch and can manage its user accounts. The other branches are periodically synchronized from remote servers and are read-only. Once the account is created, the user can access any of the Grid'5000 sites or services (monitoring tools, wiki, deployment, etc.). User data are kept local to every site and distribution to remote sites is done by the user through classical file transfer tools (rsync, scp, sftp, etc.). Data transfers from and to the outside of Grid'5000 are restricted to secure tools and done on gateway servers.

At cluster level, users submit their resource reservations and experiment jobs using the OAR [6] reservation engine and batch scheduler. OAR provides most of the important features implemented by other batch schedulers such as priority scheduling by queues, advance reservations, backfilling and resource match making. OAR relies on a specialized parallel launching tool named Taktuk[1] to manage all large-scale operations like parallel tasks launching, nodes probing or monitoring. At the Grid level (Cluster of Clusters) a simple broker collocates the resources of several Grid'5000 sites by submitting reservations to the local OAR schedulers. Currently, if one reservation is refused, all previously accepted reservations are canceled. This simple meta-reservation approach is acceptable when the platform workload is moderate. Clearly it should be replaced by a more sophisticated approach when normal workload leads to many meta reservation cancellations.

To reconfigure the software stack on every reserved node, the users run the Kadeploy2[5] deploying the user defined software environment on a disk partition of selected nodes. The software environment contains all software layers from the OS to application, in addition to experimental condition injectors and measurements tools. Deployment begins by rebooting all nodes on a minimal system through a network booting sequence. This system prepares the target disk for deployment (disk partitioning, partition formatting and mounting). Then the environment is broadcast to the selected nodes using a pipelined transfer with on the fly image decompression. At this point, some adjustments must be done on the broadcasted environment in order to be compliant with node and site policies (mounting tables, keys for authentication, information for specific services that cannot support auto-configuration). The last deployment step consists in rebooting the nodes on the deployed system from a network loaded bootloader.

4 Grid'5000 Key Results Examples

The main objective of Grid'5000 and its associated software set is to ease the deployment, execution and result collection of large scale Grid experiments. Currently about 400 experiments are planned or realized. The topics of these experiments cover all the layers of the software stack between the user and the Grid resources.

More than 50 experiments are planned at the networking layer. This includes research on high speed protocols, monitoring, distributed measurements, high performance protocols for MPI on the Grid, high bandwidth data transfer analysis and modeling, traffic isolation, stress of 10G WAN links, realistic Internet traffic replay, Grid collective communications, transfer time prediction, etc.

More than 100 experiments are proposed for the middleware layer. This set of experiments includes tests on Globus, OGSA-DAI, fault tolerant MPI, distributed storage systems, automatic Grid infrastructure deployment, rapid and dynamic virtual cluster creation, Desktop Grid environments, data management and scheduling in Desktop Grids, P2P DHT, meta and hierarchical Grid schedulers, fully distributed batch schedulers, automatic Grid execution checkpointing, JXTA performance and scalability, resource discovery systems, etc.

More than 50 experiments are planned for the programming layer of the software stack. For this layer, the research concerns the design, implementation, tests and evaluation of Grid programming environments, such as Workflow description and runtime tools (YML, OpenWP, etc.), Grid versions of MPI implementations (MPICH and OpenMPI), Grid RPC environment such as DIET and OmniRPC, combinatorial optimization environment such as PARADISEO-G, Object oriented parallel and distributed computing in Java with ProActive, Component model environments, disruptive programming approach like Chemical computing, etc.

The application layer receives a strong interest with more than 100 experiments (done or planned). The main purpose of this research is to evaluate the performance of applications ported on the Grid and test alternatives or design new algorithms and new methods for these applications. The application domains cover life sciences (mammograms comparison, protein sequencing from tandem mass spectrometry, gene prediction, virtual screening funnel, conformation sampling and docking, etc), physics (seismic imaging, parallel solvers for two phase flows, hydrogeology, simulation of self-propelled solids in a viscous incompressible fluid, Particle Image Velocimetry, seismic tomography, geophysical inverse problem, climate modeling, 3D discrete ordinates neutron transport: SWEEP3D, fluid mechanics, external aerodynamics, radiative transfer coupled to hydrodynamics, etc.), applied mathematics (sparse matrix computation, combinatorial optimization solvers, parallel and distributed model checkers, PDE problem solving with asynchronous iterations, etc.), chemistry (molecular simulation, estimation of thickness and optical constants of thin films, etc.), industrial processes, financial computing, etc.

Other experiments concern operating systems (XtreemOS), virtualization techniques and software tools to be used as Grid'5000 mechanisms: heterogeneity emulators, experimental condition injectors, monitoring tools, fast software deployment and reconfiguration tools, etc.

To highlight the interests and benefits of using Grid'5000, we present two experiment examples. The first example concerned the programming environment layer and combinatorial optimization algorithms. Optimally solving large instances of combinatorial optimization problems using a parallel Branch and Bound (B&B) algorithm requires a huge number of computational resources. In [8], the authors proposed a gridification of the parallel B&B algorithm, based on new ways to efficiently deal with some crucial issues, mainly dynamic adaptive load balancing, fault tolerance, global information sharing and termination detection of the algorithm. A new efficient coding of the work units (search sub-trees) distributed during the exploration of the search tree is proposed to optimize the involved communications. The algorithm has been implemented following a large scale idle time stealing paradigm (Farmer-Worker) and experimented on the Flow-Shop NP-hard scheduling problem instance (Ta056) (scheduling of 50 jobs on 20 machines). The algorithm allowed to improve the best known solution by providing the optimal solution with proof of optimality. The problem was solved within 25 days using about 1900 processors belonging to 6 clusters of the Grid5000 and to 3 clusters from Université de Lille1. During the resolution, the worker processors were exploited with an average of 97% while the farmer processor was exploited only 1.7% of the time. These two rates are good indicators of the efficiency of the proposed approach and its scalability. This result can be considered as a success story since the problem instance has never been solved exactly before.

The second example concerns the application level. Since the last 30 years, many research works in geophysics (seismology) focused on seismic tomography to reveal the structure of Earth interior. To solve the resolution limitation of tomographic models, an irregular model is used, which adapts locally to the density of seismic information. This method computes the seismic tomography from the huge amount of data, based on the seismicity of the world from the years 1964 to 1995 (approximately 82000 seisms and 12 millions of rays, about 1.2 millions significant rays after pre-processing). To speed-up the computation, several specific parallel MPI programs have been developed. The experiment [4] concerned the first step of a method, which consists in ray-tracing the seismic rays (the waves' paths) from the recorded seismic events. This step is highly parallel since every ray can be traced independently. However the method eventually requires an all-to-all communication phase, which is a real bottleneck on many hardware platforms. In July 2006, Grid'5000 was used for a tomography using the full dataset. Several configurations were tested to assess the application scalability, with 32, 64, 128, 192 and up to 458 processors, on 1, 2, 3 or 5 sites. Despite a considerable volume of data exchanged in the all-to-all phase (7 GB, 15 GB and 20 GB for 32, 128 and 458 processors resp.) the speedup stays nearly linear.

Moreover the application performance does not significantly decrease when using 3 sites instead of 2. The global method takes 227s on 458 processors, 3164s on 32 processors (a single cluster) and more than 36 hours on a single PC. This experiment demonstrates: 1) this class of applications scales extremely well on Grid'5000, 2) MPI applications can run efficiently on a Grid, 3) a platform like Grid'5000 is a very useful tool to evaluate the scalability and performance of parallel applications on the Grid.

5 Impact As A Research Instrument

Designing, constructing and running a Computer Science Grid raises many technical issues and has a significant cost. Grid'5000 should be evaluated as research tools, the quality and quantity of the scientific results they have produced and their impact on the research community.

One of the most significant signs of success of Grid'5000 is its number of users. We currently have about 400 active users who present their experiment context and report on the Grid'5000 web site. We frequently receive requests from foreign colleagues to get an account and use Grid'5000, despite the fact that Grid'5000 access is rather restricted, because all funding is supported by France. Foreign colleagues can get an access to Grid'5000 through collaborations with a French research team. This is the case for the participants of several European projects of the European Frame Work program 6 (Grid4all, QosCos, XtreemOS, etc.). In total, the users are from 70 computer science laboratories worldwide.

Beyond the attractiveness, the main result is the number of publications: in about 2 years of exploitation, Grid'5000 has been used for 4 HDR (a diploma in France that could be obtained 4 or 5 years after the Ph. D.), 15 Ph. D., tens of Master theses and hundreds of publications. We continuously observe an increase of the number of master students, Ph. D. candidates and researchers using it.

A nice measure of the Grid'5000 usefulness is the activity level: in normal situations the workload is around 50\% of the total capacity. However this workload can exceed 70% in the month preceding important conference deadlines such as the one of SC, GRID, CCGRID, IPDPS, etc. When the workload exceeds 90\% (this situation was observed in the Orsay site, the month preceding the SC deadline in 2006), users begin to complain, simply because they are not able to get enough resources or they get them after the deadline.

In addition to its service for research purpose, Grid'5000 is also used for education. A winter school was organized in 2006. 117 participants coming from different communities (computer science, physics, life science) attended courses where they learned how to use Grid'5000, how to run Globus GT 4 on it, how to deploy and run MPI applications on several sites, and how to reconfigure Grid'5000.

Grid'5000 is also used for large scale events. The Grid Plugtests (N-Queens and Flowshop Contests) has used Grid'5000 in a dedicated mode for several days in 2005 and 2006. The purpose of these events is to bring together users of the Proactive middleware and to test the deployment and interoperability of ProActive Grid applications. The Grid Plugtests, which consist of 2 competitions: the N-Queens Contest(find the number of solutions to the N-queens problem, N being as large as possible) and the Flowshop Contest. In 2006, the Grid Plugtests used more than 2600 CPUs during 2 days. For the second consecutive year, Grid'5000 has provided by far the largest number of CPUs among the participating Grids.

6 Conclusion

Grid'5000 belongs to a novel category of research tools for Grid and P2P research: large scale distributed platforms that can be easily controlled, reconfigured and monitored. The main difference between Grid'5000 and the previous real life experimental platforms is their degree of reconfigurability, allowing researchers to reconfigure the software stack (Grid'5000).

The construction of such multi-generations large scale instruments is new in computer science and the community is not used to deal with all the administrative, technical and scientific details related to the design, construction, exploitation, maintenance, upgrade and dismantlement of such platforms. Physicists involved in high energy physics and Astrophysicists have a long history of instrument construction behind them. This is a precious source of inspiration for computer scientists.

In addition to be an instrument to study Grid research problems, Grid'5000 belongs to a novel kind of facilities for computer scientists: platforms with resources opened and shared by a large community of users (typically hundreds). Computer scientists find in these platforms more than just resources they would not be able to access in other circumstances: they find a sophisticated environment involving supporting engineers, specific software, dedicated hardware to ease their experiments and also a social context in which they can share their problems, questions and solutions.

Acknowledgements

Many people are contributing to Grid'5000. Michel Cosnard initiated the project at the national level. Thierry Priol and Brigitte Plateau are respectively the Director of the ACI Grid and the President of the

Scientific Committee. We thank the other members of the steering committee: Michel Dayde, Frédéric Desprez, Emmanuel Jeannot, Yvon Jegou, StÈphane Lanteri, Nouredine Melab, Raymond Namyst, Olivier Richard, Pascale Vicat-Blanc Primet and Dany Vandromme. Pierre Neyron is the head of the Technical committee.

We would like to thank the French Ministry of research and the ACI Grid and ACI Data Mass incentives, INRIA, CNRS, RENATER, regional councils of Aquitaine, Bretagne, Ile de France and Provence Alpe Côte d'Azur, Alpes Maritimes General Council and the following Universities: University of Paris Sud, Orsay, University Joseph Fourier, Grenoble, University of Nice-Sophia Antipolis, University of Rennes 1, Institut National Polytechnique de Toulouse / INSA / FERIA / Université Paul Sabatier, Toulouse, University Bordeaux 1, University Lille 1 / GENOPOLE, Ecole Normale Supérieure de Lyon.

References

[1] P. Augerat, C. Martin, and B. Stein. "Scalable monitoring and configuration tools for grids and clusters". In Proceedings of the 10th Euromicro Workshop on Parallel, Distributed and Network-based Processing. IEEE Computer Society, 2002

[2] H. Casanova, A. Legrand, and L. Marchal. "Scheduling distributed applications: the simgrid simulation framework". In Proceedings of the third IEEE International Symposium on Cluster Computing and the Grid (CCGrid'03), may 2003

[3] C. Dumitrescu and I. Foster. "Gangsim: A simulator for grid scheduling studies". In Proceedings of the IEEE International Symposium on Cluster Computing and the Grid (CCGrid'05), Cardiff, UK, may 2005

[4] S. Genaud, M. Grunberg, and C. Mongenet. "Experiments in running a scientific MPI application on grid'5000". In IEEE International Workshop on Grid Computing (HPGC), IPDPS 2007. IEEE Society Press, march 2007

[5] Y. Georgiou, J. Leduc, B. Videau, J. Peyrard, and O. Richard. "A tool for environment deployment in clusters and light grids". In Second Workshop on System Management Tools for Large-Scale Parallel Systems (SMTPS'06), Rhodes Island, Greece, April 2006

[6] Y. Georgiou, O. Richard, P. Neyron, G. Huard, and C. Martin. "A batch scheduler with high level components. In Proceedings of CCGRID'2005". IEEE Computer Society, 2005

[7] X. Liu, H. Xia, and A. Chien. "Validating and scaling the micrgrid: A scientific instrument for grid dynamics". The Journal of Grid Computing, Volume 2(2):141-161, 2004

[8] M. Mezmaz, N. Melab, and E.-G. Talbi. "A Grid enabled Branch and Bound Algorithm for Solving Challenging Combinatorial Optimization Problems". In Proc. of 21th IEEE Intl. Parallel and Distributed Processing Symp., Long Beach, California, 26–30 Mar. 2007

[9] A. Takefusa, S. Matsuoka, K. Aida, H. Nakada, and U. Nagashima. "Overview of a performance evaluation system for global computing scheduling algorithms". In HPDC '99: Proceedings of the The Eighth IEEE International Symposium on High Performance Distributed Computing, page 11, Washington, DC, USA, 1999. IEEE Computer Society

Distributed Computational Methods for Coupled Fluid Structure Thermal Interaction Applications*

E. Aulisa[1] S. Manservisi[2] P. Seshaiyer[3,a] A. Idesman[4]

1 Department of Mathematics & Statistics, Texas Tech University, Lubbock, TX 79409, USA

2 DIENCA-Lab. di Montecuccolino, Via dei Colli 16, 40136 Bologna, Italy

3 Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030, USA

4 Department of Mechanical Engineering, Texas Tech University, Lubbock, TX 79409, USA

Abstract

The problem of efficient modeling and computation of the nonlinear interaction of fluid with a solid undergoing nonlinear deformation has remained a challenging problem in computational science and engineering. Direct numerical simulation of the non-linear equations, governing even the most simplified fluid-structure interaction model depends on the convergence of iterative solvers which in turn relies heavily on the properties of the coupled system. The purpose of this paper is to introduce a distributed multilevel algorithm with finite elements that offers the flexibility and efficiency to study coupled problems involving fluid-structure interaction. Our numerical results suggest that the proposed solution methodology is robust and amenable to parallelism.

Keywords: Fluid Structure Interaction, Finite Element Methods, Domain Decomposition, Multigrid Methods.

1 Introduction

Distributed computing has evolved rapidly in the last decade that has helped develop new computational methodologies to solve complex multi-physics problems involving fluid-structure interactions (FSI), efficiently. The efficient solution of such a coupled system provides predictive capability in studying complex nonlinear interactions that arise in several applications such as blood flow interaction with arterial wall to computational aeroelasticity of flexible wing micro-air vehicle, where the structural deformation and the flow field interact in a highly non-linear fashion. The direct numerical simulation of this highly non-linear system, governing even the most simplified FSI, depends on the convergence of iterative solvers which in turn relies on the characteristics of the coupled system.

Domain decomposition techniques with non-matching grids have become increasingly popular in this regard for obtaining fast and accurate solutions of problems involving coupled processes. The mortar finite element method[1,2] has been considered to be a viable domain decomposition technique that allows coupling of different subdomains with nonmatching grids and different discretization techniques. The method has been shown to be stable mathematically and has been successfully applied to a variety of engineering applications[3,4,5]. The basic idea is to replace the strong continuity condition at the interfaces between the different subdomains by a weaker one to solve the problem in a coupled fashion. In the last few years, mortar finite element methods have also been developed in conjunction with multigrid techniques[6,7,8,9,10]. One of the great advantages of the multigrid approach is in the grid generation process wherein the corresponding refinements are already available and no new mesh

* This work is supported in part by the National Science Foundation DMS 0610026 and the Advanced Research Program, THECB ARP 0212-44-C399.

[a] Corresponding Author Email: pseshaiy@gmu.edu

structures are required. Also, multigrid method relies only on local relaxation over elements and the solution on different domains can be easily implemented over parallel architectures. The purpose of this paper is to introduce a distributed multigrid algorithm that can be used to study different physical processes over different subdomains involving non-matching grids with less computational effort. In particular, we develop the method for a problem that involves a Fluid-Structure-Thermal interaction (FSTI). In section 2, the coupled model is described together with their finite element discretization. In section 3 the multigrid domain decomposition algorithm is discussed. Section 4 outlines the distributed computational methodology. In section 5, we present the numerical experiments for the benchmark application presented and follow that with discussion and conclusions in section 6.

2 Model Problem

In this section, we present a model for the interaction of a nonlinear structural domain interacting with a fluid medium. Note that for simplicity of presentation, we consider a model with a structural element to be a nonlinear beam and the methodology presented herein, can be extended to more complicated structural elements as well. Moreover, the methodology is described for a two-dimensional problem and can be extended to higher dimensions also.

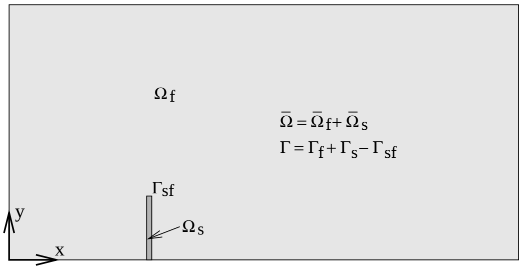

Ω_f
$\bar{\Omega} = \bar{\Omega}_f + \bar{\Omega}_s$
$\Gamma = \Gamma_f + \Gamma_s - \Gamma_{sf}$
Γ_{sf}
y
Ω_s
x

Figure 1 Computational domain for the FSTI problem

Let the computational domain $\Omega \subset \Re^2$ be an open set with boundary Γ. Let Ω be decomposed into the two disjoint open sets, a fluid subdomain Ω_f and a solid subdomain Ω_s with respective boundaries Γ_f and Γ_s. Let Γ_{sf} be the interior boundary between Γ_f and Γ_s, $\Gamma_f^e = \Gamma \cap \Gamma_f$ be the fluid exterior boundary and $\Gamma_s^e = \Gamma \cap \Gamma_s$ be the solid exterior boundary. For simplicity assume that the only boundary which can change in time is Γ_{sf}. The unsteady Navier-Stokes equations for incompressible flows are considered in the fluid domain, while the energy equation is solved in the whole domain. In the solid region, the nonlinear Euler-Bernoulli beam equation is considered. In this approximation plane cross sections perpendicular to the axis of the beam are assumed to remain plane and perpendicular to the axis after deformation [11] and under these hypotheses only a one-dimensional model is required for describing the axial and transverse deflections of the beam. We will denote by Λ the beam axis and by (ξ,η) a local reference system oriented with the ξ-axis parallel to Λ. Let δ and L denote the thickness and length of the beam respectively, Γ_{sf} is in $(\xi, \pm\delta/2)$ for $0 \le \xi \le L$ and in (L,η) for $-\delta/2 \le \eta \le \delta/2$. Let Γ_1 be the part of the fluid exterior boundary where Dirichlet boundary conditions are imposed for the velocity field $\vec{u} = (u_1, u_2)$; Neumann homogenous boundary conditions are considered on the remaining part. Similarly, let $\Gamma_2 \subset \Gamma$ be the part of the boundary where Dirichlet boundary conditions are imposed for the temperature T, while Neumann homogenous boundary conditions are considered on the rest of Γ. In $\xi = 0$, Dirichlet zero boundary conditions are imposed for the solid displacements and its appropriate derivatives. Conditions of displacement compatibility and force equilibrium along the structure-fluid interface are satisfied.

3 Decomposition & D iscretization

Let the domain Ω be partitioned into m non-overlapping sub-domains $\{\Omega^i\}_{i=1}^m$ such that the intersection of any two sub-domains is empty, a vertex, or a collection of edges of the respective domains. In the latter case, we denote this interface by Γ^{ij} which consists of individual common edges from the domains

Ω^i and Ω^j . Over each subdomain, a fully coupled system is solved for the solution variables, namely, the velocity, the pressure, the stress vector, the temperature, the heat flux and the beam displacements.

In order to account for the changing nature of the fluid and solid subdomains, one must define a dynamic mesh for the space discretization. However, to avoid extreme distortion, we choose to move the mesh independently of the fluid velocity in the interior of the fluid domain. Such a scheme, called arbitrary Lagrangian-Eulerian (ALE) formulation, is commonly applied when studying fluid-structure interaction [14, 15, 16, 17]. The structural equation is discretized in time by a using Newmark integration scheme [11]. In particular, the constant-average acceleration method was employed which is known to be stable for each time step and conserves energy for free vibration problem.

We then introduce a finite element discretization in each subdomain Ω^i through the mesh parameter h which tends to zero. Let $\{\Omega_h^i\}_{i=1}^m$ be the partition of the discretized domain Ω_h . Now, by starting at the multigrid coarse level

l = 0, we subdivide each Ω_h^i and consequently Ω_h into triangles or rectangles by families of meshes $T_h^{i,0}$. A typical refinement is illustrated in Figure 2.

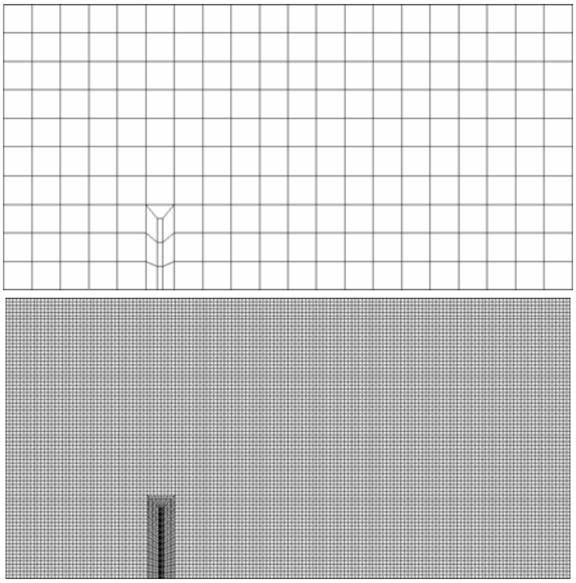


Figure 2 Coarse Mesh Level (top) and fine mesh Level (bottom) built with three consecutive refinements

Based on a simple element midpoint refinement different multigrid levels can be built to reach the finite element meshes $T_h^{i,n}$ at the top finest multigrid level l = n. At the coarse level, as at the generic multigrid level l, the triangulation over two adjacent subdomains, Ω_h^i and Ω_h^j, obeys the finite element compatibility constraints along the common interfaces. For details on multigrid levels and their construction one may consult [18, 19]. By using this methodology, we construct a sequence of meshes for each multigrid level in a standard finite element fashion with compatibility enforced across all the element interfaces built over midpoint refinements. In every subdomain Ω_h^i the energy equations can be solved over a different level mesh, generating a global solution over Ω_h , consisting mesh solutions at different levels over different subdomains. Let $\Omega_h^{i,l}$ be the subdomain i where the solution will be computed at the multigrid level *l*. It should be noted that the multigrid levels at which the solution is computed over adjacent subdomains, $\Omega_h^{i,l}$ and $\Omega_h^{j,k}$ may be different from each other $(l \neq k)$, with no compatibility enforced across their common interface.

4 Distributed Algorithm

The solution to the associated fluid-structure-thermal interaction problem is then achieved via an iterative strategy, where four systems of equations are solved separately and in succession, always using the latest information, until convergence is reached. An iterative multigrid solver in conjuction with a Vanka type smoother is used for the Navier-Stokes, the energy equation and the grid velocity equation systems. For the solution of the non-linear beam equation a direct nonlinear solver is used. The distributed computational algorithm employed is summarized in Figure 3. The Navier-Stokes, energy and grid-velocity systems are solved using a fully coupled iterative multigrid solver[20] with a Vanka-type smoother. Multigrid solvers for coupled velocity/pressure system compute simultaneously the solution for both the pressure and the velocity field, and they are known to be one of the best classes of solvers for laminar Navier-Stokes equations

[18,19]. The Vanka smoother employed in our multigrid solver involves the solution of a small number of degrees of freedom given by the conforming Taylor-Hood finite element discretization used. For this kind of element the pressure is computed only at the vertices while the velocity field is computed also at the midpoints. Examples of computations with this kind of solver can be found in [6,7,8,18,19]. In order to increase the convergence rate, the considered Vanka-type smoother has been coupled with a standard V-cycle multigrid algorithm. The multigrid does not change the nature of the solver, but allows the information to travel faster among different parts of the domain. A rough global solution is evaluated on the coarsest mesh *l = 0* and projected on the finer grid *l =1*, where Vanka-loops are performed improving its details. The updated solution is then projected on the mesh level *l = 2* and improved. The procedure is repeated until the finest mesh is reached. Solving the equation system in fine meshes improves solution details, but at the same time reduces the communication speed over the domain. However, this does not affect the global convergence rate since a considerable information exchange among different parts of the domain has been already done when solving in coarser mesh levels. All these considerations can be directly extended to the energy equation solver, where the same element block is considered.

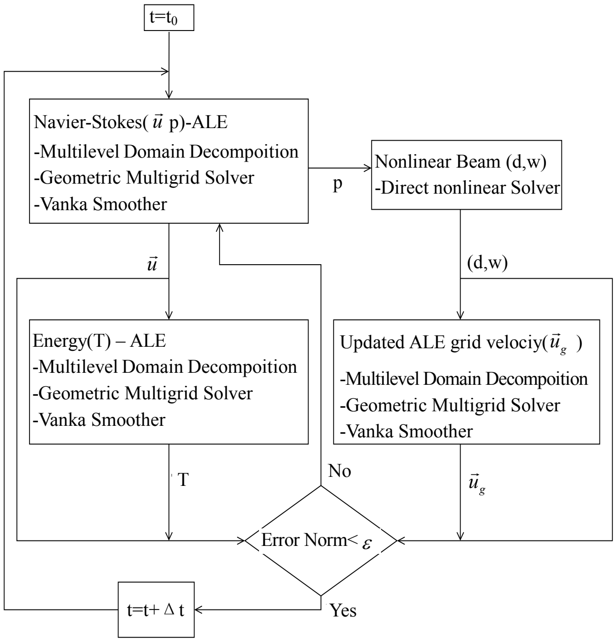

$t=t_0$
Navier-Stokes($\vec{u}$ p)-ALE
-Multilevel Domain Decompoition
-Geometric Multigrid Solver
-Vanka Smoother
p
Nonlinear Beam (d,w)
-Direct nonlinear Solver
$\vec{u}$
(d,w)
Energy(T) – ALE
-Multilevel Domain Decompoition
-Geometric Multigrid Solver
-Vanka Smoother
Updated ALE grid velociy($\vec{u}_g$)
-Multilevel Domain Decompoition
-Geometric Multigrid Solver
-Vanka Smoother
T
No
$\vec{u}_g$
Error Norm< ε
$t=t+\Delta t$
Yes

Figure 3 Distributed Computational Methodology

5 Numerical Results

In this section we test the performance of distributed multilevel formulation for the FSTI application presented in the paper. Let the rectangular region $\Omega=[4m]\times[2m]$ be the computational domain with boundary Γ (Figure 1). The solid region Ω_s consists of a beam clamped at point (*1m, 0*) with length equal to *0.5m* and thickness equal to *0.04m*. On the left side of the domain inflow boundary conditions are imposed for the velocity field $\vec{u}=(u_1,u_2)$ with parabolic profile $u_1=0.1y(2-y)$ *m/s* and $u_2=0$. On the right side of the domain outflow boundary conditions are imposed while on the remaining part of the boundary non-slip conditions are considered. The temperature is set equal to zero in the inlet region and to 100^{o}C on the solid boundary where the beam is clamped. Adiabatic conditions are imposed on the rest of the domain. The initial conditions for both the temperature and the velocity field are zero. The fluid and the solid properties are chosen in order to produce a large deformation of the beam. This choice implies strong interactions among all the parts of the system and test the reliability of the solver in challenging situations. In the Navier-Stokes system, the fluid density, the viscosity, the volumetric expansion coefficient and the reference temperature are equal to 100 kg/m^3, 0.01 kg/ms, 0.01 K^{-1} and 0^{o}C, respectively. In the energy equation the solid density is 200 kg/m^3, while the heat capacity and the heat conductivity, are 100 J/kg K and 10 W/m K in the fluid region, and 10 J/kg K and 400 W/m K in the solid region respectively. The stiffness for unitary length of the beam is equal to $ 1 kg m^2/s^2. In all the simulations the same time step $\Delta T=0.01$s is used, for a total of 500 time steps (5 seconds). Only the four level meshes, l_0, l_1, l_2, l_3, are considered and in Figure 2, the two different level meshes, l_0 and l_3 are shown. The coarse mesh level l_0 has 207 elements, while the mesh level l_3 obtained after three consecutive midpoint refinements has 13248 elements. The one-dimensional mesh on the beam axis follows the same midpoint refinement algorithm used for the two-dimensional computational domain. On the coarse level l_0 three elements are available, while on the fine grid l_3 after 3 refinements, the number of elements becomes 24. Since the number of unknowns is quite small, the solution of the non-linear beam equation is always evaluated on the finest mesh using a direct non-linear solver.

The results obtained with our coupled model (case C) are compared with the results obtained for the same geometry with a rigid beam, and zero buoyancy force (case A), and with the results obtained neglecting only the effects of the non-linear term in the beam equation (case B). All the computations are done at the time t = 5 s and over the finest level l_3. In Figure 4 on the top, the beam bending and the corresponding grid deformation are displayed, showing the strong influence of the pressure load on the beam shape. Figure 4 on the bottom shows the velocity field map and clearly indicates that the stationary solution is not reached since new vortices are constantly created and advected towards the outflow region.

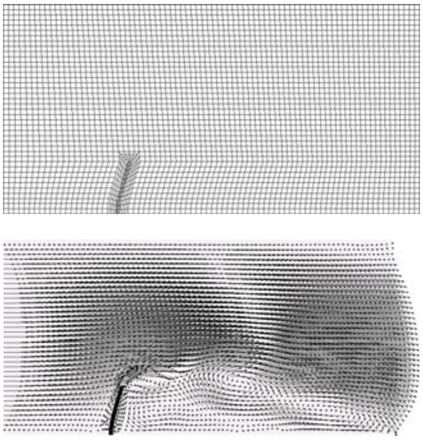


Figure 4 Beam bending and grid movement (top); Velocity Field Map (bottom)

In Figure 5a, the two components of the velocity field profiles evaluated over the section *y=0.5* for *x* in [0, 4], are shown for all the three cases. Figure 5b plots similar results for the pressure (top panel) and the temperature (bottom panel) profiles. The combined effect of the beam deflection and of the buoyancy force (case B and C) modify considerably all the profiles obtained in case A. Even the differences between case B and C are not negligible.

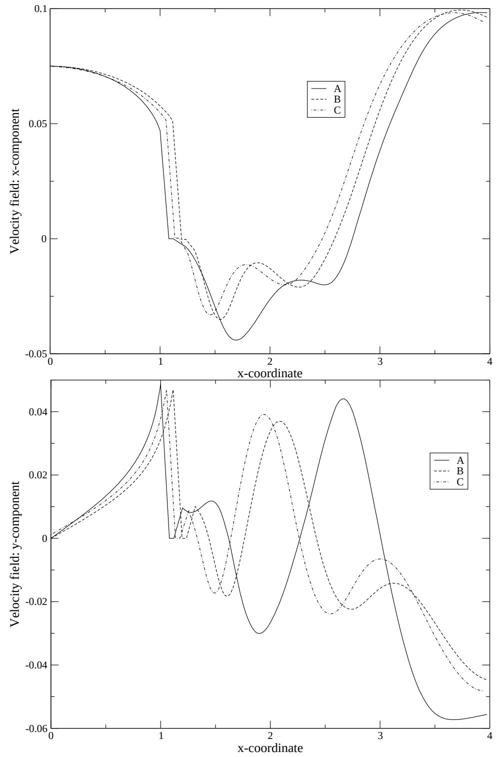


Figure 5a Profiles of the two velocity components along y=0.5 for x in [0, 4] at the time t=5 for Cases A, B, C

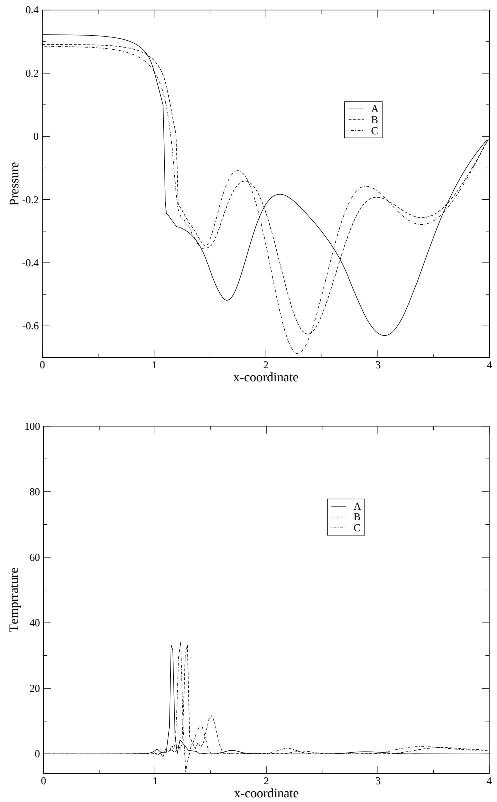


Figure 5b Pressure (top)& Temperature (bottom) profiles along y=0.5 for x in [0, 4] at the time t=5 for Cases A, B, C

The presence of the nonlinear term in the beam equation has considerable effects on all the solution profiles, pointing out how sensitive nature of the interaction among all the parts of the coupled system. The number of unknowns (including velocity field, pressure, temperature and displacement), involved in the computation at the mesh level l_3 is quite large, approximately 94000. However Figure 4 on the bottom illustrates that the only part of the system, subjected to high vorticity is the region downstream of the beam. In the region upstream of the beam and in the upper part of domain, the velocity field is almost stationary.

We remark that in order to obtain more efficient distributed computations, the solution must be evaluated at varying mesh levels. For instance, performing the computations at mesh levels l_2 or l_1 in parts of the domain where the mesh level l_3 is not needed can save a lot of degrees of freedom. To evaluate the efficiency of the computations we split the computational domain Ω into three subdomains Ω^1, Ω^2, Ω^3 over which three different non-conforming meshes are built, respectively. The subdomains and the three different non-conforming partitions are shown in Figure 6a and 6b respectively.

In the subdomain Ω^3 which is the solid domain the mesh level l_1 is always used. In the first configuration, P_1 (top), the mesh levels, l_2 and l_1, are considered for the subregions Ω^2 and Ω^1, respectively. The different couplings of level meshes, l_3 - l_2 and l_3 - l_1 are used in the same subregions for the second configuration P_2 (middle), and a third configuration P_3 (bottom-right). The numbers of nodes is greatly reduced for all the three non-conforming configurations. In particular approximately 11000, 39000 and 28000 are the new number of unknowns for the new configurations P_1, P_2 and P_3, respectively. The computational CPU time and the memory allocation expenses are consequently reduced. In Figure 7, the deflection of the beam extreme point is compared (for the three conforming meshes l_0, l_1 and l_3, and for the 3 non-conforming meshes P_i, i=1..3 for both the linear (top-panel) and the non-linear (bottom-panel) cases. The results show clear advantages of the non-conforming discretizations over the conforming ones. Obviously the path obtained with the finest mesh l_3 can be considered the most accurate. The l_2 path is very close to the l_3 in the first second but differences appear as soon as the time increases. The l_1 path is always below the l_3, showing too much stiffness in the beam response. The beam oscillation obtained with the non-conforming configuration P_1 perfectly overlaps the result obtained with the conforming mesh l_2, and the result obtained with the configuration P_2

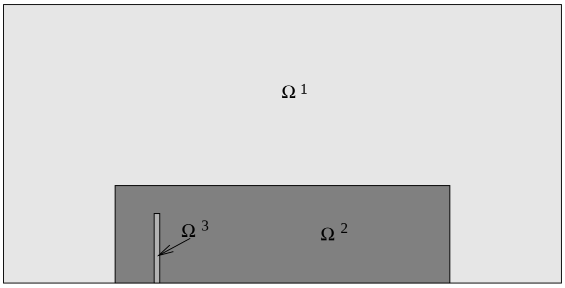


Figure 6a Domain decomposition for efficient computation

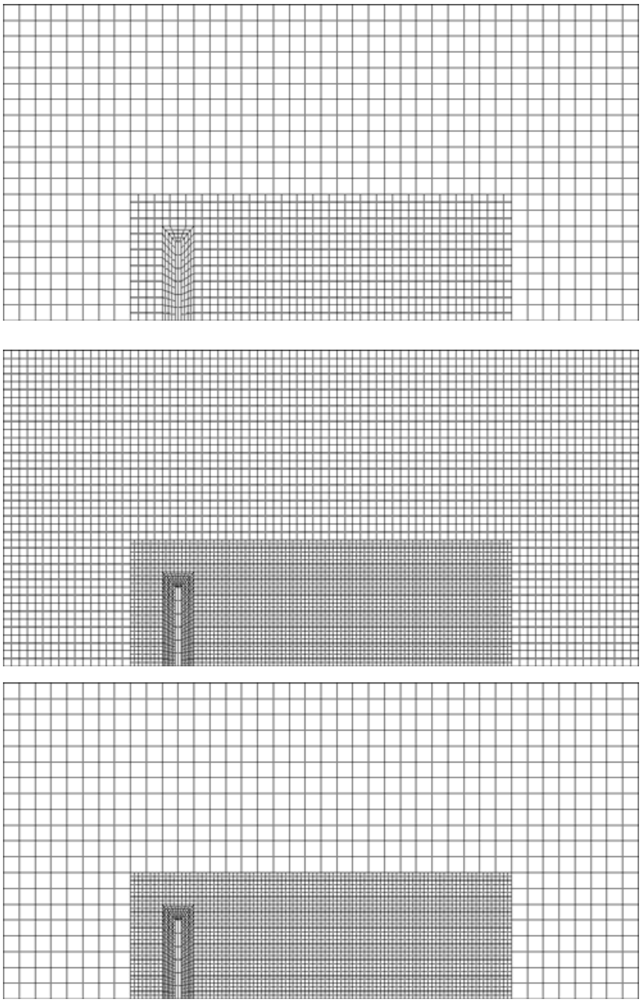


Figure 6b Different non-conforming configurations P_1 (top panel), P_2 (middle panel), P_3 (bottom panel)

perfectly overlaps the result in l_3. It is possible to find very small differences between the path in l_3 and the path in P_3, where there are two mesh levels between the two adjacent regions Ω^1 and Ω^2. These results clearly indicate how one can use the non-conforming multilevel partitioning to preserve the same accuracy in regions of interest, reducing at the same time the number of degrees of freedom in other parts of the domain. It must also be pointed out that the nonlinear beam case (bottom-panel) yields a deflection that is much stiffer than the linear beam case (top-panel). This suggests the importance of the influence of the coupling between the axial and transverse beam deflection when considering the overall coupled system.

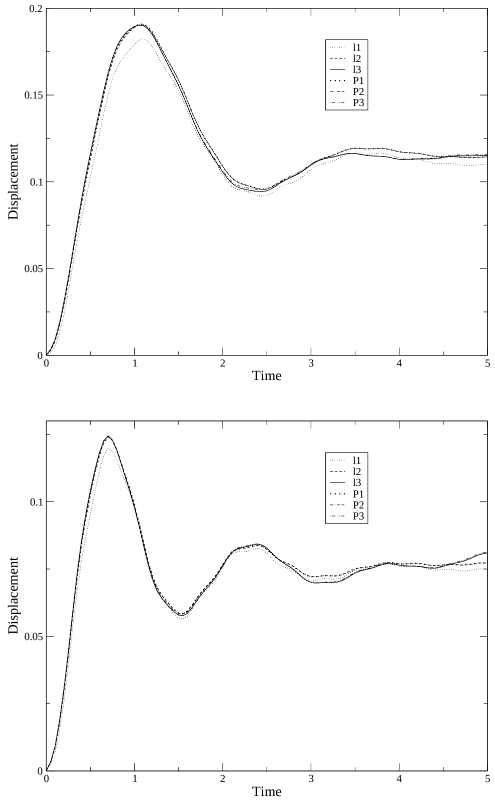


Figure 7 Deflection of the beam extreme point for conforming and nonconforming meshes: Linear (top-panel) and Nonlinear (bottom-panel)

6 Discussion And Conclusions

This paper presents a distributed computational methodology for solving Fluid-Structure-Thermal interaction problems. A benchmark application that models the interaction of a nonlinear beam structure in a fluid medium along with temperature equations has been studied and tested. Our computational results clearly indicate that the methodology described in conjunction with the multilevel multigrid method leads to a fast and flexible algorithm to solve the associated coupled FSTI problem.

In a recent work [6], we had tested the performance of the computational methodology presented in a parallel environment to study flow through a L-shaped channel. We considered four different configurations of meshes over the L-shaped domain, namely Case A: fine grid l_3, Case B: coupled levels l_2 - l_3, Case C: coupled levels l_1-l_3 and Case D: coupled levels l_0 - l_3. We compute and compare the solutions of the problem obtained by using 1, 2, 4, 8, and 16 processors respectively. The results are presented in Figure 8. We hope to study a similar performance of the methodology presented in this paper to more complex applications involving fluid-structure interaction in terms of load balancing and scalability with a parallel infrastructure, which will be the focus of a forthcoming paper.

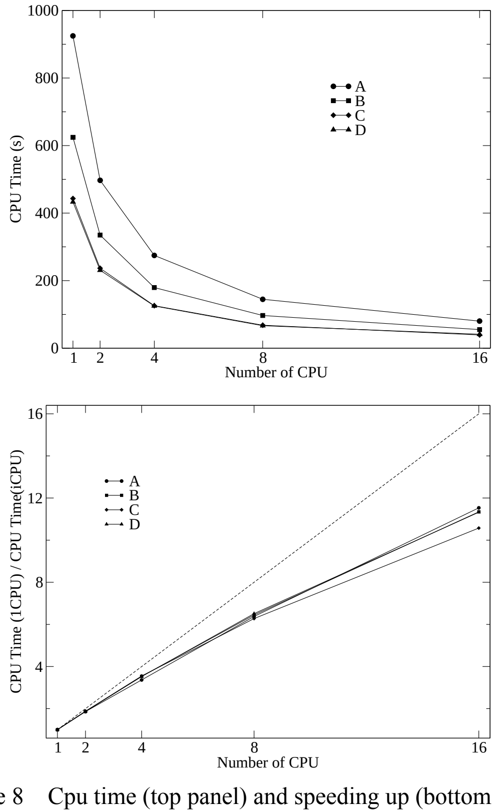


Figure 8 Cpu time (top panel) and speeding up (bottom panel) with fixed norm residual (10^{-13}) as a function of the number of the processors for the different Cases *A*,*B*,*C* and *D* with data exchange at the end of each block relaxations

References

[1] C. Bernardi, Y. Maday and A. Patera, “Domain decomposition by the mortar element method”, Asymptotic and numerical methods for partial differential equation with critical parameters. H.K. et al. eds, Reidel, Dordecht, pp. 269-286, 1993

[2] F. Ben Belgacem, “The mortar finite element method with Lagrange Multipliers”, Numer. Math., vol. 84(2), pp. 173-197, 1999

[3] P. Seshaiyer and M.Suri, “hp submeshing via non-conforming finite element methods”, Comput. Meth. Appl. Mech. Eng., vol. 189, pp. 1011-1030, 2000

[4] F. Ben Belgacem, L.K. Chilton and P. Seshaiyer, “The hp-Mortar Finite Element Method for Mixed elasticity and Stokes Problems”, Comput. Math. Appl., vol. 46, pp. 35-55, 2003

[5] F.Casadei G.Fotia, E.Gabellini, F.Maggio and A.Quarteroni, “A mortar spectral/finite element method for complex 2D and 3D elastodynamic problems”, Comp. Methods Appl. Mech. Engnrg., vol. 191, pp.5119-5148, 2002

[6] E. Aulisa, S. Manservisi and P. Seshaiyer, “A computational multilevel approach for solving 2D Navier-Stokes equations over non-matching grids”, {Comput. Meth. Appl. Mech. Eng.}, vol 195, pp 4604-4616, 2006

[7] E. Aulisa, S. Manservisi and P. Seshaiyer, “A non-conforming computational methodology for modeling coupled problems”, Nonlinear Analysis, vol. 6, pp. 1445-1454, 2005

[8] E. Aulisa, S. Manservisi and P. Seshaiyer, “A multilevel domain decomposition approach to solving coupled applications in computational fluid dynamics”, International Journal for Numerical Methods in Fluids, vol. 56, pp. 1139-1145, 2008

[9] D. Braess, W. Dahmen and C. Wieners, “A Multigrid Algorithm for the Mortar Finite Element Method”, SIAM J. Num. Anal., vol. 37(1), pp. 48-69, 1999

[10] J. Gopalakrishnan and J.E. Pasciak, “Multigrid for the Mortar Finite Element Method”, SIAM J. Num. Anal., vol. 37(3), pp. 1029-1052, 2000

[11] J.N. Reddy, “An Introduction to Nonlinear Finite Element Analysis”, Oxford University Press, Oxford, 2004

[12] V. Girault and P. Raviart, “The Finite Element Method for Navier-Stokes Equations: Theory and Algorithms”, Springer, New York, 1986

[13] R.~Temam, “Navier-Stokes equation”, North-Holland, Amsterdam, 1979

[14] E.W. Swim and P. Seshaiyer, “A nonconforming finite element method for fluid-structure interaction problems”, Comput. Meth. Appl. Mech. Eng., vol. 195(17-18), pp. 2088-2099, 2006

[15] J. Donea, S. Giuliani, J. Halleux, “An arbitrary Lagrangian Eulerian finite element method for transient fluid-structure interactions”, Comput. Meth. Appl. Mech. Eng., vol. 33, pp. 689-723, 1982

[16] C. Grandmont, V. Guimet, Y. Maday, “Numerical analysis of some decoupling techniques for the approximation of the unsteady fluid structure interaction”, Math. Models Methods Appl. Sci., vol. 11, pp. 1349-1377, 2001

[17] T. Hughes, W. Liu, T. Zimmermann, “Lagrangian Eulerian finite element formulation for incompressible viscous flows”, Comput. Methods Appl. Mech. Engrg., vol. 29, pp. 329-349, 1981

[18] M. Schafer and S. Turek, “The benchmark problem: flow around a cylinder, in Flow simulation with high performance computers II”, Notes on Numerical Fluid Mechanics , E.H. Hirschel ed., vol. 52, pp. 547, 1996

[19] S. Turek, “Efficient solvers for incompressible flow problems: an algorithmic and computational approach”, Lecture Notes in computational science and engineering, Springer, vol. 6, 1999

[20] S.Vanka, “Block-implicit multigrid calculation of two-dimensional recirculation flows”, Comput. Meth Appl. Mech. Eng., vol. 59(1), pp. 29-48, 1986

The Improvement of the Distributed Gmres(m) Algorithm and Application in Elastic BEM

Chunfeng Liu[1,2] Aimin Yang[1] Xinghua Ma[1] Nan Ji[1] Yanbing Liang[1]

1 College of Science, Hebei Polytechnic University, Tangshan, Hebei, 063000, China

2 Hebei University of Technology, Tianjin , 300130 ,China

Email:liucf403@163.com; aimin@heut.edu.cn

Abstract

The GMGRE(m) algorithm is used to solve linear equations set. With the rapid development of high-speed network technology, the distributed methods have become the important way to solve the actual questions. The distributed GMRES(m) algorithm based on Galerkin principle and boundary element method is improved and applied to solve the large-scale elastic problem. Taking the single object elasticity object which has a hole as an example, the application of the distributed GMGRE(m) algorithm to the BEM is presented. The outcome demonstrates that using distributed GMGRE(m) algorithm in BEM, the higher computing and accuracy can be obtained. Because the methods make the partition of boundary nods more optional, the example implies the methods more accurate and efficient than the finite element.

Keywords： distributed GMRES(m) algorithm; the Boundary Element Method; numerical calculation

1 Introduction

As a forceful numerical analytical method, for its more accuracy and low dimension, BEM has already become an accurate, efficient engineering numerical analytical method during the past 30 years. It’s largely used in the analytical computing of many subjects, especially in the field of computing mathematics and computing mechanics. It is considered as the most important supplement. But the coefficient Matrix of the equation sets obtained through the BEM is usually dissymmetrical. When there are more cells, the computing and memory will increase greatly if using the BEM to solve large-scale problem. Therefore, it largely limits the application of BEM in large-scale problems(Please read the reference [1-4]).

Parallel computing is a good way to improve computing speed and enlarge solving scale. It provides powerful means to save large-scale BEM problem. Therefore, the research of parallel BEM is important in improving the solving scale of BEM and decreasing the computing time, and also enlarging the fields which use the BEM(Please read the reference [5-7]).

Most problems of science and engineering technology are solved by linear system of equations. In 1986 GMRES was put forward by Yousef Saad and Martin H.Schultz, which is an iterative algorithm to solve large linear algebra system equations whose coefficient matrix is asymmetrical (Please read the reference [8-10]). Now the algorithms of Arnoldi, GMRES and GRMES (m), based on Galerkin Theory, are the fundamentals to the new algorithms, and GMRES algorithm is regarded as one of effective solutions to solve large asymmetrical linear system equations. Saad and many other professionals have given a comprehensive introduction of GMRES algorithm, whose astringency and utility have been proved by numerical experiments. But because of a large number of floating-point arithmetic involved in GMRES algorithm resulting in calculation amount increases in exponential order, the research of the parallel GMRES method is of great significance to get the GMRES method to solve the practical problems(Please read the reference [11]).

2 The GMRES (M) method

Suppose the system of equations is $Ax=b$, in which A is a nonsingular large matrix, $b \in R^n$ is a known vector and the norm herein after is 2-norm. K_m and L_m are m dimensional subspaces, which are generated from $\{v_i\}_{i=1}^{m}$ and $\{w_i\}_{i=1}^{m}$. Supposing $x_0 \in R^n$ is a random vector and $x = x_0 + z$, $Ax=b$ is equivalent to $Az = r_0$ in which $r_0 = b - Ax_0$.

Galerkin Theory used in $Az = r_0$ can be stated that approximate result z_m is sought in the subspace K_m so as to get the residual vectors $r_0 - Az_m$ and all vectors in L_m reach orthogonality.

If we choose $L_m = K_m$, we call this Galerkin Method Arnoldi Algorithm; if we choose $L_m = AK_m$, we call this Galerkin Method as GMRES Algorithm. GMRES Algorithm has been improved greatly with the efforts from many professionals. It also has become the main method to solve large asymmetrical linear system equations through being integrated with various pretreatment technologies.

On the basis of the analysis of the upward section, we choose $K_m = span\left\{r_0, Ar_0, \cdots A^{m-1}r_0\right\}$, so we can find a set of standard orthogonal bases in K_m. Then

$$\|r_0 - Az\| = \|r_0 - AV_m y\| = \|r_0 - AV_{m+1}\bar{H}_m y\| = \left\|V_{m+1}\left(\beta e_1 - \bar{H}_m y\right)\right\|$$

is got.

Because $V_{m+1}^T V_{m+1} = I$, $\|r_0 - Az\| = \|\beta e_1 - \bar{H}_m y\|$.

So minimizing $\|r_0 - Az\|$ in R^n equals to minimizing $\|\beta e_1 - \bar{H}_m y\|$ in K_m, which can be eventually concluded into solve least squares equation $\min\|\beta e_1 - \bar{H}_m y\|$.

The calculation process of GMRES Method can be concluded into,

(1) Select x_0, then calculate $r_0 = f - Ax_0$ and $v_1 = r_0 / \| r_0 \|$;

(2) Iterate $For \quad j = 1, 2, \cdots, k, \cdots$ till meeting the needs of ***do***

$$h_{ij} = (Av_j, v_i) \quad (i = 1, 2, \cdots, j)$$

$$\hat{v}_{j+1} = Av_j - \sum_{i=1}^{j} h_{ij} v_i \,; \quad h_{j+1,j} = \| \hat{v}_{j+1} \|$$

$$v_{j+1} = \hat{v}_{j+1} / h_{j+1,j}$$

(3) Construct an approximate solution

$$x_k = x_0 + V_k y_k$$

in which y_k satisfies $\min J(y)$ ($J(y) = \| \beta e_1 - \bar{H}_k y_k \|$).

Theoretically speaking, if $\left\{A^i r_0\right\}_{i=0}^{n-1}$ near independence, while $m = n$, GMRES(m) algorithm should offer the accurately solution, but when m is very big, all the $\left(v_i\right)_{i=1}^{m}$ must be saved in the calculation, which will cause memory empty more larger to large scale problem, so it is unpractical. And when $k \to \infty$, not only internal memory and the amount of calculating are increasing, but also the orthogonalily of each array in the matrix V_k becomes relatively poor, this time the solution will oscillation in a small domain. While, after the original algorithm is pretreated, the difficulty is overcome when the technology of over again opening is supplied, then the GMRES(m) algorithm is obtained.

The concrete realized steps of the GMRES(m) algorithm are:

(1) let

$$x_0 = 0,\ r_0 = b - Ax_0,\ \beta = \| r_0 \|,\ v_1 = r_0 / \beta,\ V_1 = \{ v_1 \}$$

(2) iteration: $For \quad j = 1, 2, \cdots, m \quad do$

$$h_{ij} = (Av_j, v_i)\,(i = 1, 2, \cdots, j),$$

$$\hat{v}_{j+1} = Av_j - \sum_{i=1}^{j} h_{ij} v_i$$

$$h_{j+1,j} = \| \hat{v}_{j+1} \|, \quad v_{j+1} = \hat{v}_{j+1} / h_{j+1,j}$$

$$V_{j+1} = (V_j, v_{j+1}),$$

$$\bar{H}_j = \begin{pmatrix} \bar{H}_{j-1} & h_{ij} \\ 0 & h_{j+1,j} \end{pmatrix}_{(j+1)\times j}$$

$\bar{H}_j$ is a upper Hessenberg matrix，when $j = 1$, the first array is omission, and $AV_m = V_{m+1}\bar{H}_m$。

(3) solve the least square problem

$$\| r_m \| = \min_{y_m \in R^m} \| \beta e_1 - \bar{H}_m y_m \|,$$

and y_m is obtained;

(4) conform the proximately solution

$\boldsymbol{x_m = x_0 + V_m y_m}$

(5) calculate the modulo of the residual vector

$\| \boldsymbol{r_m} \| = \| \boldsymbol{b - A\, x_m} \|$

(6) judge of again activation

$$\| r_m \| \le \varepsilon \begin{cases} yes : x = x_m \ and\ stop \\ no\ : x_0 = x_m \ and\ turn\ to \end{cases} \qquad (1)$$

ε is the established reliance of convergent judgment.

3 The distributed GMRES(M) method

For the dissymmetrical matrix $A \in R^{n\times n}$, to solve the equation $Ax = b$, the steps of the distributed GMRES(m) algorithm are as follows:

(1) $\forall X_0 \in R$, setup parameter ξ, α, β, m 。

(2) calculate $r_0^{(i)} = b - A_i X_0$ in each CPU $P_i (i = 1, 2, \cdots, P)$, get $r_0 = \sum_{i=1}^{P} r_0^{(i)}$ and $\|r_0\|$ through communication, then sent out r_0 and $\|r_0\|$ to P_i.

(3) iteration: DO $k = 1$, n Calculate $A_i v_k$ in P_i, get $Av_k = \sum_{i=1}^{P} A_i v_k$ through communication.

Such calculation will be run in P_i as:

```
$h_{ik} = (Av_k, v_i), i = 1, 2, \cdots, k$
$\hat{v}_{k+1} = Av_k - \sum_{i=1}^{k} h_{ik} v_i$
$h_{k+1,k} = \| \hat{v}_{k+1} \|$
$v_{k+1} = \hat{v}_{k+1} / h_{k+1,k}$
let $\alpha_0 = \max_i \{ \| v_{k+1} \|, \| v_i \| \} \quad i = 1, 2, \cdots, k$
$f_k = | \bar{e}_m^T g_m |$    (to distributed GMRES(m) algorithm)
IF ($f_k < \xi$) THEN
$X_k = X_0 + V_k y_k$
GOTO (4)
END IF
IF ($k = m$ and $\alpha_0 > \alpha$) THEN
$X_k = X_0 + V_k y_k$
let $\boldsymbol{X_0 = X_k}$
GOTO (2)
        END IF
        let $\beta_0 = f_k - \min_i f_i \quad (i = 1, 2, \cdots, k)$
        IF ($\beta_0 > \beta$), THEN
            let $l$:
$\min_i f_i = f_l$
            $X^{(l)} = X_0 + V_l y^{(l)}$
        $X_0 = X_l$
        GOTO (2)
END IF
END DO
```

(4) the calculation will be independently accomplished in P_i,.

In these steps, the uppercase letter express matrix, the lowercase letter express vector.

4 The application of the distributed GMRES(M) algorithm to the BEM

In the numerical simulation of the engineering problem, the discrete equation matrix must be formed whether using BEM or FEM. After the boundary integral equation is dispersed, the affecting coefficient matrix is often dissymmetrical dense matrix. But the distributed GMRES(m) algorithm is based on the Arnodi algorithm, and it is a numerical method to solve linear equation sets with dense coefficient matrix, the time of the computing is proportional to $\boldsymbol{N}$, so it is very effective to use distributed GMRES(m) algorithm to solve large scale BEM coefficient equation sets. Taking the elasticity problem as an example, the application of the distributed GMRES(m) algorithm to the BEM problem is analyzed.

When using BEM to solve elasticity problem, firstly the boundary surface of the object is defined, then the surface is dispersed into several quadrangle linear cells or other kind of cell. At the same time, the composbe given. The serial number of the nodes should beition of the nodes of the cell and the coordinate of the nodes should counter-clockwise, and the outer low vector of the cell should beside the boundary. According to the already known conditions of the force and the displacement, we can confirm the boundary conditions of each node and cell. Then solve the problems repeatedly using the distributed GMRES(m) algorithm. Finally construct the new familiar solution. If the familiar solution satisfies the requirement of the accuracy, step repeating, or continue to repeat till it satisfies the requirement.

5 Computing example

In order to validate the dependability and the accuracy of the distributed GMRES(m) algorithm, an example is given as follows: the object is a single elasticity non with holes, and its one side is fixed.

The computing model and its dispersed framework model of the elasticity object A (200mm×200mm×25mm, 30mm×25mm) are separately as Fig.1 and Fig.2. The dispersed date of the computing model can be seen in table 1.

The elasticity modulus of the object A is $E = 210GPa$, the ratio of Possion is $v = 0.3$, the even load is $P = 10MPa$. From Fig.3 and 4, we can see the displacement of the nodes (along Z) among the elastic object A and the surface force and compare with the FEM. During the computing, we set $m = 32$ in the distributed GMRES(m) algorithm, the total computing time is 4 hours, 20 minutes, 30 seconds and 7 milliseconds, but the time of the FEM is 40 hours, 29 minutes, 7 seconds and 32 milliseconds.

According to the outcome of this example, it is noticed that the displacement and the surface force of the elasticity object distribute reasonably, which proves the accuracy of the model and stability of the solution. That is to say, using distributed GMRES(m) to solve the BEM problems, a model with higher computing efficiency, accuracy and stability can be obtained. When comparing with FEM, it is noticed that the solution is nearer to exact solution, and more tallies with the fact.

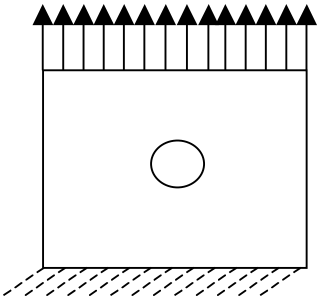


Figure1 Computing model (partical)

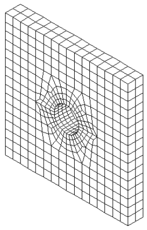


Figure2 Discrete framework model (partical)

Table 1 Discrete date

total nodes	total cells	total outer nodes	total nodes in the hole
936*10^3	936*10^3	18*10^3	32*4*10^3

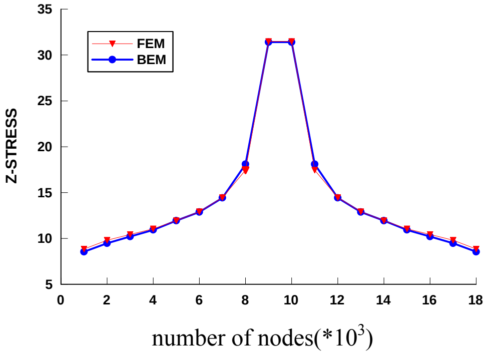


Figure3 Z-displace of A

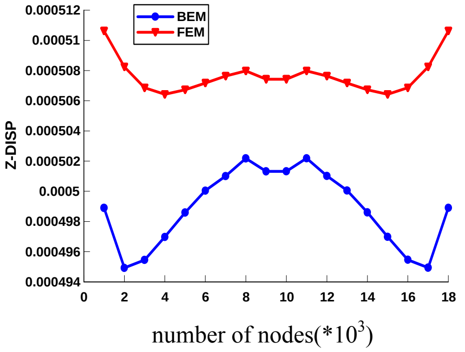


Figure4 Z-stress of A(The comparison of BEM and FEM)

6 Conclusion

In this paper, the distributed GMGRE(m) algorithm is used to solve linear equations set. Taking the single object elasticity object which has a hole as an example, the application of the distributed GMGRE(m) algorithm to the BEM is presented. The outcoe demonstrates that using distributed GMGRE(m) algorithm in BEM, the higher computing and accuracy can be obtained.

Acknowledgements

This paper is supported by the Natual Scientific Fund Project of Hebei Province (E2004000245), Scientific Technology Research and Development Plan Project (07160219B-6), Scientific Fund Project of Hebei Polytechnic University (z0808).

References

[1] Minghua XU, The Reoffered GMRES(m) Algorithm with Proper Parameters, Academy Newspaper of Jiangsu Petrochemical College, 1999, 11(3): 52-55

[2] W. Michacl, The Research and Development of Parallel Computation, Parallel Theory and Practice, 2000

[3] Xiaomei LI, Parallel Algorithm, Science and Technology Press, Changsha, Hunan, 1992

[4] Yan ZHANG, Distributed Parallel Algorithm Designing, Analysis and Realization Doctor Thesis of Electronic Science and Technology University, 2001

[5] Xin YIN, Three-dimensional Elastic BEM Parallel Calculation and Its Engineering Application, Engineering Doctor Degree Papers from Tsinghua University, 2000: 1-30

[6] Tian CHEN, Frictional Contact Analysis of Multiple Cracks by Incremental Displacement and Resultant Traction Boundary Integral Equations, Engn. Anal.Bound.Elem, 1998, 21:385-392

[7] R.W.Hockney, The Science of Computer Benchmarking, The Society for Industrial and Applied Mathematics, Philadelphia, 1996:6-12

[8] G.F.Psister, Clusters of Computers: Characteristics of an Invisible Architecture, IEEE Int'l. Parallel Processing Symp, Honolulu, 1996

[9] Saad Y, Schultz M H, GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetrical Linear System, SIAMJ Sci Comp, 1986,7(3):856-869

[10] Xiao HAN, Development of Simulator for Rolling In: M.Tanaka and Z.Yao eds. Boundary Element Method, Proceedings of the 7th Japan-China Symposium in Boundary Element Methods, Fukuoka Japan, 1996:287-296

The Mobile Agent Technology-Based Mobile Electronic Commerce

Zhengjian Ding[1] Bin Han[2]

1 School of Computer and Communication Lanzhou University of Technology, Lanzhou 730050, China
Email: Dingzj@lut.cn

2 School of Computer and Communication Lanzhou University of Technology, Lanzhou, 730050, China
Email:Phoebe5678@sina.com

Abstract

Today, Electronic Commerce based on Internet has become an important business running pattern. Meanwhile, mobile telecommunication is incorporating Internet. With the increasing number of mobile users, providing mobility and customized services has become a new trend of the development of electronic commerce. These applications can collectively be termed "Mobile Electronic Commerce". The mobile Electronic Commerce not only strengthens the applications of Internet and mobile telecommunication, but also is a complement to the traditional business activities. This paper examines about the role of Mobile Agent connected with Mobile Electronic Commerce at present, propose a 4-level architecture for the mobile E-commerce base on the Mobile Agent and the problems that still exist in the filed of the mobile E-commerce base on the Mobile Agent.

Keywords: Agent, Mobile Agent, Mobile E-Commerce, Ontology, UDDI

1 Introduction

Mobile Electronic Commerce is an emerging new computing environment incorporating both wireless and high speed networking technologies. Users equipped with personal digital assistants or PDAs will have access to a wide variety of services that will be made available over national and international communication networks. Mobile users will be able to access their data and other services such as electronic mail, electronic news, map services, and electronic banking while on the move. To receive these services, mobile users will be connected to fixed networks via wireless networks (or mobile networks)[1].

This paper will present a comprehensive solution to Mobile Electronic Commerce and some key technologies that used to implementing the Mobile Agent-based Mobile Electronic Commerce.

1.1 Mobile electronic commerce

The meaning of Electronic Commerce is very rich, including online trading, online searching, certificate authority, security, advertisement and etc.. These services provide practical, convenient, secure and complete environment of transaction to two parts of the deal in the pattern of B2B and B2C.

The information that comes from Internet is increasing drastically. The environment of network is more and more complicated, which makes it uncertain for the separated entities to participate in the Electronic Commerce. Enterprise hopes to build more intimate relationship over internet and reduce traditional delay in the transaction. For individuals, they hope to obtain more intelligent, active, and customized services. All of these requirements become the challenge to the traditional Electronic Commerce. In short, currently traditional Electronic Commerce can't achieve our tasks effectively[2].

At present, follow the development of wireless technologies and the increasing of mobile users, the

Mobile Commerce is presented to us. The development and the conformity of the mobile commerce industry chain will be accelerated, and will enrich the applications. The mobile electronic commerce is the development of the electronic commerce. The applications we can use including the mobile payment, mobile ERP, mobile stock market, mobile bank, mobile office, and so on. It can realize the personal information management (PIM), bank operation, business, shopping, location-based service, and so on.

1.2 Agent and mobile agent

If examine a comprehension about agent technology and agent services shortly, is as following. Characteristics that agent is different form exist applications are autonomy that do one’s action Elevation ability of intelligence through experience studying, and from contact part with human necessarily necessary sociality etc. The advent of agent technology and platform independent computer language gives people new methods to implement high-efficiency Electronic Commerce. If we summarize agent’s special quality, they are as following [3][4][5]:

1) Personalized

2) Adaptive

3) Cooperative

4) Autonomous

5) Reactivity

Mobile agents can be defined as a “computing paradigm in which a program, in the form of a software agent, can suspend its execution on a host computer, transfer itself to another agent-enabled host on the network, and resume execution on the new host.”[6]

Current mobile agent systems are based on agent architectures that are partially or fully implementation programming language-specific. Mobile agent implementation in a specific programming language has usage limitations as the inter agent communication and agent migration to other agent hosts needs support for the same language.

Mobile agent systems provide the benefits of agent migration, multi-agent communication and negotiation. Such characteristics are well suited to mobile enterprise environments where coordinated distributed tasks can be carried out without a continuous wireless connection with mobile devices in the system. This paradigm can also help software development and deployment speed and flexibility in fields like e-commerce, network management and mobile computing [7]. However the lack of a standardized programming model prevents the widespread deployment of this potentially useful technology [8]. Programming language specific mobile agent systems presume a common execution environment. The basic features of mobile agent are mobility, whose purposes are to reduce network traffic and implement asynchronous interaction.

2 The Architecture of Mobile E-Commerce Based-on Mobile Agent

In order to help the application and technologies handle mobile Electronic Commerce, we propose the architecture as shown in Figure1. This architecture will help user to strategize and effectively implement Mobile Electronic Commerce applications. By using following architecture, the entity or individual is not force to do everything to build Mobile Electronic Commerce system, rather, they can build on the functionalities provide by others, they can build on the functionalities provided by others. This will help the development of Mobile Electronic Commerce applications as designers and developers can assume certain functions provided by lower layers and need not focus on capabilities and constraints of individual devices and networks.

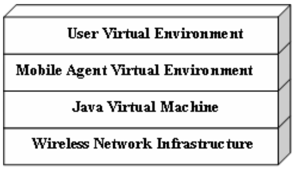

User Virtual Environment
Mobile Agent Virtual Environment
Java Virtual Machine
Wireless Network Infrastructure

Figure 1 Architecture of Mobile Electronic Commerce based-on Mobile Agent

The architecture (Figure 1) has four layers: Wireless Network Infrastructure (WNI), Java Virtual Machine (JVM), Mobile Agent Virtual Environment (MAVE), User Virtual Environment (UVE).

2.1 Layer 1: wireless network infrastructure

Networking support from wireless network is crucial in realizing mobile commerce applications. There have been significant advances in wireless and mobile networks in the last few years in terms of protocols, standards, technologies.

Here we discuss three technologies that play important role in Mobile Electronic Commerce. SMS (Short Messaging Service) is being offered by GSM networks allows users to transmit short messages to other users of GSM. It is a very important part of Mobile Electronic Commerce, especially for transactions involving short messages that can tolerate few seconds of delay. Bluetooth is a low power wireless standard for allowing a short range communications among multiple devices. It has become a global specification supported by many enterprises and can be widely deployed for conducting short range Mobile Electronic Commerce transactions between devices in a short distance. One major advantage of Bluetooth is that it allows the design of low power, small sized, and low cost radios that could be fit in many handheld devices. GPRS (General Packet Radio System) is a service by GSM and the expected bit rates art 160 Kbps. GPRS supports quality of service and dynamic IP address allocation. Due to its packet switched nature, it is suited to Mobile Electronic Commerce transactions, which may occur infrequently. Due to GSM presence in most of countries, GPRS will be made available to a large number of customers. [9]

2.2 Layer 2: java virtual machine (JVM)

JVM can be regards as "virtual machine" which can execute Java code. JVM is implemented by the software, and is constructed on the real hardware. The most important function of JVM is that it realizes the platform independence, as long as we transplant the interpreter to specific computer according to the JVM specification, we can guarantee that any java code which has been compiled can be executed on this computer. So the complied Java code can be executed on the local computer, or be transferred through the Internet and executed on other different types of computers. So, it is basic platform for mobile agent.

2.3 Layer3: mobile agent virtual environment

Mobile Agent Virtual Environment provides the main functions of Mobile Agent, this include Agent addressing, tracking, migration, registration, messaging and so on. The Object Management Group (OMG) Mobile Agent System Interoperability Facility (MASIF) (2000) defines the four main concerns about agent interoperability: A standard way to create, suspend, resume and termination; A common mobility infrastructure to support multi-agent communication and migration to other mobile systems; Standardized mechanism for naming and addressing agents and agent systems; Standardized location syntax for finding agents.

The use of a XML-based agent representation and UDDI provides ready ways to achieve the above. In our proposal each agent offers a Web service interface itself. That is, it has a set of operations that can be called (collectively a Web service(s)) for which there is a corresponding Web Services Description Language (WSDL) file. This service is how external entities communicate with the agent. The life-cycle of the system execution is described below to explain further how this will work (Figure 2).

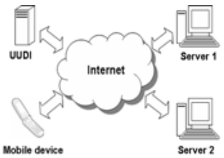


Figure 2 Structure of the system

We assume an underlying computing infrastructure where some of the computing hosts are mobile devices with wireless connectivity. In this scenario the mobile device end-user initiates a request and an agent migrates from the mobile device to the service provider server. Each server in the system hosts an agent container. Once the agent migrates to the server the connection with the mobile device is terminated. From this point the agent itself communicates with other agents and makes Web Service calls to gather and prepare the required information or perform the requested tasks. At the end of its tasks, the gathered information or result of the assigned task is sent back to this originating server.

Fundamentally agents and containers will be addressed via a URL. In this way, a unified, universally used and already deployed addressing mechanisms is utilized. Each agent, to achieve unique identification, will be assigned a unique ID by the container when it migrates from the mobile device and this ID will be incorporated into the unique URL that an agent will have at any point of time in its lifecycle. The mobile agent then registers itself with the UDDI registry and makes an entry there, which represents its current location on that host. This registration will involve publishing to UDDI the WSDL file that describes the set of operations the agent provides. The unique ID assigned to the agent is used as the 'service name' in the UDDI registry. This enables other agents to identify and discover the agent through the service name. Each time the agent migrates to another host this UDDI entry gets updated so the next messages sent to this agent get delivered to its current host/ location. When an agent wants to communicate with another agent it can invoke an operation or operations on that agent. Such a call might typically be preceded by a UDDI lookup to determine the target agent's current address/URL.

Assigning a unique identification to agents and making is available for other agents and containers can make it possible to find and negotiate with agents easily. The user identification part of the unique ID can be used to send back the results to the mobile device where the agent was initiated. Containers also provide services that can be addressed with a unique URL to allow agents to communicate with containers.

Usage of UDDI to store the addresses of agents and containers will help in making the system simpler due to widespread support and usage by industry. It will also be useful for inter-agent messaging to make sure that the message is delivered at the correct place even in the event of an agent migrating.

2.4 Layer 4: user virtual environment

Mobile Electronic Commerce provides the interface between User and Mobile Electronic Commerce. Currently, Mobile Electronic Commerce provides a variety of customized services the main services including Wireless advertisement, Mobile Bank, Financial transaction, Small payment and so on.

3 The Key Technologies

Some key technologies are necessary to the Mobile Agent System, such as the mobility, tracking, registration, migration and so on.

3.1 Mobility

The most important issue of the mobile agents is their mobility. There are two basic models of migration: the weak and the strong migration. The weak migration is transfer of only the agent's code and data. The agent restarts on the new host from the beginning but with its data. The agent must prepare for the transfer so that all the necessary information is in the data. The strong migration transfers agent's state, and the agent restarts from the point where it stopped. The weak migration is commonly used in agent systems today, since strong one can be difficult to implement into the Java environment or costly in performance.

3.2 Tracking

An agent track scheme is required to enable remote inter-agent interaction in the Mobile Electronic Commerce context. In our tracking scheme, the entire agent environment is subdivided into regions as outlined in Mobile Agent System Interoperability Facilities Specification. In each region, there is one agent server acting as agent name server for that region, which contains an agent registry. Tracking agents in a distributed environment includes three separate phases;

3.3 Registration

When it is created, an agent registers its name and the server on which it was created (home server) in the registry of the region (home region).

3.4 Migration

If the agent migrations within the same region, it only updates its location with the registry of its current region using the new server address. If the agent migrates to a different region, it needs to update its location with the registry of its home region using the new region address, as well as the registry of its current region using the new server address.

3.5 Locating

This is a two-step tracking. When an agent is tracked, the registry of its home region is contacted first. It contains the current region address. The name of home region can be extracted from the agent name. The registry of current region is then contacted to obtain the address of the server the agent is on.

4 Built-in Problem of the Mobile Agent-Based Mobile E-Commerce

In Mobile Electronic Commerce environment, different application domains have their own semantic models. How can one ensure that sellers and buyers have the same understanding on the essential issues in business process? The solution is ontologies. The ontologies play an important role in e-commerce in that they offer solutions to the integration of heterogeneous and distributed information sources about a knowledge domain. Ontology is a representation of knowledge about the chosen domain. The construction of ontology is complex in that it involves individuals, organizations and even geographic considerations.

5 Conclusion

In this paper, we discuss several issues that is important for Mobile E-Commerce. We have introduced the architecture of Mobile E-Commerce base on Mobile agent and the method for solving the problem that exit in the Mobile agent based Mobile E-Commerce and some key technologies that are necessary in this field. However, there are many technical and non-technical hurdles that need to be overcome before this technology excerts its effectiveness on the field of E-Commerce. We believe, our architecture will allow interoperability of Mobile E-Commerce applications and products from different providers. This would help in the adoption of Mobile E-Commerce on a global scale.

References

[1] Agrawal, R. and Chrysanthis, P. K. , “Efficient data dissemination to mobile clients in e-commerce applications. Advanced Issues of E-Commerce and Web-Based Information Systems,” WECWIS 2001, Third International Workshop, June 2001, pp. 58-65

[2] Lyytinen, K.:M-commerce-mobile commerce: “a new frontier for E-business. System Sciences,” Proceedings of the 34th Annual Hawaii International Conference, Jan 2001, pp. 3509-3509

[3] James Hendler, “Agent and the Semantic Web,” University of Maryland

[4] M.N.Huhns, “Agents as Web Services,” IEEE Internet Computing, Vol. 6, No. 4, Jul-Aug. 2002, pp. 93-95

[5] Pattie Maes, Robert H. Guttman, Alexandros G. Moukas, Agents that Buy and Sell: Transforming Commerce as we Know It.”

[6] Jansen W., “Mobile Agent Security,” NIST Special Publication 800-19, Nov. 2004

[7] Steele, R. , “A Web Services-based System for Adhoc Mobile Application Integration,” IEEE Intl. Conf. on Information Technology: Coding and Computing ’03, 2003

[8] Schoeman M. and Cloete E., “Architectural Components for the Efficient Design of Mobile Agent System,” ACM International Conference Proceeding Series, 2003, pp. 48-58

[9] Upkar Varshney and Ron Vetter, “A Framework for the Emerging Mobile Commerce Applications,” Proceedings of the 34th Hawaii International Conference on System Sciences, 2001

[10] Robert Steel, Tharam Dillon, Path Pandya, Yuri Vensov., “XML-based Mobile Agent, the International Conference on Information Technology: Coding and Computing”

A New Algorithm for All-pairs Shortest Paths of Given Source-destination Pair

Tianzhi Li   Fengsheng Xu

Department of Computer Science and Technology, Dezhou University, Dezhou, Shandong, 253023, China
Email: ltz@dzu.edu.cn

Abstract

A new algorithm to find all the shortest paths from a specified source vertex to each other vertex is proposed in this paper through analyzing the most efficient algorithms - Dijkstra's Algorithm. The data structure used in this algorithm is simple and realizes easily .The running time of this algorithm is $O(n^2)$ as well as Dijkstra's Algorithm.

Keywords: shortest path, Dijkstra's algorithm, predecessors list, successors list, contrary predecessors list, graph

1 Introduction

The development, computational testing, and efficient implementation of shortest path algorithms have remained important research topics within related disciplines such as operations research, management science, geography, transportation, and computer science. These research efforts have produced a number of shortest path algorithms as well as extensive empirical findings regarding the computational performance of the algorithms [1-10].

Dijkstra's Algorithsm [1], introduced in 1959 provides one the most efficient algorithms for solving the shortest-path problem. The algorithm maintains a tentative cost d_v for each vertex v, such that some path from s to v has total cost d_v. As the algorithm proceeds, the tentative costs decrease, until at the termination of the algorithm, for each vertex v, d_v is the cost of a minimum-cost path from s to v. Initially $d_v = 0$ and $d_v = \infty$ for every $v \neq s$ .The algorithm maintains a partition of the vertices into two states: unlabeled vertices, those with infinite or finite tentative costs whose minimum cost is not yet known; labeled vertices, those whose minimum cost is known. Initially, s is labeled and all other vertices are unlabeled. The algorithm maintains a predecessor P_v of vertex v on the shortest path. The algorithm consists of repeating the following step until all vertices are scanned:

Scan a Vertex. Select an unlabeled vertex v such that d_v is minimum and declare v labeled. For each $vw \in E$, if $d_v + W(v,w) < d_w$, set $d_w = d_v + W(v,w)$.

But Dijkstra's algorithm only finds one of the shortest paths between a specified source vertex and all other vertex in a weighted directed or undirected graph.

In practice, one may wish to find all shortest paths between a specified source node and a destination node. How do we find all the shortest paths? Many papers study algorithms for finding all the shortest paths connecting a given source-destination pair [2, 3 and 4].But these algorithms are perplexing in data structure and algorithm describing.

A new improved algorithm is proposed in this paper based on the analysis of all the probability of edges contributed to the construction of the new shortest paths. The proposed algorithm not only finds all the shortest paths connecting a given source-destination pair, but also reduces the computational complexity required to construct the shortest paths.

2 Background

Definition 1 (Path in Graph) Given a directed weighted graph $G = (V, E, W)$, a path p from a vertex x to a vertex y in G is a sequence of edges $xv_1, v_1v_2, \ldots, v_{k-1}y$ (possibly $v_1 = y$). The rank of path is the number of edges in the path p. The weight of path p is defined to be $W(p) = \sum_{e \in P} W(e)$.

Definition 2 (Shortest Path) The shortest path from x to y is a path of minimum weight from x to y.The shortest path from x to y is defined as $\delta(\mathrm{x,y}) = \min\{\mathrm{W(p): p\ is\ a\ path\ from\ x\ to\ y}\}$. If there is no path from x to y then $\delta(\mathrm{x,y}) = \infty$.

Definition 3 (Adjacency Matrix) Let an $n \times n$ weight matrix $W = a_{ij}$ represents a graph G. The weight matrix $W_{n \times n} = [W_{ij}]$ has the following entries:

Lemma 1 (Sub-paths of Shortest Paths Are Shortest Paths)

$$W_{ij} = \begin{cases} \mathrm{W(i,j)} & \text{if } \mathrm{v_i} v_j \in E \\ \infty & \text{otherwise} \end{cases} \quad \textit{for} \quad 1 \le \mathrm{i,j} \le \mathrm{n}$$

Let P be a shortest path from vertex s to vertex v in a graph $G = (V, E)$, and let P_{xy} be a sub-path of P from vertex x to vertex y. Then, P_{xy} is a shortest path from x to y.

Lemma 2 For all $x, y, u \in V$, we have $\delta(x,y) \le \delta(x,u) + \delta(u,y)$.

3 Data Structure and Algorithm

The algorithm in this paper borrows from the following properties of Dijkstra's algorithm: k_v, d_v, P_v.

k_v, the bool-valued flag which indicates that the shortest path to vertex v is labeled. Initially, $k_v = \mathit{false}$ for all $v \in V$.

d_v, the length of the shortest known path from source node s to v. When the algorithm begins, no shortest paths are known. The distance d_v is a tentative distance. During the course of the algorithm candidate paths are examined and the tentative distances are modified.

Initially, $d_v = \infty$ for all $v \in V$ such that $s \ne v$, while $d_s = 0$.

P_v, the set representing the predecessors of vertex v on the shortest path from s to v.

Initially, $\mathrm{p_v} = \varnothing$ for all $v \in V$.

This algorithm proceeds in phases as well as Dijkstra's.

Algorithm 1 Generate Predecessors_list

```
Input:weighted graph G
Output: predecessors list P
Method:
1. Select a vertex v having minimum tentative distance and $k_v = false$.
2. Set $k_v = true$.
3.
for all $(v, w) \in E$ do
begin
if $k_w = false$ and $d_v + W(v,w) < d_w$ then
begin
$d_w = d_v + W(v,w)$
$P_w = \varnothing$
$P_w = \{v\}$
End
if $k_w = false$ and $d_v + W(v,w) = d_w$ then
begin
$P_w = P_w \cup \{v\}$
end
end
```

In each pass exactly one vertex has its k_v set to true. The algorithm terminates after $|V|$ passes are completed.

In this algorithm, the length of the predecessors set is uncertainty, the memory maybe needed which can later be released and then allocated again. So we use dynamic memory allocation - linked List to represent the predecessors set. The structure for predecessors list is defined as following:

```
typedef struct ArcNode{
        int adjvex;      // index of vertice
      struct ArcNode *next;     //next parent on the
                                          //predecessors list
   } ArcNode;

typedef struct Vnode {
        VertexType data;         // vertex info
ArcNode *firstparent; // first parent on the
                                   //predecessors list
}ParentList[N];   // N is the number Overtices
```

This algorithm generates a predecessors list. See Fig# 1 and Fig# 2 for examples.

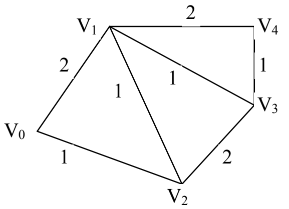


Figure 1 An undirected weighted graph

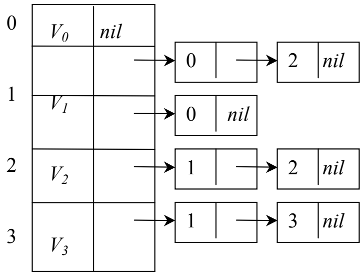


Figure 2 Predecessors list p for Figure 1. Source node is V_0

According predecessors list *P*, we can get all the shortest path with Backtracking Mechanism. To get each shortest path we should traverse all the vertexes on the path. However, such work is often difficult and time consuming. If we convert the predecessors list *P* into a successors list *S* (contrary predecessors list), this problem would become easily to solve.

The successors_list structure is the same with the predecessors_list's, just replace the attribute “firstparent ” in Vnode with “firstchild”.

Algorithm2 Predecessors_list to Successors_list

```
Input: predecessors list P
Output: successors list S
Method:
1. Initially, S[v]-> firstchild =nil   for all $v \in V$.
2. for i=0 to |V|-1 do
begin
   t=P[i]->firstparent
   while t   do
      begin
         j=t-> adjvex;
         q=new  (ArcNode)     //allocate memory for
                              //node q
         q->adjvex=i
         q->next=S[j]->firstchild
         S[j]->firstchild=q   //insert q into successor
                              //list S[j]
      t=t->next
   end
end
```

See Fig# 3 for example.

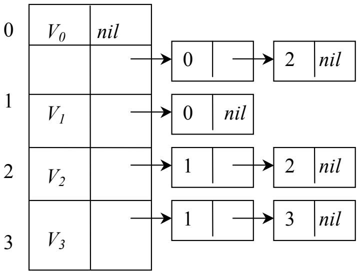


Figure 3 Successors list *S* for Fig#1.Source node is V_0

According to the successors list, all the shortest paths between a specified source node and a destination node can be computed easily. If we have computed all the shortest paths P_y from x to y, then partial shortest paths from x to every node u on the successors list of y can be computed by: $p_u = p_y \cup y$, for every $p_y \in P_y$.The solving process is a directed graph with no directed cycles.

For example, according to Fig#3, all the shortest paths from v_0 to each other nodes in Fig#1 are shown order by rank as following: $V_0 V_1$, $V_0 V_2$, $V_0 V_1 V_3$, $V_0 V_1 V_4$, $V_0 V_2 V_1$, $V_0 V_2 V_3$, $V_0 V_1 V_3 V_4$, $V_0 V_2 V_1 V_3$, $V_0 V_2 V_1 V_4$, $V_0 V_1 V_3 V_4$, $V_0 V_2 V_1 V_3 V_4$.The solving process is illustrated in Fig#4.

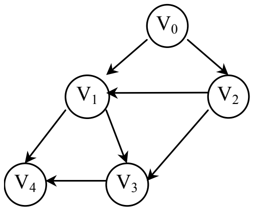


Figure 4 Solving process for Fig#1.Source node is V_0

4 Analysis

The running time of this algorithm include the time of creating predecessors list , the time of generating successor list and the time prints shortest paths. The procedure of creating predecessors list runs in $O(n^2)$ [1].The time of generating successor list is impacted by the length of the predecessors list, but the Time complexity of it is no more than $O(n^2)$ .The time of prints shortest paths is determined by the sum of all the shortest paths. So the total running time of this algorithm is $O(n^2)$.

5 Conclusions

In this paper, a new efficient algorithm has been presented for finding all the shortest path from a source vertex s to each other vertex v .The new algorithm is simple and efficient which runs in $O(n^2)$ time as well as Dijkstra's algorithm in the worst case. This algorithm is suitable for weighted directed and undirected graph.

References

[1] Sara Baase, Allen Van Gelder, "Computer Algorithms: Introduction to Design and Analysis, 3rd", Beijing: Higher Education Press, pub, 2001

[2] Fengsheng Xu, Tianzhi Li, "A New Algorithm for Finding All the Shortest Path", Computer Engineering & Science, Vol. 28,No.12, Dec 2006,pp.83-84

[3] Fengsheng Xu , "The New Algorithm for Finding the Shortest Path", Computer Engineering & Science, Vol. 28,No.2, Feb 2006, pp.84-85

[4] Fengsheng Xu, "Algorithm for Finding the Shortest Path", Computer Applications, Vol. 24, No.5, Feb 2004, pp.88-89

[5] Guangzheng Long, Jianju Yang, "Improved Algorithm of Short-Cu", Systems Engineering and Electronic, Vol. 24, No.6, Feb 2002, pp.106-108

[6] D. Frigioni, A. Marchetti-Spaccamela, and U. Nanni, "Fully dynamic output bounded single source shortest path problem," in Proc. 7th Annu. ACM-SIAM Symp. Discrete Algorithms, (Atlanta, GA), pp. 212-221, 1998

[7] F. Benjamin Zhan, " A Comparison Between Label-Setting and Label-Correcting Algorithms for Computing One-to-One Shortest Paths", Journal of Geographic Information and Decision Analysis, Vol. 4,No.2, Feb 2000,pp.1-11

[8] Noga Alon, Zvi Galil, Oded Margalit, "On the Exponent of the All Pairs Shortest Path Problem", Journal of Computer and System Sciences , Vol. 54,No.2,April 1997 ,pp.255-262

[9] Uri Zwick, "All pairs shortest paths using bridging sets and rectangular matrix multiplication", Journal of the ACM (JACM), Vol.49, No.3, May 2002, pp.289-317

[10] T. Takaoka, "A faster algorithm for the all-pairs shortest path problem and its application",In Proc. 10th Int. Conf. Comput. Comb., Lect. Notes Comput. Sci., 2004, pp. 278-289

[11] Uri Zwick, "All Pairs Shortest Paths in weighted directed graphs exact and almost exact algorithms", Foundations of Computer Science, 1998. Proceedings.39th Annual Symposium on Volume, Issue, 8-11, Nov 1998, pp.310 - 319

The Research of the Parallel Section Pursant Method to Solve Trinal-Angles Linear Equations

Xinghua Ma[1] Aimin Yang [*1] Chunfeng Liu[1] Yibin Zhu[2] Yanbing Liang[1]

1 College of Science, Hebei Polytechnic University, Tangshan, Hebei, 063000 China

2 Tangshan Central Radio and TV University, Tangshan, Hebei, 063000 China
Email:aimin_heut@163.com

Abstract

With the rapid development of high-speed network technology, the cluster systems have been the main platform of parallel algorithm. Because of the delay of their high communication, some parallel algorithms of fine grain are not fit to run in this environment. Therefore, it is necessary to study their parallel achievements in cluster systems. In terms of that, through analyzing the main influential characters of algorithmic parallel efficiency, the parallel pursuant method to solve trinal-angles linear system equations is raised on the basis of the separation theory of the segmentation strategy. And then when being put into practice based on MPI in cluster system to compare speed-up ratio and efficiency of parallel method with original serial method, this result shows that the method has higher calculation efficiency.

Keywords: Trinal-Angles Linear System Equations, section parallel pursuant method, cluster system

1 Introduction

With the rapid development of the technology of the network, cluster system has become a main platform cluster system for parallel algorithm, which using high-speed universal network to dispatch a group of high-performance working stations or PCs integrally, assigned relevant supporting software, such as MPI, PVM, etc., constitutes a high efficient parallel processing system. Although the high-speed network has shortened the communication delay greatly, in fact it still influences the efficiency of the parallel algorithm. So the algorithm used in the cluster system only applies to the parallel of medium grain and above, which makes it necessary to design coarse grain parallel algorithm suitable to the network parallel. On the basis of the above-mentioned point of view, this paper provides a research of the parallel pursuant method in the cluster system in MPI according to the method of measurement of communication expense raised by B.K. Schmidt (Reference [1,2,]).

In actual computing, the solution of most basic number questions is reduced to solve trinal-angles linear system equations at last. Such as: the computing of the ter-spline function (Reference [3-5]) and the computing of the questions of border value in frequently differential equation(Reference[6]). While parallel pursuant method is the most wealthy appeal method to solve trinal-angles linear system equation(Reference [7,8]), in this paper, the branch pursuant method must be retrospected firstly, then obtained the parallel pursuant method (Reference [9-11])to solve the trinal-angles linear system equations, and the testing outcome is given lastly.

2 Doolittle Decomposition

First we describe the Doolittle decomposition of normal matrix, suppose:

$$A=\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ & \cdots & & \cdots & \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix}=LU$$

$$=\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ l_{31} & l_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & l_{n3} & \cdots & 1 \end{bmatrix}\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{22} & \cdots & u_{1n} \\ & & \ddots & \\ & 0 & & u_{nn} \end{bmatrix}$$

through the multiplication of matrix and matrix, we can obtain

$$\begin{cases} u_{1j}=a_{1j} & (j=1,2,\cdots,n) \\ l_{i1}=a_{i1}/u_{11}, & (i=2,3,\cdots,n) \\ u_{ij}=a_{ij}-\sum\limits_{k=1}^{i-1} l_{ik}u_{kj} & (i=2,\cdots,n,j=i,\cdots,n) \\ l_{ij}=(a_{ij}-\sum\limits_{k=1}^{j-1} l_{ik}u_{kj})/u_{jj} & (j=1,2\cdots,n,i=j+1,\cdots,n) \end{cases}$$

Suppose $Ax=b$, $A=LU$, solving $Ax=b$ is equal in value to solve $Ly=b,Ux=y$, because

$$Ly=\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_{21} & 1 & 0 & \cdots & 0 \\ & \cdots & & \cdots & \\ l_{n1} & l_{n2} & l_{n3} & \cdots & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}=\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

we can obtain firstly

$$y_k=\begin{cases} b_1 & (k=1) \\ b_k-\sum\limits_{j=1}^{k-1} l_{kj}y_j & (k=2,\cdots,n) \end{cases}$$

then

$$Ux=\begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ & \cdots & & \cdots & \\ 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

we can obtain lastly

$$x_k=\begin{cases} y_n/\mathrm{u_{nn}} & (k=n) \\ (y_k-\sum\limits_{j=k+1}^{n} u_{kj}x_j)/\mathrm{u_{kk}} & (k=n-1,\cdots,1) \end{cases}$$

3 The Branch Pursuant Method

Suppose the trinal-angles linear system equations to solve is

$$\begin{cases} a_1x_1+ & b_1x_2 & & & = & r_1 \\ c_2x_1+ & a_2x_2+ & b_2x_3 & & = & x_2 \\ & \cdots & \cdots & \cdots & & \\ & c_{n-1}x_{n-1}+ & a_{n-1}x_{n-1}+ & b_{n-1}x_n & = & r_{n-1} \\ & & c_nx_{n-1}+ & a_nx_n & = & r_n \end{cases}$$

$$\begin{bmatrix} a_1 & b_1 & & & \\ c_2 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & c_n & a_n \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}=\begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{n-1} \\ r_n \end{bmatrix}$$

$$\boldsymbol{Ax=r} \qquad (1)$$

where

$$A=\begin{bmatrix} a_1 & b_1 & & & \\ c_1 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & c_n & a_n \end{bmatrix}$$

$$x=(x_1,x_2,\cdots,x_n)^T\,, \quad r=(r_1,r_2,\cdots r_n)^T$$

and when $|i-j|>1$, $c_{ij}=0$, at the same time

$$\begin{cases} |a_1|>|b_1|>0 \\ |a_i|\geq|b_i|+|c_i| \\ |a_n|>|c_n|>0 \end{cases} \quad b_ic_i\neq 0(i=2,3,\cdots,n-1) \ \triangleleft$$

The formula of branch pursuant method to solve the Eq.(1) is:

(1) Doolittle decomposition

$$A=\begin{bmatrix} a_1 & b_1 & & & \\ c_1 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-1} & a_{n-1} & b_{n-1} \\ & & & c_n & a_n \end{bmatrix}$$

$$
=\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_2 & 1 & 0 & \cdots & 0 \\ 0 & l_3 & 1 & \cdots & 0 \\ & & & \cdots & \\ 0 & 0 & 0 & l_n & 1 \end{bmatrix}
\begin{bmatrix} u_1 & v_1 & & & \\ & u_2 & v_2 & & 0 \\ & & \ddots & \ddots & \\ & & & \ddots & \ddots \\ & 0 & & u_{n-1} & v_{n-1} \\ & & & & u_n \end{bmatrix}
$$

obviously:

$$
\begin{cases} a_1 = u_1 \ , v_i = b_i \\ c_i = l_i u_{i-1} \qquad (i = 2,3,\cdots,n) \\ a_i = v_{i-1} l_i + u_i \end{cases}
$$

If $u_{ii} \neq 0 (i = 1,2,\cdots,n-1)$, then

$$
\begin{cases} u_1 = a_1 \\ l_i = c_i / u_{i-1} \qquad (i = 2,3,\cdots,n) \\ u_{\mathrm{i}} = a_i - b_{i-1} l_i \\ v_{\mathrm{i}} = b_{\mathrm{i}} \end{cases}
$$

$$
A=\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ l_2 & 1 & 0 & \cdots & 0 \\ 0 & l_3 & 1 & \cdots & 0 \\ & & & \cdots & \\ 0 & 0 & 0 & l_n & 1 \end{bmatrix}
\begin{bmatrix} u_1 & b_1 & & & \\ & u_2 & b_2 & & 0 \\ & & \ddots & \ddots & \\ & & & \ddots & \ddots \\ & 0 & & u_{n-1} & b_{n-1} \\ & & & & u_n \end{bmatrix}
$$

(2) Solve the linear equations

Suppose $Ly = r$ and $Ux = y$, then

$$
\begin{cases} y_1 = r_1 \\ y_k = r_k - l_k y_{k-1} \end{cases} (k = 2,3,\cdots,n)
$$

$$
\begin{cases} x_n = y_n / u_n \\ x_k = (y_k - b_k x_{k+1}) / u_k \end{cases} (k = n-1, n-2,\cdots,1)
$$

reduce the two equations above, then

$$
\begin{cases} p_j = c_j q_{j-1} + a_j & (q_0 \equiv 0, j = 1,2,\cdots,n) \\ q_j = -b_j / p_j & (j = 1,2,\cdots,n-1) \\ u_j = {(r_j - c_j u_{j-1})}\Big/{p_j} & (u_0 \equiv 0, j = 1,2,\cdots,n) \end{cases} \qquad (2)
$$

$$
\begin{cases} x_n = u_n \\ x_j = q_j x_{j+1} + u_j \end{cases} (j = n-1,\cdots,1) \qquad (3)
$$

Eq.(2) is named “pursue” computing process, Eq.(3) is named “drive” computing process.

4 The Parallel Section Method

By the medium of analysis, the mainly calculation of the pursuant method includes the calculation of inner product for vectors, the calculation of matrix timing vector, the calculation of matrix timing matrix and etc. For the large linear problem, it is necessary to calculate using parallel methods of these segments in pursuant method. In the process of designing these parallel methods, we elementary base on the principle of separately, divide the original matrix or vector into some blocks, then distribute each block into various node machines, which will run the submission in dependently (Please read the reference [1]). It is a much better proposal to the cluster system which has no more nodes.

The key point of the parallel computing is the execution of the divide and rule strategy, the complexity question is decomposed a number of independent computing units which are more small, irrelevant one another and can independent compute. But the branch pursuant method above is linear pass-push method, which only can pursue and drive one by one, the conclusion is that the branch method has no parallelism. So the key point to design the parallel method is how to split, the solving way is leading into a inverse pursuant computing method.

$$
\begin{cases} P_i = b_j Q_{j+1} + a_j \\ \qquad (Q_{n+1} \equiv 0, j = n, n-1,\cdots,1) \\ Q_j = -c_j / Pj \\ \qquad (j = n, n-1,\cdots,2) \\ U_j = (r_j - b_j U_{j+1}) / P_j \\ \qquad (u_{n+1} \equiv 0, j = n, n-1,\cdots 1) \end{cases} \qquad (4)
$$

$$
\begin{cases} x_1 = U_1 \\ x_{j+1} = Q_{j+1} x_j + u_{j+1} (j = 1,2,\cdots,n-1) \end{cases} \qquad (5)
$$

We can build the section computing formula of pursuant method from the thought of develop the above two pursuant methods.

The destination to lead into a inverse pursuant computing method is pursued from the tail of the equations. So if we suppose the linear systems equations is divided two sections, and two CPU are given in cluster system, then we can pursue at the same time from head and tail. The first CPU gives effect to the Eq.(2)(there is $j=1,2,\cdots,[n/2]$) in the first kind of formulas, and the second CPU gives effect to the Eq.(4) (there is $j=n,n-1,\cdots,[n/2]+1$) in the second kind of the formulas. That is the parallel pursue computing process.

When pursuing to the middle of the equations, the result is obtained of the above parallel computing, then we can obtained a 2×2 linear algebraic equations:

$$\begin{cases} x_{\lceil n/2\rceil} = q_{\lceil n/2\rceil} x_{\lceil n/2\rceil+1} + u_{\lceil n/2\rceil} \\ x_{\lceil n/2\rceil+1} = Q_{\lceil n/2\rceil+1} x_{\lceil n/2\rceil} + U_{\lceil n/2\rceil+1} \end{cases} \tag{6}$$

$$\begin{cases} x_{\lceil n/2\rceil} = \dfrac{q_{\lceil n/2\rceil]} U_{\lceil n/2\rceil+1} + u_{\lceil n/2\rceil}}{1 - Q_{[\lceil n/2\rceil+1} q_{\lceil n/2\rceil}} \\ x_{[n/2]} = \dfrac{Q_{[\lceil n/2\rceil+1} u_{\lceil n/2\rceil]} + U_{\lceil n/2\rceil+1}}{1 - Q_{\lceil n/2\rceil+1} q_{\lceil n/2\rceil}} \end{cases} \tag{7}$$

So the $x_{\lceil n/2\rceil}$ and $x_{\lceil n/2\rceil+1}$ can compute at the same time. Then the two CPU drive at the same time form middle to both ends. The first CPU gives effect to the "drive" Eq.(3) (there is $j=\lceil n/2\rceil-1,\cdots,2,1$), and at the same time the second CPU gives effect to the "drive" Eq.(5) (there is $j=\lceil n/2\rceil+2,\cdots,n-1,n$). So we can obtain the solution $x_1,x_2,\cdots,x_n$ of the trinal-angles linear system equations.

Obviously, when the communication overhead isn't attention in the factual computing, the complex degree of the parallel section pursuant method is half size of the branch pursuant method through analyzing from theory, and the complex degree is O(*n/2*). Of course, if the linear system equations are divided to more sections, and if the CPU is enough, the computing speed is more quickly, and the speed-up is more higher.

5 Simulation of the Algorithm

Use MPI to simulate the above method in the internet of 1000Mbps, choose the equity model configuration ,and realize it using MPI+Fortran. In the 8-hodes cluster system, we separately use 2 nodes ,4nocks and 8nodes to simulate the parallel section pursuant method, and compare it with serial runtime. In the cluster each node is p_4 2.6GHZ. Assure the child takes of the QR decomposition is separately 2, 4 and 8.

The outcome of experiment is as table 1. In the table n expresses the rank of the matrix, P expresses the number of CPU, K expresses the number of the divided assignment

we can see from the table that under the loom cluster environment, when p=2, the parallel algorithm gets a certain acceleration will also increase, but the acceleration ratio when k=4 is less than the acceleration ratio when k=8 .this is because when k increase, the date that transported also increase, so it causes the increasing of the communication. This is fit to the theoretical and practical analysis.

6 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.

Acknowledgements

This paper is supported by the Natual Scientific Fund Project of Hebei Province (E2004000245), Scientific Technology Research and Development Plan Project (07160219B-6), Scientific Fund Project of Hebei Polytechnic University (Z0808).

Table 1 The outcome of the parallel section pursuant methed solving the system of linear equations

n	Serial time (s)	K=2 P=2		
		Parallel time (s)	Acceleration ratio	Efficiency
200	25.36	14.30	1.77	0.89
400	47.25	26.25	1.80	0.90
600	69.89	38.83	1.80	0.90

N	Serial time (s)	K=4 P=4		
		Parallel time (s)	Acceleration ratio	Efficiency
200	25.36	10.08	2.52	0.63
400	47.25	18.17	2.60	0.65
600	69.89	26.47	2.64	0.66

n	Serial time (s)	K=8 P=8		
		Parallel time (s)	Acceleration ratio	Efficiency
200	25.36	10.00	1.62	0.28
400	47.25	16.22	1.90	0.49
600	69.89	20.18	2.04	0.47

References

[1] Illiams G Fox R, Messina P, Parallel Computing Works! Morgan Kaufman, 1994

[2] Chunfeng LIU, Aiming YANG, The Parallel Algorithm of QR Decomposition of Matrix in Cluster System World Academic Press, 2006:107-111

[3] Aimin YANG and Yiming CHEN, MPI Parallel Programming Environment and Programming Research, Academy Newspaper of Heibei Polytechnic University, 2005,3: 41-44

[4] Jianfei ZHANG and Hongdao JIANG, Direct Blocking parallel algorithm of large scale BF equation sets, Application Mechanics transaction 2003,20(4):129-133

[5] Yan ZHANG, Distributed Parallel Algorithm Designing, Analysis and Realization Doctor Thesis of Electronic Science and Technology University 2001

[6] R.W.Hockney. The Science of Computer Benchmarking. The Society for Industrial and Applied Mathematics, Philadelphia, 1996

[7] G.F.Psister, Clusters of Computers: Characteristics of an Invisible Architecture. IEEE Int'l, Parallel Processing Symp, Honolulu, 1996

[8] Saad Y, Schultz M H.,GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetrical Linear System[J]. SIAMJ Sci Comp, 1986,7(3):856-869

[9] Yiming CHEN, Network Parallel BEM for Band Precision Rolling, COMPUTATIONAL MECHANICS, WCCM VI in conjunction with APCOM'04, Beijing, China. Tsinghua University Press & Springer-Verlag, 2004,9: 5-10

[10] Zhihui DU, High-performance MPI programming Technology, Tinghua University Press, August of 2001

[11] W. Michacl. The Research and Development of Parallel Computation, Parallel Theory and Practice, 2000

Design and Implementation of Autonomic Computing System for Server Cluster

Wenjie Liu[1]   Yuntao Zhou[2]

1 Department of Software and Theory (School of Computer), Northwestern Polytechnical University, Xi’an, 710072, China

Email：liuwenjie@nwpu.edu.cn

2 Department of Lab and Device Management, Northwestern Polytechnical University, Xi’an, 710072, China

Email：zhouyuntao@nwpu.edu.cn

Abstract

Aiming at the problem that servers cluster cannot rapidly deploys system and maintenance cost is high, on the basis of studying the autonomic computing and analyzing the features of cluster system, this paper designed and implemented an autonomic computing system for servers cluster. The designed system can mask the complexity of hardware and auto-control the cluster system, which realized the software and hardware cooperative work. Agent technique is used to collect the cluster status information and report it to autonomic computing system automatically, which then realizes the system self- check and self-recovery. After describing the system structure and modules functions, the autonomic computing features are described, which can auto-deploying the system and recovering the fault, then reduce the cost and realize the system self-management.

Keywords：Servers cluster; Autonomic Computing; Multi-agent System; Cooperative work; Self-Management

1 Introduction

In the past, when enterprises users build their application system, there are only two kinds of architecture that can be selected. One is based on mainframe; the other is based on client/server cluster. The first kind of architecture has high performance, high flexibility and high availability, but it costs much in buying hardware devices, and there are many functions that are never used, therefore resources are wasted. The second kind of architecture allows users to add hardware devices according their needs, but this kind of system is not real cluster, and it lacks of necessary availability and manageability, which makes users pay much in application upgrading and management.

With the occurrence of network, a new architecture, which has high performance/price comparison, comes into being and becomes the mainstream - distributed cluster architecture. When users want to accomplish their tasks, this architecture provides more computing ability and transparent data access ability, at the same time, realizes the high performance and high reliability.

But one coin has two sides, there are many problems in this kind of structure, such as, after hardware fault occurs, the SPOF (Single Point of Fault) status lasts long, which will take long time to check the cluster system configuration, to find out the fault point, to install software patches or find the difference of hardware/software. So it cannot deal with the increased load. To solve the above problems, higher availability management and more effective hardware utility are needed.

IBM’s senior vice president, Paul Horn, proposed the concept of autonomic computing in March 2001. He noted that autonomic computing system must have four features: self-configuration, self-optimization, self-healing and self-protection. The aim of autonomic computing environment is to make IT system reach the level of RAS (High Reliability, High Availability, and High Serviceability) [1-2]. The core of the concept is to use software to auto-control the complex hardware system therefore reduce the management cost.

This paper takes the servers cluster as management object, designed and implemented an autonomic computing system. By using this system, servers cluster can configure itself, find the hardware fault and recover itself automatically, which realizes the autonomic features and accomplish the aim of self-management.

2 Cluster system analysis

Cluster system contains many homogeneous or heterogeneous computers, which are connected to accomplish specified tasks cooperatively. It provides high performance services continuously. Servers cluster are a group of independent servers, which can be regarded as a single server in the network, and can be managed as a single system. The single system provides the client station with high reliable service. A cluster system contains many servers that share data storage, each server communicates with others across inner LAN. When fault occurs in one of the servers, the application running in this server will be taken over by another server automatically. Severs cluster can provide quite high performance none-stopping service, because each server can take on part of the computing task. As cluster system owns the performance of many servers, the system computing ability will increase also. At the same time, when fault occurs in one of the servers, system can separate this server from others by using special software, realize the new load balance by the load shift mechanism among servers, and simultaneously notify the administrators by signals [3-4].

Practice proves that cooperative work in the cluster has much higher computing ability than mainframe, super computer and fault tolerance system, moreover owns the lower cost.

But there are also disadvantages in cluster system as following [5].

It cannot rapidly deploy software: For large-scale cluster system, to deploy software rapidly on each node of the system needs much time and human power.

It cannot rapidly shift roles: Application requires that each server can shift to another role in different time. For example, in data center, sometimes more web service is required; sometimes more video service is needed. This requires servers to shift rapidly between two services to meet the needs of different time.

Maintenance costs are high: With the increase of nodes count, the fault chances also increase, and system recovery costs become high. How to reduce the time of recovering system and maintenance costs becomes a problem in cluster system.

This paper designed and implemented an autonomic computing system for cluster system, which can solve the above problems and reduce the management cost. The designed system has the following features:

1) Rapidly deploy the system software;
2) Automatically recover system when fault occurs;
3) Optimize system automatically;
4) Protect system when illegal invasion occurs.

3 Design of autonomic computing system

At present, there are two kinds of ways to build cluster system. One is to connect the backup server to the main server, when the main server is failure, backup server will take over all the tasks. The other is to connect multiple servers together, all the servers work cooperatively to do the same task, therefore improve the response time of large -scale application. Also, each server takes on some fault tolerance task, when fault occurs in one server, system will separate that server from others and accomplish new load balance [6]. PC servers usually use two servers to build cluster system.

UNIX system often uses 8 servers to build cluster, and the OpenVMS of Compaq can support 96 servers cluster. The system we designed is aiming at the UNIX server's cluster, by adding a management server to the cluster system to manage the cluster, by adding agents to the servers of cluster to get the health status of servers and communicate with the management server, by using the Ignite-UX to realize the software deployment across network [7-8].

To realize the system self-management and autonomic features, we designed the following system architecture.

3.1 Architecture

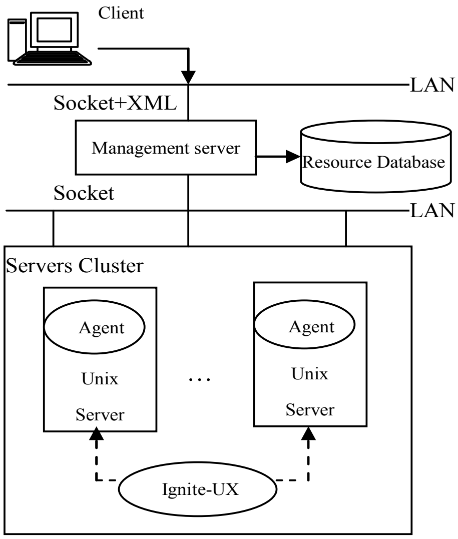


Figure1 Architecture of Autonomic System

The system is divided into three layers, client, management server and servers cluster. The three layers communicate with each other by LAN. Client communicates with management server by socket and XML; management server communicates with cluster system by socket. To communicate and operate the servers in the cluster, Ignite software must be installed on each UNIX server. Ignite-UX uses "pull" and "push" to deploy OS software across network, and can rapidly deploy system in the first time, or copy this configuration to other systems across network. This ability can save the time of the administrators therefore reduce the cost. By this way, rapid cluster system deployment can be achieved. Moreover, an agent is installed on each server to communicate with management server, deal the request and return the result. It is actually a background daemon process, by which management server can get the servers status and send request. It is the core of the autonomic computing system.

Client sends varied requests to management server, such as OS installation, software/patches installation, system backup, system recovery, system hardware status inquiry and so on.

Management server receives the requests and executes different operations according to request type. If the request is to query resource information, management server will get the information from resource database, and return the results with XML format. If the request is to install software on cluster servers, management server will firstly divide the servers of cluster into object servers and Ignite-UX server. For example, 8 nodes cluster can be divided into 7 object servers and one Ignite-UX server. Then management server will use the image file on the Ignite-UX server to deploy the object servers simultaneously. After the installation finished, results will be returned to client and resource database will be updated.

Resource database stores the OS type and version of each node of the cluster, the software and patches version installed in each node, current OS status and disk storage size, etc. The information in the database will be dynamically updated according to the node status.

3.2 Modules Constitution

In the autonomic computing system, the core is the management server. It is the brain of the whole system. It receives the client request, parses the XML data and sends commands to cluster servers, updates the resource database information according to the server status. The realization of this part contains the following modules.

1) Service Control

2) Communication Control

3) Resource Control

4) Events Control

5) Monitor Control

Moreover, agent is also the core module in the system. It is installed on UNIX server, which is to receive commands from management server, execute the commands and return the results.

The autonomic computing system modules constitution is as following:

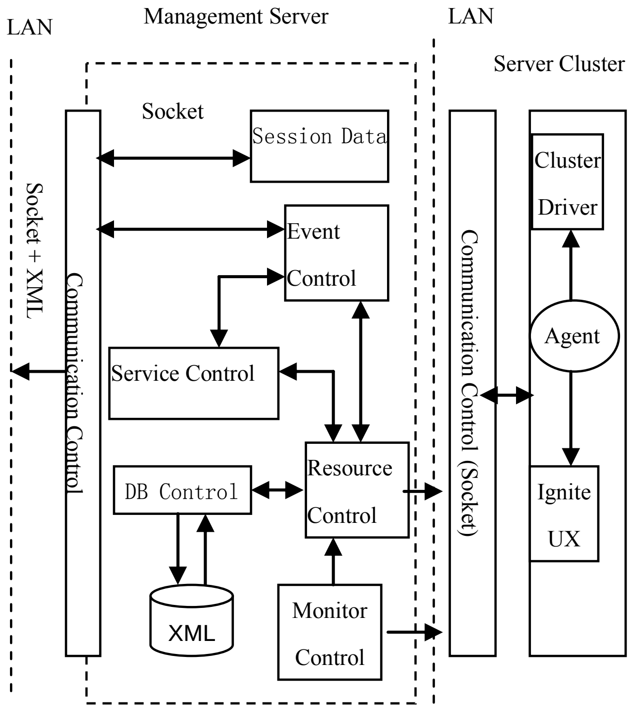


Figure 2 Modules Constitution

Service Control: Service control is the scheduling center, which is monitoring the events from events control module all the time. If new event is received, this module will extract XML data from the event, parse the information and judge the request type. If the request type is to query resource information, then service control will access XML database across resource control module, package the result information and send it to event control module, at last return the result to the client user. If the request type is to operate the servers of cluster, such as system backup or recovery, Unix script execution and etc., service control will call resource control module to parse the XML data to string type, send these strings in bytes to the agents installed in the servers of cluster by socket, then the agents will call the different shell scripts to execute the operations on the actual servers. The result information will be returned to resource control module by socket. After receiving the result from resource control, Service control will package the result information into events, and send events to event control module; finally result information will be returned to the client user across communication control module.

Communication Control: Communication control is to build connection between client and management server, receive the requests from clients, forward the XML data and return the results. To deal the requests of multiple clients, we build a Session Data for each client to save the status, port and IP information. As there are multiple clients in the system, many users may operate one server, for example, one user sends a command of "OS Start", and then the other user sends a command of "OS Stop", after executing the two commands, the server OS is stopped. If the first user wants to know the final status of that server, he should refresh the client GUI manually. To solve the problem, we use polling mechanism to get server status every five seconds, return the information to client, and then refresh the client GUI automatically.

Resource Control: Resource control is to manage the server hardware and status information, including adding data, modifying data, and deleting data from XML database. To assure the validation and integrity of the data, before one client wants to get the server status, the server resources will be locked. After one request is dealt, the server resource will be unlocked.

Events Control: Event control is to manage the events in the system. Each client has a session data in the system to store events information, and the client requests will be sent to management server as different events. All the events from different clients will be queued. Event control will get the event information from session queue and build the events queue. The events that are not dealt will be sent to service control module, and the events dealt will be sent to communication control module. The result information will be again saved in session data, and then returned to client by communication control module. The events control procedure is as following:

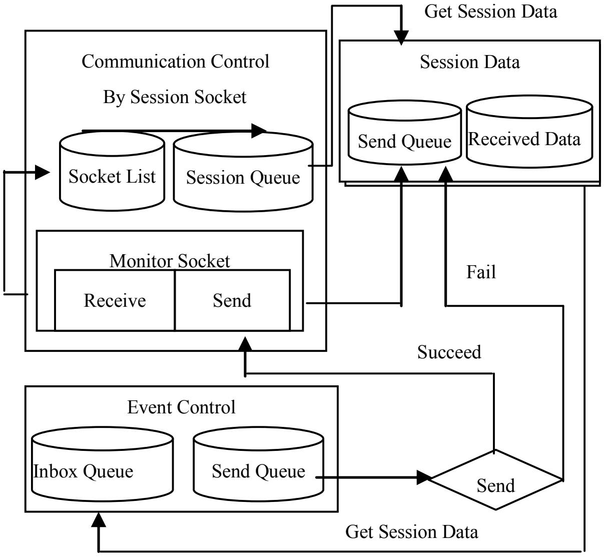


Figure3 Events Control Procedure

As figure3 shows, the Session Data stores the received data from client and the return result in Send Queue. Event control gets the session data into Inbox Queue, and sends the dealt events in Send Queue to Session Data. If succeed to send, the communication control will receive the events by monitoring socket and then send the dealt events to the Send Queue in Session Data. If failed to send, the events that are not dealt will be added into the Send Queue of Session Data directly. This means the socket error may occur. The Communication control builds a socket list and Session Queue to communicate with multiple users.

Monitor Control: Monitor control is to monitor the server's status in the cluster. It receives the report from the agents installed in the servers. If one server status has changed, the agent on it will send new status report to monitor control. The monitor control will update the server status information in XML database and send "Status Change" event to the event control module. At last the event will be return to the client and Client GUI will be updated too.

Agents: The agent on each server is a demon running in background. It is the crucial module, which makes the system own the ability of self-check, self-recovery. If we say the management server is the brain of the cluster system, then the agent is the nerve of each server. It perceives the health status of each server and reports it to management server. It receives the commands from management server and executes them, which is like that the brain tells the arms to stretch out or hold down. The agent is monitoring the request at one port, if new request comes, it will get the information, parse it into different command, and execute the command by cluster communication driver and Ignite-UX. The command type can be OS installation, software backup, patches update and so on. The final realization on each server is across shell scripts provided by UNIX core. Moreover, agent will check the hardware information every 30 seconds, then send health status report to monitor control module. If fault or error occurs in the server, the error information will be returned to management server at once.

3.3 Autonomic Computing Features

The servers in the cluster system accomplish one task cooperatively, providing a high performance environment for end users. If the cluster system owns the autonomic computing features that are, self-configuration, self-optimization, self-healing and self-protection, the system will work more effectively, and the system management will be much easier, therefore system management cost will be reduced [9-10].

The autonomic computing system we designed in this paper is for cluster system to realize self -management. The designed system has the following four features:

Self-configuration: Management server maintains two tables, one stores the server's status information, and the other stores the server's load information. The data in these two tables will be updated according to the real time information provided by monitor control module. When management server finds the workload of one server has exceeded the upper limit threshold value, it will find another server whose workload is normal to replace that server.

Self-healing: Health check is an important function in cluster system, which provides the automatic error check mechanism. This function is done by monitor control module. After some time, monitor control module will send request to collect hardware information of servers, the agent in each server will receive the request and collect the status information of each server and send health report to monitor control module. If error is found on one server, management server will use the backup image file on Ignite server to recover that server.

Self-optimization: This function is done by monitor control module. It will send the commands of software upgrade and patches update to the cluster system after some time. The servers will download the new software/patches from Ignite server and update themselves, therefore optimize themselves.

Self-protection: In the autonomic computing system, we divide the user into two groups, common user and administrators. Common users cannot execute the operations such as "OS install", "System recovery", "User Management”, and so on. Moreover, as to system installation and system recovery, password is requested before operation execution, which can reduce the risk and protect system from unsafe invasion.

4 Summary and future work

On the basis of analysis the problem in current cluster system, this paper designed and implemented an autonomic computing system for servers cluster based on multi-agent technique, which can make the cluster system work more effectively, make the complex system management become much easier, and make the system has the ability to manage themselves. The system designed has been put into practice, which has been proved to realize rapid software configuration, rapid error recovery and rapid application shift. But as the business needs vary quickly, the confirmed factors in system may vary accordingly. So now there is some incomplete design in system, which is the subject to study and improve in the future.

References

[1] ERAPHIN B. CALO and DINESH VERMA, “Toolkit for Policy Enablement in Autonomic Computing”, ICAC-04, International Conference on Autonomic Computing. IEEE Computer Society, National Science Foundation, IBM Corporation, SUN Microsystems, April 2004

[2] JEFF O. KEPHART, DAVID M. CHESS, “The Vision of Autonomic Computing”, Computer Journal, IEEE Computer Society, January 2003 issue

[3] Yoshihiro Tohma. Incorporating Fault Tolerance into an Autonomic-Computing Environment. IEEE Computer Society Vol. 5, No. 2; February 2004

[4] Zoran Constantinescu. Towards an Autonomic Distributed Computing System. Proceedings of the 14th International Workshop on Database and Expert Systems Applications (DEXA’03). IEEE Computer Society, 2003

[5] Rajkumar Buyya, “High Performance Cluster Computing”, Architecture and System, Beijing, Electronic Polytechnic Publishing Company，2001, pp. 15-17

[6] Kephart JO, Walsh WE. An artificial intelligence perspective on autonomic computing policies. In: Verma D, Devarakonda M, Lupu E, Kohli M, eds. Proc. of the 5th IEEE Int’l Workshop on Policies for Distributed Systems and Networks. New York: IEEE Computer Society, 2004. 3-12

[7] CATHERINE H. CRAWFOND and ASIT DAN. EModel: Addressing the Need for a Flexible Modeling Framework in Autonomic Computing. Computer Journal, IEEE Computer Society, January 2002 issue

[8] Tianfield H. Multi-Agent autonomic architecture and its application in E-medicine. In: Liu JM, Faltings B, Zhong N, Lu RQ, Nishida T, eds. Proc. of the IEEE/WIC Int’l Conf. on Intelligent Agent Technolocy (IAT 2003). Los Alamitos: IEEE Computer Society, 2003. 601-604

[9] R. Sterritt, “Towards Autonomic Computing: Effective Event Management”, Computer Journal, IEEE Computer Society, January 2003 issue

[10] CATHERINE H. CRAWFOND and ASIT DAN, “E-Model: Addressing the Need for a Flexible Modeling Framework in Autonomic Computing”, Computer Journal, IEEE Computer Society, January 2002 issue

The Research and Realization of the Parallel Gmres(m) Algorithm

Aimin Yang*[1] Shaohong Yan[1] Chunfeng Liu[1,2] Yamian Peng[1]
Yanbin Sun[1] Yan Yan[1]

1 College of Science, Hebei Polytechnic University, Tangshan, Hebei, 063000, China

2 Hebei University of Technology, Tianjin, 300130, China

Email：aimin@heut.edu.cn

Abstract

With the rapid development of high-speed network technology , the cluster systems have been the main platform of parallel algorithm. Because of the delay of their high communication, some parallel algorithms of fine grain are not fit to run in this environment. Therefore, it is necessary to study their parallel achievements in cluster systems. In terms of that, this paper aims at the internal parallel of the GMRES (m) method in order to find the solution of the linear equation groups and obtains coarse grain parallel algorithms, and more, we devise the program of this method using Fortran. At last, the example expresses that the designing parallel algorithm has much higher speedup in this cluster system.

Keywords: Cluster; GMRES (m) method; Parallel algorithm

1 Introduction

With the rapid development of the technology of the network, cluster system has become a main platform cluster system for parallel algorithm, which using high-speed universal network to dispatch a group of high-performance working stations or PCs integrally, assigned relevant supporting software, such as MPI, PVM, etc., constitutes a high efficient parallel processing system. Although the high-speed network has shortened the communication delay greatly, in fact it still influences the efficiency of the parallel algorithm. So the algorithm used in the cluster system only applies to the parallel of medium grain and above, which makes it necessary to design coarse grain parallel algorithm suitable to the network parallel. On the basis of the above-mentioned point of view, this paper provides a research of the parallel GMRES (m) in the cluster system in MPI according to the method of measurement of communication expense raised by B.K.Schmidt.

Most problems of science and engineering technology are solved by linear system of equations. In 1986 GMRES was put number of floating-point arithmetic involved in GMRES algorithm resulting in calculation amount increases in exponential order, the research of the parallel GMRES method is of great significance to get the GMRES method to solve the practical problems.

GMRES method referred in Reference [2-5] only ranges over the parallel algorithm of multiplying matrix by vectors and QR decomposition, which is not complete for integral parallel GMRES method. So according to the portioning principle, this paper makes a research of parallel pretreated GMRES（GMRES(m)）method and its parallel realization. In connection of GMRES (m) method inherent parallelism, this paper also raises parallel algorithm based on the cluster system.

2 Galerkin theory of linear system equations

Suppose the system of equations is $Ax=b$, in which A is a nonsingular large matrix, $b \in R^n$ is a known vector and the norm herein after is 2-norm. $\boldsymbol{K_m}$ and $\boldsymbol{L_m}$ are m dimensional subspaces, which are generated from $\{v_i\}_{i=1}^{m}$ and $\{w_i\}_{i=1}^{m}$. Supposing $x_0 \in R^n$ is a random vector and $x = x_0 + z$, $Ax = b$ is equivalent to $Az = r_0$, in which $r_0 = b - Ax_0$ .Galerkin Theory used in $Az = r_0$ can be stated that approximate result z_m is sought in the subspace K_m so as to get the residual vectors $r_0 - Az_m$ and all vectors in L_m reach orthogonality. That is to say, if $z_m \in K_m$ and $\forall w \in L_m$, we will get

$$(r_0 - Az_m, w) = 0$$

Then we can express Galerkin Theory in the symbol of matrix to solve large asymmetrical linear system equations. Saad and many other professionals have given a comprehensive and GMRES algorithm is regarded as one of effective solutions Galerkin Theory, are the fundamentals to the new algorithms, introduction of GMRES algorithm, whose astringency and equations whose coefficient matrix is asymmetrical. Now the algorithms of Arnoldi, GMRES and GRMES (m), based on utility have been proved by numerical experiments. But because forward by Yousef Saad and Martin H.Schultz, which is an iterative algorithm to solve large linear algebra system of a largeSuppose $V_m = (v_1, v_2, \cdots v_m)$ and $W_m = (w_1, w_2, \cdots w_m)$, in which $\{v_i\}_{i=1}^{m}$ and $\{w_i\}_{i=1}^{m}$ are the bases of K_m and L_m separately. So we can express z_m into $z_m = V_m y_m$, in which $y_m \in R^m$. Then $(r_0 - Az_m, w) = 0$ can be shown

$$\left(W_m^T A V_m\right) y_m = W_m^T r_0$$

Supposing $W_m^T A V_m$ is a nonsingular matrix, we can get an approximate resul

$$z_m = V_m \left(W_m^T A V_m\right)^{-1} W_m^T r_0$$

3 GMRES (M) Method

If we choose $L_m = K_m$, we call this Galerkin Method Arnoldi Algorithm; if we choose $L_m = AK_m$, we call this Galerkin Method as GMRES Algorithm. GMRES Algorithm has been improved greatly with the efforts from many professionals. It also has become the main method to solve large asymmetrical linear system equations through being integrated with various pretreatment technologies(Please read the reference [6-8]).

On the basis of the analysis of the upward section, we choose $K_m = span\{r_0, Ar_0, \cdots A^{m-1} r_0\}$, so we can find a set of standard orthogonal bases in K_m. Then $\|r_0 - Az\| = \|r_0 - AV_m y\| = \|r_0 - AV_{m+1}\bar{H}_m y\| = \|V_{m+1}(\beta e_1 - \bar{H}_m y)\|$ is got.

Because $V_{m+1}^T V_{m+1} = I$, $\|r_0 - Az\| = \|\beta e_1 - \bar{H}_m y\|$.

So minimizing $\|r_0 - Az\|$ in R^n equals to minimizing $\|\beta e_1 - \bar{H}_m y\|$ in K_m, which can be eventually concluded into solve least squares equation $\min\|\beta e_1 - \bar{H}_m y\|$.

The calculation process of GMRES Method can be concluded into,

(1) Select x_0, then calculate $r_0 = f - Ax_0$ and $v_1 = r_0 / \| r_0 \|$;

(2) Iterate $For \quad j = 1, 2, \cdots, k, \cdots$ till meeting the needs of ***do***

$$\text{G}\, h_{ij} = (Av_j, v_i) \quad (i = 1, 2, \cdots, j)$$

$$\hat{v}_{j+1} = Av_j - \sum_{i=1}^{j} h_{ij} v_i$$

$$h_{j+1,j} = \| \hat{v}_{j+1} \|$$

$$v_{j+1} = \hat{v}_{j+1} / h_{j+1,j}$$

(3) Construct an approximate solution $x_k = x_0 + V_k y_k$ in which y_k satisfies $\min J(y)$ ($J(y) = \| \beta e_1 - \bar{H}_k y_k \|$).

Theoretically speaking, if $\{A^i r_0\}_{i=0}^{n-1}$ near independence, while $m = n$, GMRES(m) algorithm should offer the accurately solution, but when m is very big, all the $(v_i)_{i=1}^{m}$ must be saved in the calculation, which will cause memory empty more larger to large scale problem, so it is unpractical. And when $k \to \infty$, not only internal memory and the amount of calculating are increasing, but also the orthogonalily of each array in the matrix V_k becomes relatively poor, this time the solution will oscillation in a small domain. While, after the original algorithm is pretreated, the difficulty is overcome when the technology of over again opening is supplied, then the GMRES(m) algorithm is obtained.

The concrete realized steps of the GMRES(m) algorithm are:

(1) let
$x_0=0,\ r_0=b-Ax_0,\ \beta=\|r_0\|,\ v_1=r_0/\beta,\ V_1=\{v_1\}$

(2) iteration： $For\ \ j=1,2,\cdots,m\ \ do$

$$h_{ij}=(Av_j,v_i)\,(i=1,2,\cdots,j),$$

$$\hat{v}_{j+1}=Av_j-\sum_{i=1}^{j}h_{ij}v_i$$

$$h_{j+1,j}=\|\hat{v}_{j+1}\|,\quad v_{j+1}=\hat{v}_{j+1}/h_{j+1,j}$$

$$V_{j+1}=(V_j,v_{j+1}),$$

$$\overline{H}_j=\begin{pmatrix}\overline{H}_{j-1} & h_{ij}\\ 0 & h_{j+1,j}\end{pmatrix}_{(j+1)\times j}$$

$\overline{H}_j$ is a upper Hessenberg matrix， when $j=1$, the first array is omission, and $AV_m=V_{m+1}\overline{H}_m$;

(3) solve the least square problem
$$\mathrm{e}\|r_m\|=\min_{y_m\in R^m}\|\beta e_1-\overline{H}_m y_m\|,$$
and y_m is obtained；

(4) conform the proximately solution
$$x_m=x_0+V_my_m;$$

(5) calculate the modulo of the residual vector
$$\|r_m\|=\|b-Ax_m\|;$$

(6) judge of again activation
$$\|r_m\|\le\varepsilon\begin{cases}yes: x=x_m\ and\ stop\\ no\ : x_0=x_m\ and\ turn\ to\ (1)\end{cases}$$

ε is the established reliance of convergent judgment, and often recommendable $\varepsilon=1.0\times10^{-6}$.

In (3), $\overline{H}_m$ must be changed into F_i $(i=1,2,\cdots,m+1)$ through plane rotation transformation in order to get y_m, in other words the QR decomposition must be proceeded to $\overline{H}_m$, that is
$$Q_m\overline{H}_m=R_m$$
in which $Q_m=F_1F_2\cdots F_{m+1}$ is a $(m+1)\times(m+1)$ matrix, R_m is a $(m+1)\times m$ upper triangular matrix (the elements of the last line are all zero), then
$$\min\|\beta e_1-\overline{H}_my_m\|=\|Q_m(\beta e_1-\overline{H}_my_m)\|$$
$$=\|g_m-R_my_m\|$$
in which $g_m=Q_m\beta e_1$. That is
$$\|r_m\|=\|b-Ax_m\|=\left|\overline{e}_m^Tg_m\right|$$
in which $\overline{e}_m^T$ is a $m+1$ dimension unit vector.

4 The parallel GMRES(M) algorithm

By the medium of analysis, the mainly calculation of the GMRES(m) algorithm includes the calculation of inner product for vectors, the calculation of matrix timing vector, the calculation of matrix timing matrix, the calculation of QR decomposition to solve the least square problem and etc. For the large linear problem, it is necessary to calculate using parallel methods of these segments in GMRES(m) algorithm. In the process of designing these parallel methods, we elementary base on the principle of separately, divide the original matrix or vector into some blocks, then distribute each block into various node machines, which will run the submission in dependently (Please read the reference [1]). It is a much better proposal to the cluster system which has no more nodes.

Research the realization of GMRES(m) algorithm in parallel cluster system, firstly we must apply some parallel methods (Please read the reference [9-11]).which have been designed, then construct the parallel GMRES(m) algorithm fitting cluster system.

In large linear problem, it is the foremost segment of the GMRES(m) algorithm to establish the test condition matrixes, to calculate the test conditions, and to solve the least square problem using QR decomposition. So it is necessary to make them parallel. The parallel GMRES(m) algorithm is organic combination of the three segments, concrete includes:

(1) In the orthogonal process of forming the V and H matrixes, the parallel methods of calculating inner product and matrix timing vector will be transferred;

(2) In the process of solving the least square problem, the parallel methods of QR decomposition, matrix timing matrix, and matrix timing vector will be transferred.

If let
$$A=(A_1^T,A_2^T,\cdots,A_p^T)^T,\quad b=(f_1^T,f_2^T,\cdots,f_P^T)^T$$

which is the form of dividing blocks, each block will be distributed various node, the parallel iterate algorithm of (1-1) will be accomplished under the parallel GMRES(m) algorithm.

In order to get the convergent solution of (1-1) more quickly, the matrix $\bar{H}_k$ will be formed over again, in every step, but its exponent will be increasing continuously. The particularly calculative steps are:

(1) $\forall X_0 \in R$, setup parameter ξ , α , β , m 。

(2) calculate $r_0^{(i)} = b - A_i X_0$ in each CPU $P_i (i = 1, 2, \cdots, P)$, get $r_0 = \sum_{i=1}^{P} r_0^{(i)}$ and $\|r_0\|$ through communication, then sent out r_0 and $\|r_0\|$ to P_i

(3) iteration: DO $k = 1$, n

Calculate $A_i v_k$ in P_i , get $A v_k = \sum_{i=1}^{P} A_i v_k$ through communication.

Such calculation will be run in P_i as:

```
$h_{ik} = ( A v_k , v_i ) , i = 1 , 2 , \cdots , k$
$\hat{v}_{k+1} = A v_k - \sum_{i=1}^{k} h_{ik} v_i$
$h_{k+1,k} = \| \hat{v}_{k+1} \|$
$v_{k+1} = \hat{v}_{k+1} / h_{k+1,k}$
let $\alpha_0 = \max_i \{ \| v_{k+1} \| , \| v_i \| \}$   $i = 1 , 2 , \cdots , k$
$f_k = | \bar{e}_m^T g_m |$   (to GMRES(m) algorithm)
IF ($f_k < \xi$) THEN
$X_k = X_0 + V_k y_k$
GOTO (4)
END IF
IF ($k = m$  $and$  $\alpha_0 > \alpha$) THEN
$X_k = X_0 + V_k y_k$
let $X_0 = X_k$
GOTO (2)
          END IF
          let $\beta_0 = f_k - \min_i f_i$  $(i = 1, 2, \cdots, k)$
          IF ($\beta_0 > \beta$) , THEN
               let $l$: $\min_i f_i = f_l$
               $X^{(l)} = X_0 + V_l y^{(l)}$
           $X_0 = X_l$
           GOTO (2)
END IF
   END DO
```

(4) the calculation will be independently accomplished in P_i,.

In these steps, the uppercase letter express matrix, the lowercase letter express vector.

There are some question must be attentive in calculation, including:

(1) Use $Q_k R_k$ decomposition of $\overline{H}_k$ when calculating f_k , get g_k from βe_1 , then judge x_k fitting the precision requisition whether or not, if fitting y_k will be obtained quickly, at last x_k is out.

(2) The proximate solution obtained must be fitted: when $n \to \infty$, $\| r_m \| < \xi$ (insure the precision requisition), $\max_{1 \le i \le m} \| v_i \| \le \alpha$ (insure the orthogonalily of v_i), $\max \left| \| r_k \| - \| r_i \| \right| \le \beta$ (insure the stability of process).

The above is the elementary calculating style of parallel GMRES(m) algorithm, which uses the technology of dividing blocks in parallel and the alternately technology of internal and outer storage, makes the scale of solving problem increased, make the calculating speed more quickly, make the analysis time descend, so the parallel GMRES(m) algorithm designed is more fit for calculating the system of the linear equations in the large engineering problems, and it offers a good kind of method to large supply of the system of equations.

5 Simulation of the algorithm

Use MPI to simulate the above algorithm in the internet of 1000Mbps, choose the equity model configuration ,and realize it using Fortran. In the 8-hodes cluster system, we separately use 2 nodes ,4nocks and 8nodes to simulate the parallel GMRES(m) algorithm, and compare it with serial runtime. In the cluster each node is p_4 2.6GHZ. Assure the child takes of the QR decomposition is separately 2 , 4 and 8.

The outcome of experiment is as table 1. In the table n expresses the rank of the matrix, P expresses the number of CPU, K expresses the number of the divided assignment we can see from the table that under the loom cluster environment, when p=2, the parallel algorithm gets a certain acceleration will also increase, but the acceleration ratio when k=4 is less than the acceleration ratio when k=8 .this is because when k increase, the date that transported also increase, so it causes the increasing of the communication. This is fit to the theoretical and practical analysis.

6 Conclusion

The parallel GMRES(m) algorithm 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 cluster environment.

Table 1 The outcome of the GMRES(m) algorithm solving the system of linear equations

n	Serial time (s)	K=2 P=2		
		Parallel time (s)	Acceleration ratio	Efficiency
600	90.66	58.49	1.55	0.78
800	130.58	81.61	1.60	0.80
1200	190.53	119.08	1.60	0.80

n	Serial time (s)	K=4 P=4		
		Parallel time (s)	Acceleration ratio	Efficiency
600	90.66	43.56	2.13	0.53
800	130.58	53.08	2.46	0.61
1200	190.53	76.21	2.50	0.62

n	Serial time (s)	K=8 P=8		
		Parallel time (s)	Acceleration ratio	Efficiency
600	90.66	44.22	2.05	0.26
800	130.58	62.18	2.10	0.26
1200	190.53	91.60	2.08	0.26

Acknowledgements

This paper is supported by the Natual Scientific Fund Project of Hebei Province (E2004000245), Scientific Technology Research and Development Plan Project (07160219B-6), Scientific Fund Project of Hebei Polytechnic University (z0808).

References

[1] 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

[2] Jianfei ZHANG and Hongdao JIANG, Direct Blocking parallel algorithm of large scale BF equation sets, Application Mechanics transaction 2003,4: 129-133

[3] W. Michacl. The Research and Development of Parallel Computation, Parallel Theory and Practice, 2000

[4] Xiaomei LI and Rongzeng JIANG, Parallel Algorithm, Science and Technology Press, Changsha, Hunan, 1992

[5] Yan ZHANG, Distributed Parallel Algorithm Designing, Analysis and Realization Doctor Thesis of Electronic Science and Technology University, 2001

[6] Yiming Chen. Network Parallel BEM for Band Precision Rolling, COMPUTATIONAL MECHANICS, WCCM VI in conjunction with APCOM'04, Beijing, China. Tsinghua University Press & Springer-Verlag, 2004,9: 5-10

[7] Zhihui DU, High-performance MPI programming Technology, Tinghua University Press, 2001

[8] Chunfeng LIU, Aimin YANG. The Parallel Algorithm of QR Decomposition of Matrix in Cluster System. World Academic Press. 2006:107-111

[9] R.W.Hockney. The Science of Computer Benchmarking. The Society for Industrial and Applied Mathematics, Philadelphia, 1996

[10] G.F.Psister. Clusters of Computers: Characteristics of an Invisible Architecture. IEEE Int'l. Parallel Processing Symp. Honolulu, 1996

[11] Saad Y, Schultz M H.GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetrical Linear System[J]. SIAMJ Sci Comp, 1986,7(3):856-869

PQPSO: A New Parallel Quantum-Behaved Particle Swarm Optimization*

Yan Ma[1] Yang Liu[1] Yupin Chen[2] Xiuzhen Li[3]

1 Department of Information Science and Technology, Taishan University, Tai’an, Shandong, 271021, China
Email: mayan1616@163.com

2 Jiangsu Wuxi Institute of Communications Technology

3 Department of Radiology, Taishan Medial University

Abstract

Quantum-behaved Particle Swarm Optimization is a new particle Swarm Optimization algorithm. Compared with Standard Particle Swarm Optimization (SPSO), it guarantees that particles converge in global optimum point in probability and this algorithm has better performance and stability. This paper based on Sun’s work [1, 2] introduces an improved Quantum-behaved Particle Swarm Optimization Algorithm, multi-swarm Parallel Quantum-behaved Particle Swarm Optimization (PQPSO). In this algorithm, employs the co-evolution model to avoid pre-maturity and improve global search performance. This approach is tested on several accredited benchmark functions and the experiment results show much advantage of PQPSO to SPSO and QPSO, and the running time is also decreased in linear.

Keywords: parallel; quantum; particle; swarm; co-evolution

1 Introduction

Numerical optimization has been widely used in engineering to solve a variety of NP-complete problems in areas such as structural optimization, neural network training, control system analysis and design, and layout and scheduling problems to name but only a few. In these and other engineering disciplines, two major obstacles limiting the solution efficiency are frequently encountered. First, large-scale problems are often computationally expensive, requiring significant resources in time and hardware to solve. Second, engineering optimization problems are often plagued by multiple local optima and numerical noise, requiring the use of global search methods such as population-based algorithms to deliver reliable results.

The Quantum Particle Swarm Optimization (QPSO) algorithm is one of emerging global search methods [1, 2] based on the quantum theory. It is a kind of new swarm intelligence algorithm after Ant Algorithm and is particularly suited to continuous variable problems and has received increasing attention in the optimization community. It has been successfully applied to large-scale problems in several engineering disciplines and, as a population based approach, is readily parallelizable [3]. It also has fewer algorithm parameters than either GA or PSO algorithms. Furthermore, Generic settings for these parameters work well on most problems [4].

In this paper, the Quantum-Behaved Particle Swarm Optimization (QPSO) is briefly described. In the next section, put the parallelisms crude of QPSO and high speed of computer together, island model is introduced. The Parallel Quantum-Behaved Particle Swarm Optimization (PQPSO) is reported. In section 3, the test problems and the results of experiments are reported in Section4. The paper ends with the

* This work was supported by the Taishan University Scientific Research Projects under Grant No. Y07-2-12 and Shangdong Province Education Department Research Projects under Grant No. J07WJ25.

conclusion and ideas for future research in Section 5.

2 QPSO

In a PSO system, a particle corresponds to individual of the organism, which depicted by its position vector $\vec{x}$ and its velocity vector $\vec{v}$, is a candidate solution to the problem. That is the trajectory of the particle is determined. Then the optimal solution of the probability of moving out the trajectory is ignored. Therefore, in general, PSO can obtain good solutions in high-dimensional spaces but the ignorance of optimal solution does exist and PSO stumbles on local minima.

Keeping to the philosophy of PSO, we proposed a Delta potential well model of PSO in quantum world (QPSO) [2]. Because $\vec{x}$ and $\vec{v}$ of a particle are not determined simultaneously according to uncertainty principle, the term trajectory is meaningless in quantum world.

2.1 Classic PSO algorithm

In a classical PSO system proposed by Kennedy and Eberhart, the particles are manipulated according to the following equation:

$$v_i(t+1) = wv_i(t) + \varphi_1(p_i - x_i(t)) + \varphi_2(p_g - x_i(t)) \tag{1}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{2}$$

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

Where *x* and *v* denotes the position and velocity of particle *I* among the population correspondingly, $\varphi_1{}^T$ and $\varphi_2{}^T$ are two random vectors in the range [0,1].

In Eq.(1), vector p_i is the best position (the position giving the best fitness value) of the particle *i* and vector p_g is the position of the best particle among all the particles in the population. Parameter w is the inertia weight [4], which does not appear in the original version of PSO [3]. In [5], M. Clerc and J. Kennedy analyze the trajectory and prove that, whichever model is employed in the PSO algorithm, each particle in the PSO system converges to its local point *p*, whose coordinates are $p_d = (\varphi_{1d} p_{id} + \varphi_{2d} p_{gd}) / (\varphi_{1d} + \varphi_{2d})$ so that the best previous position of all particles will converge to an exclusive global position with $t \to \infty$.

2.2 QPSO

In Quantum-Behaved Particle Swarm Optimization (QPSO) [4], the particles move according to the following equation:

$$mbest = \frac{1}{M}\sum_{i=1}^{M} p_i = \left(\frac{1}{M}\sum_{i=1}^{M} p_{i1}, \frac{1}{M}\sum\nolimits_{i=1}^{M} p_{i2}, \ldots, \frac{1}{M}\sum\nolimits_{i=1}^{M} p_{id}\right) \tag{3}$$

$$p_{id} = \varphi * p_{id} + (1-\varphi) * p_{gd}, \varphi = rand() \tag{4}$$

$$X_{id} = p_{id} \pm \beta * \left|mbest_d - X_{id}\right| * \ln(\frac{1}{u}), u = rand() \tag{5}$$

where *mbest* is the mean best position among the particles. φ and *u* are a random umber distributed uniformly on [0,1] respectively and β is the only parameter in QPSO algorithm. The only parameter β for QPSO is set from 1.2 to 0.5, which decreased linearly.

3 Parallel Particle Swarm Algorithm And Implement

3.1 Parallelization

The development in microprocessor technology and network technology has led to increased availability of low cost computational power through clusters of low to medium performance computers. To take advantage of this, communication layers such as MPI and PVM have been used to develop parallel optimization algorithms, the most popular being gradient-based, Genetic Algorithm (GA), and Simulated Annealing (SA). In biomechanical optimizations of human movement, for example, parallelization has allowed previously intractable problems to be solved in a matter of hour. As a result of QPSO algorithm is based on evolution algorithm of swarm, contains dormant parallel mechanism, so it can be paralleled and not limited to number of computers[7, 8].

Parallel Quantum-Behaved Particle Swarm Optimization algorithm put the high speed of Parallel computers with the parallel of quantum-behaved particle swarm optimization, promoting the computational speed and quantities of QPSO very distinctly.

3.2 Parallel particle swarm optimization

Like many other evolutionary algorithms, the major problem confronts Quantum-Behaved Particle Swarm Optimization Algorithm is premature convergence, which results in great performance loss and sub-optimal solutions. With QPSOs the fast information flow between particles seems to be the reason for clustering of particles [6]. Diversity declines rapidly, leaving the QPSO algorithm with great difficulties of escaping local optima. Another problem with QPSO in multi-modal optimization is computational cost. With the diension of the optimization problem increasing, the population size must be enlarged to ensure the algorithm have a good performance, which, however, makes the algorithm computationally expensive.

To solving the aforementioned problems, we propose in this paper a Parallel Quantum-Behaved Particle Swarm Optimization (PQPSO) the multi-stage portfolio optimization. The PQPSO employ the co-evolution model, in which the global population (swarm) is partitioned into q subpopulations, where q is the number of PPEs (physical processing elements). The PPEs communicate periodically to exchange the gbest and the communication is a synchronous voting that the gbest of a subpopulation is broadcast to all the PPEs. The subpopulation stores the gbests received from other counterparts in its local memory and randomly selects a gbest at each iteration to and adjust its position according to the equation (5).

The period of communication for the PPEs is set to be T number of generations (iterations), which follows an exponentially decreasing sequence: initially $\lceil N_g/2 \rceil$, then $\lceil N_g/4 \rceil$, and so on, where N_g is the maximum number of generations. The rationale is that at the beginning of the search, the diversity of the global population is high. At such early stages, exploration is more important than exploitation; therefore, the PPEs should work on the local subpopulation independently for a longer period of time. When the search reaches the later stages, it is likely that the global population converges to a number of different gbests. Thus, exploitation of more promising position is needed to avoid unnecessary work on optimizing the local gbest positions.

With the above design considerations, the parallel Quantum-Behaved Particle Swarm Optimization Algorithm is outlined below.

```
1) Initialize a population (array) which including m particles, For the ith particle, it has random location Xi in the problem space.
2) Partition the population into q subpopulations with size equal to M/q and assign to the different PPEs, respectively. M is the size of the whole population;
3) i=0;
4) Do
5) T=0;
6) Do
7)     For each of the subpopulation do
8)         Evaluate the desired optimization fitness function for each particle;
9)         Compare the evaluated fitness value of each particle with its pbest. If current value is better than pbest, then set the current location as the pbest location. Furthermore, if current value is better than gbest, then reset gbest to the current index in particle array;
10)        Change the location of the particle according to the Eq.(3), Eq.(4) and Eq.(5), respectively;
11)    Endfor
12) Until ++T=$\lceil N_g/i \rceil$;
13) Accept the gbest position from a remote PPE to replace the gbest of the subpopulation.
14) $i = i \times 2$
15) Until the total number of generations elapse equal to $N_g$
```

4 Experiments And Results

4.1 Test function

To test the performance of parallel particle swarm algorithm, three benchmark functions are used here for comparison with SPSO, the two functions are:

Rosenbrock Function, shaffer's Function. The two functions are all minimization problems with minimum value zero. For the purpose of comparison, Table 1 lists the initialization ranges and $V_{\max}$ and $X_{\max}$ values for all the functions, respectively. The fitness value is set as function value.

Table 1 Test function

	Functions	Initial Range	$X_{\max}$	$V_{\max}$
f1	$\sum_{i=1}^{n}100(x_{i+1}-x_i^2)^2+(x_i-1)^2)$	[15,30]	-100	100
f2	$f(x)_3=0.5+\frac{(\sin\sqrt{x^2+y^2})^2-0.5}{(1.0+0.001(x^2+y^2))^2}$	[30,100]	-100	100

4.2 The comparison of the mean fitness value and the run-time

We had 50 trial runs for every instance and record mean best fitness values and run-time for 50 runs of each functions in Table 2. to Table 7.. In order to investigate the scalability, different population sizes M are used for each function with different dimensions. The population sizes are 40 and 80. Generation is set as 1000, 1500 and 2000 generations corresponding to the dimensions 10, 20 and 30 for the first function. Dimensions of the second function are two. in most cases the performance of the PQPSO is better than GA (Genetic Algorithms) PGA (parallel Genetic Algorithms)[9,10], (parallel Binary Genetic Algorithms), DGA(Decimal Genetic Algorithms),PDGA(Parallel Decimal Genetic Algorithms),SPSO, PSPSO(parallel standard particle swarm optimization) and QPSO.

The mean fitness value of Rosenbrock Function is showed from Table 2. to Table 3..

Table 2 Rosenbrock function, the mean fitness value of 40 population sizes

CPU	DIM	Gmax	PGA	PDGA	PPSO	PQPSO
1	10	1000	249.563	85.166	9.4043	16.155
	20	1500	467.584	262.933	142.438	43.8411
	30	2000	552.457	344.563	508.957	167.229
2	10	1000	114.951	44.4003	18.658	8.07985
	20	1500	2960.62	152.422	143.921	26.1934
	30	2000	37284.6	218.74	210.585	91.4359
4	10	1000	2532.21	90.8629	9.4043	8.7288
	20	1500	169745	250.92	142.438	13.462
	30	2000	805365	478.589	408.957	62.138

Table 3 Rosenbrock function, the mean fitness value of 80 population sizes

CPU	DIM	Gmax	PGA	PDGA	PPSO	PQPSO
1	10	1000	253.442	57.5812	37.3747	5.8455
	20	1500	271.97	182.064	83.6931	23.5401
	30	2000	481.841	216.589	202.072	71.1963
2	10	1000	87.5399	29.8941	10.3677	4.71275
	20	1500	128.577	140.064	46.5886	15.9374
	30	2000	372.719	134.792	74.7392	30.7117
4	10	1000	130.824	42.0861	8.1202	3.98273
	20	1500	2612	54.4882	30.8569	15.7344
	30	2000	24420.9	217.842	91.8591	32.0939

Table 4　The mean fitness value of Shaffer's Function

M	CPU	M	Gmax	PGA	PDGA	PPSO	PQPSO
2	1	40	2000	0.0513525	0.168217	0.00062	5.05799e-4
		80	2000	0.04883636	0.158856	0.000286	1.97642e-4
	2	40	2000	0.0475789	0.127803	0.0005761	4.39723e-4
		80	2000	0.0463202	0.116415	0.0001985	4.77496e-5
	4	40	2000	0.0523936	0.106613	0.00063	2.52634e-4
		80	2000	0.0450622	0.0891515	0.0001035	1.39142e-5

Table 5　Rosenbrock function, the run-time of 40 population sizes

CPU	DIM	Gmax	PGA	PDGA	PPSO	PQPSO
1	10	1000	0.496713	0.26773	0.408314	0.340196
	20	1500	1.47704	0.804395	1.25326	1.00294
	30	2000	2.99084	1.59162	2.50315	2.17644
2	10	1000	0.372596	0.142056	0.220722	0.180823
	20	1500	1.10229	0.434556	0.66117	0.544956
	30	2000	2.23922	0.877119	1.53776	1.17268
4	10	1000	0.334903	0.104019	0.205542	0.16763
	20	1500	1.00478	0.320019	0.599977	0.47056
	30	2000	2.15482	0.592764	1.244269	0.913631

Table 6　Rosenbrock function, the run-time of 80 population sizes

CPU	DIM	Gmax	PGA	PDGA	PPSO	PQPSO
1	10	1000	0.989269	0.526976	0.788501	0.67769
	20	1500	2.96494	1.58227	2.49435	1.97663
	30	2000	5.97496	3.13876	4.77487	4.32968
2	10	1000	0.739336	0.298362	0.427034	0.379494
	20	1500	2.19878	0.863061	1.27953	1.08661
	30	2000	4.49334	1.65878	2.60106	2.27469
4	10	1000	0.718366	0.234039	0.402065	0.252708
	20	1500	1.99295	0.604871	1.077497	0.707123
	30	2000	4.07133	1.2007	2.24484	1.76248

Table 7　The run-time of Shaffer's Function

M	CPU	M	Gmax	PGA time	PDGA time	PPSO time	PQPSO time
2	1	40	2000	0.250742	0.177153	0.220546	0.193113
		80	2000	0.499748	0.357093	0.4163514	0.377872
	2	40	2000	0.20154	0.0961906	0.128751	0.101304
		80	2000	0.391077	0.184431	0.234487	0.194171
	4	40	2000	0.190688	0.0780975	0.1025845	0.0864447
		80	2000	0.340708	0.140757	0.1800742	0.145721

4.3 Analyze the test result

Compare the run-time of processor in Figure 1. (Next figure express population sizes 80 are used for the first function with 30 dimensions and the second function with 2 dimensions, generation is set as 2000, and the processor number is 1, 2 and 4.)

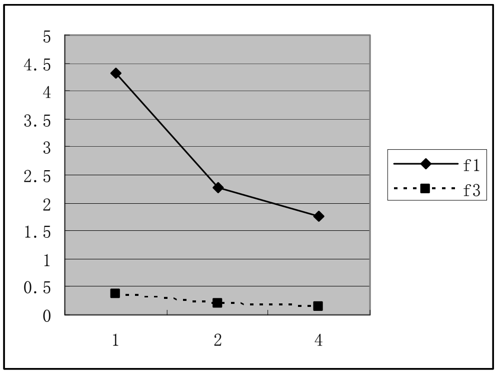


Figure 1 Run-time of processor

5 Conclusion

According to the experimental results, in most cases the performance of the PQPSO method in coping with test functions is better than QPSO and the class PSO algorithm, and than those obtained through other EAs. Saving computational time is very important. Time can be saved from parallel or distributed processing.

On the premise of better results than QPSO, according to the table 5., table 6. and table 7., Proximity linear speedup is acquired in the present test by increasing processor. But the speedup is not linear; firstly, Parallel programming is not very difficult, secondly, limited to hardware and network conditions. According to the figure 1., when the problem scale is increasing , the speed-up is better.

Future work will focus on approaches of solving large-scale problems or apply it to problems with the run-time demands strictly.

References

[1] Others: Sun, J. and Xu W.B.: A Global Search Strategy of Quantum-behaved Particle Swarm Optimization

[2] Others: J. Feng B. and Xu W.B.Particle Swarm Optimization with Particles Having Quantum Behavior. Proceedings of 2004 Congress on Evolutionary Computation. (2004) 325-331

[3] Others: J. Kennedy and R. Eberhart, “Particle Swarm Optimization”, Proc. IEEE Conf. On Neural Network, 1942-1948 (1995)

[4] Others: Y. Shi and R. Eberhart, “Empirical study of particle swarm optimization,” Proc. Congress on Evolutionary Computation, 1945-1950 (1999)

[5] Periodicals: M. Clerc and J. Kennedy, “The particle swarm: explosion, stability, and convergence in a multi-dimensional complex space,” IEEE Trans on Evolutionary Computation, vol. 6, no. 1, pp58-73 (2002)

[6] Others: Jun Sun and Wenbo Xu, “Particle Swarm Optimization with Particles Having Quantum Behaviour” IEEE Congress. Evolutionary Computation (2004)

[7] Others: K.E. Parsopoulos and M.N. Vrahatis, “Particle swarm optimization method for constrained optimization problems”, Intelligent Technologies-Theory and Application, 214-220 (2002)

[8] Others: J.F.Schutte, J.A.Reinbolt, B.J.Fregly, R.T.Haftka, A.D.George Parallel Global Optimization with the Particle Swarm Algorithm

[9] Others: Angeline,P. J. Tracking exterma in dynamic environments. Proc. Evolutionary Programming V1.Indianapolis,IN.pp.335.345,1998

[10] Others: Back, T. On the behavior of evolutionary algorithms in dynamic environments.Proc. Int.Conf. on Evolutionary Computation.Piscataway, NJ: IEEE Press, pp. 446-451, 1998

Research on Application of Distribute OLAP System with an Improved Election Algorithm*

Lijun Wang

Economics and Management Department, North China Electric Power University, Baoding, Hebei, China
Email: Email:wljazg2008@yahoo.cn

Abstract

The OLAP system can support enterprise decision well through intelligent query with data of enterprise. To deal with abundant and complicated query tasks OLAP system needs high performance server, so that the cost of OLAP system is higher for most middle and small enterprises. With the aid of distribution technology, we can deploy OLAP system on multiple minicomputers according to allocating tasks to each server, so we can attain application of OLAP with existing resource of enterprise. The system coordinator is selected by using election algorithm. In a full-connected LAN, the adopted election schemas are generally the classic Bully algorithm by Garcia-Molina or improved ones. In this study, we propose an improved election algorithm suitable to the distributed OLAP system. Using the new algorithm can improve the system performance and produce fewer messages.

Keywords: OLAP, Data Warehouse, Distributed system, election algorithm

1 Introduction

On-Line Analytical Processing (OLAP)[1] is a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user. The quantity of data that the OLAP system deals with is massive, and in the actual application of enterprise, users from various departments may submit lots of query tasks. With the increasing of numbers of query tasks, the OLAP system needs higher quality server, thus the cost increases fast. Meanwhile, submitting lots of query tasks at one moment may reduce the reliability of the system.

The distributed OLAP system can solve the problem well. According to distributed technology, we can deploy OLAP application on several minicomputers, and all the servers provide service to users as a whole [2]. The internal structure of the system is transparent to users, so we can use some minicomputers or microcomputers instead of high quality servers. In the distributed OLAP system, there is only one coordinator at a moment, when a new query task arrival, the coordinator should allocate it to an idle or low load server of the system, so the whole system can keep load balance [3].

In a distributed system that using centralized scheduling scheme, the coordinator has an important effect on system performance. In the distributed system, the coordinator not only allocates the tasks to other servers, but also performs the query task, so the coordinator should be the highest quality server of the system. When the coordinator overload, it is failed to coordinate the system, a new coordinator should be selected. The election algorithm includes bully algorithm and ring algorithm [4]. In a full connected LAN the bully algorithm is adopted usually [5]. The classic bully algorithm can select the optimal node as the coordinator, but it may cost long time and produce large messages. Particularly when the process whose

* This work is supported by Youth Foundation of North China Electric Power University (No. 200711024)

number is small initiates the election, it may appear broadcast storm in the network. To deal with these problems, bully algorithm used in actual application is improved usually. The improvement ways include changing detect method, setting standby coordinator and simplifying election. In this study, an improved bully algorithm is adopted in the distributed OLAP system, and it is modified according to actual application.

2 Distributed OLAP System Structure

Objects of the distributed OLAP system include enhancing the system's query and analysis speed, improving the system's reliability, enhancing the system's scalability and reducing the cost of deploying OLAP system.

The distributed system is composed of multi-OLAP servers, multi-data input module, and data querying module and application severs. These OLAP servers run on microcomputers that use TCP/IP protocol to communicate with local area network. The OLAP servers use SQL Server 2000/Analysis Services [6]. Figure 1 shows the system structure.

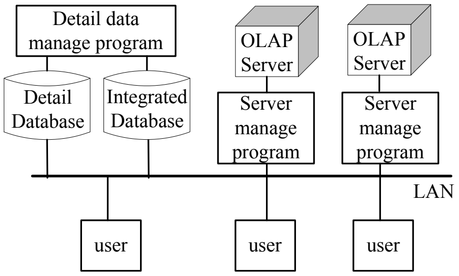


Figure 1 the distributed OLAP system structure

In the distributed OLAP system, firstly the data warehouse is established, in the detail database there are historical data and data input by departments of the enterprise. Then data is processed by data changing program, it is arranged, extracted and inducted according to the subject and stored into the Integrated Database, thus the data warehouse is finished [7]. Secondly the OLAP server begins to establish corresponding data cubes according to the kind of data warehouse. The same server manage program is deployed on every server; it charges the allocation of query tasks, the election of coordinator and basic I/O operations. In the LAN, the system performs in the LAN Because the Communication is high efficient and high reliable. The system and the users are connected by network. Each server in the system can receive task and send query result back to user. All the servers provide services to users as a whole. To the users it is just like there is only one server in the system.

3 The Distributed OLAP System Election Algorithm

3.1 Bully algorithm

A distributed system requires a process to act as coordinator. An "election algorithm" can perform the selection of this process automatically. The commonly used election algorithms are bully and ring algorithm. In a full-connected network, bully algorithm is usually adopted. For simplicity, we assume the following:

1) Processes each have a unique, positive identifier.

2) All processes know all other process identifiers.

3) The process with the highest valued identifier is duly elected coordinator.

4) When an election "concludes", a coordinator has been chosen and is known to all processes.

In a distributed system, the working of the bully algorithm is as follows:

1) When a process "notices" that the current coordinator is no longer responding (4 deduces that 7 is down), it sends out an ELECTION message to any higher numbered process and waits for respond messages.

2) If none respond, it 4 becomes the new coordinator of the system, and then it sends out a COORDINATOR message to all other processes informing them of this change of coordinator).

3) If a higher numbered process responds to the ELECTION message with an OK message, the election is cancelled and the higher-up process starts its own election (5 and 6 in this example both start, with 6

eventually winning).

4) When the original coordinator 7 comes back on-line, it simply sends out a COORDINATOR message, as it is the highest numbered process.

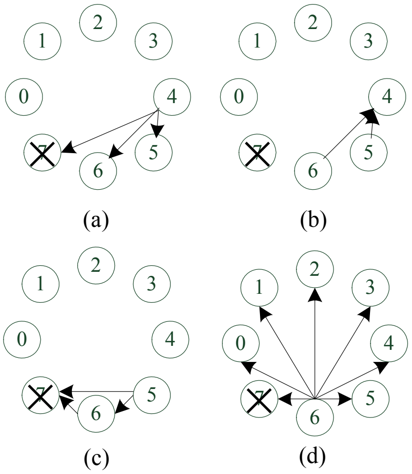


Figure 2 Process of Bully algorithm

3.2 Algorithm analysis and improvement

When the coordinator is failed the classic bully algorithm can chose the biggest number process of the other processes can as a new coordinator. But the disadvantages are that when a small number process initiates an election the communication quality is massive. For the example of 3.1, in the system with 8 processes, no.4 process initiates the new election, it chose no.6 as the new coordinator, and there are 14 passing messages. If the no.2 process initiated the election there would be 32 passing messages. When there are more processes in the system, and the coordinator fails frequently, the communication quality will increase rapidly, so the system performance decreases fast even collapses.

To deal with the problem, the paper [8] proposed an improved bully algorithm. The improvement idea focuses how to decease the communication quality. When a process notices the failed coordinator it initiates a new election, if no none respond it becomes the coordinator. If someone responds it check it own process table to find the biggest number process within the respond ones, and appoint it as the new coordinator. At last the process send message to all the other processes. The working of the improved bully algorithm is as follows:

1) When a process notices that the current coordinator is no longer responding, it sends out an ELECTION message to any higher numbered process.

2) If none respond, it becomes the coordinator.

3) If someone responds, it checks its own process table to find the biggest number process among the responding ones and appoint it as the new coordinator.

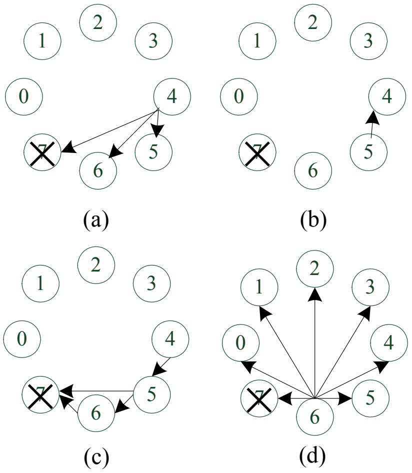
Using the improved bully algorithm can decrease the passing messages obviously, so the system can chose a new coordinator after the old one failed, meanwhile the system can keep high performance. But the improved one cannot adapt to the distributed OLAP system well. In the system, each node has a process table and the table will update dynamically. When a new node is added into the system, the other nodes update the process tables to add the new node, and when a node fails and cannot respond to others, the record should be deleted with the table. So when a election is performed as the improved bully algorithm, one node appoint another node as the new coordinator, but actually it may not be the biggest number process among the active processes. If the responders are small number process and the big number processes have not received message or the respond message lost, the appointed one is a small number process. Thus the system performance will reduce. Based the above analysis, we improved the algorithm further through adding some new steps.

4) If none respond then appoint the biggest active one of the table, and the process send no message.

5) The appointed process check it own process table, and send messages to the bigger number ones.

6) Repeat the step 5), until a new coordinator is selected.

In the Figure 3, no.4 node sends message to no.5, 6 and 7, no.6 don’t respond as network or other cause, so the no.4 node appoints no.5 as the coordinator, no.5 node continues the election process, and no.6 responds to no.5 node, then the no.6 nodes is selected as the last coordinator. The same measures are performed for the bigger number node, the method can not reduce more communication quality than the algorithm in the paper [8]. But it can ensure selecting a bigger number node as a coordinator. In the distributed OLAP system, when there are more query tasks, it is very important to select a high performance coordinator, so the new improved bully algorithm is suitable to the system.



(a) (b)

(c) (d)

Figure 3 Process of improved Bully algorithm

4 The Distributed OLAP System Election Algorithm Performance Analysis

According to the classic bully algorithm, when the system coordinator fails, the first node that notices the failed coordinator is randomized, it means that every node of the last nodes may initiate a new election process. In a distributed system with n nodes, it is assumed that only one node notices the failed one. We can analyze the messages in two cases [9].

1) The node that notices the failed coordinator is just next to the coordinator, and the node initiates an election process. It becomes the new coordinator because there is no bigger number node. The node sends n-2 messages to inform other nodes that it is the coordinator now, and then the election process end.

2) The node that notices the failed coordinator is just smallest number node, and it initiates the election. The node sends n-1 messages to other nodes, as the node number is smallest, then a new election is performed within the n-2 nodes. So the algorithm s complexity is ○(n^2) in the condition[10].

According to the new improved bully algorithm, it is also assumed that only one node notices the failed coordinator. If the node that notices the failed coordinator is the no.i and the node initiates a new election, we can calculate the messages quality as follow:

1) The no.i sends n-i messages, and receive n-i respond messages at most. Through checking the table The no.i node appoints the biggest number node of the respond ones as the new coordinator, the new coordinators then send n-1 messages to the other nodes. Thus the messages quality is 2*(n-i)+n-1.

2) The node appointed by the initializing node doesn't become the coordinator, it finds bigger number node through checking the table, and appoints the new node as a coordinator. Then the messages quality is 2*(n-i)+n+1.

From the analysis of the system performance, we can draw the conclusion: the complexity of the new improved bully algorithm is ○(n), and when the next bigger node initiates the election process, there are least messages and it's quality is n-2.

Through experiment, the election processes are performed using the two algorithms in the distributed OLAP system that has 20 nodes. We can calculate the different messages quality through controlling the node that initiates the election process.

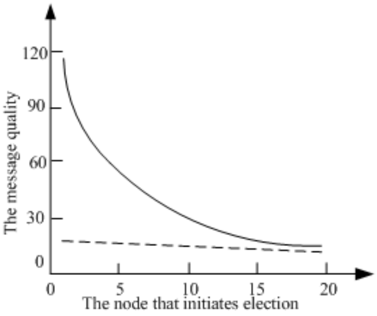


Figure 4 Comparison of Bully algorithm and the new improved one

In the fig4, the full curve shows the performance result of bully algorithm and the broken curve shows that of the new improved one. According to the bully algorithm, the system messages increase rapidly with the number of the initiating election decreasing. And the new improved one can maintain low message quality, so the system is high efficient and high stable.

5 Conclusion

With the developing continuously of decision support system, more and more enterprises begin to deploy OLAP application. But the OLAP system with one server is inefficient and expensive. The introduction distribute technology into OLAP system can assign tasks into different servers [11]. So the OLAP application can run on servers with low configuration or microcomputers. And the system can provide services to user as a whole with the help of the coordinator. In the distributed OLAP systems, the coordinator is selected according to election algorithm. In this paper we propose a new improved bully algorithm. Through the experiments analysis we find that the algorithm can select the high performance coordinator and maintain low messages quality. When there are more nodes in the system, the superiority of the algorithm can be better embody. In a less numbers of nodes system, it has the similar performance to the classic bully algorithm.

References

[1] Sen Arun. Metadata management: past, present and future. Decision Support System, 2004, 37(4):151-173

[2] George Coulouris, Jean Dollimore, Tim Kindberg. Distributed system Concept and Design. Jin Beihong. China Machine Press, Beijing,2004

[3] Ye Deqian, Ma Qinyong. Research on Improving Synthetical Performance of a Mini-OLAP System with Multi-Services. Mini-Micro Systems. 2002, 23(4), 486~488

[4] Zhang Gang, Gan Hongmin, Cai Zhiping, Zhang ying. Election algorithm based on trust degree. Journal of Shenyang University of Technology. 2007, 1, 81-85

[5] Molina, H. Garcia, Elections in distributed computing systems. IEEE Trans. On Parallel and Distributed Systems, Vol31, No1, 1982,1,47-59

[6] Reed Jacobson. SQL Server 2000 Annalysis Services[M]. Microsoft Press, 2000, 245~330

[7] Chau K.W., Cao Ying, Anson M. Application of Data Warehouse and Decision Support System in Construction Management. Automation in Construction, 2003, (3): 213-224

[8] Li Xiaoting, He Boxiong, Zhong Lianqiong. Bully algorithm and optimization in distributed systems. Journal of Xi'an Technological University. 2004, 24(3), 210-213

[9] Nakano, Koji, and Olariu, Stephan, Uniform leader election protocols for Radio networks IEEE Trans. On Parallel and Distributed Systems, Vol13, No5, 2002,5,516-529

[10] Shi Zhongzhi, Wu Chenggang. Module-Based Mobile Agent and Its Schedule Method. Journal of Software, 2002,13(8):1628-1636

[11] Jin Zhengshu, Hu Guang. Application of inquiry optimized in the Distributed database system. Computer Applications and Software. 2003,58-60

A Billing System Model Based on Parallel Processing

Yong Zheng[1] Hanhua Lu[2] Yanfei Sun[2]

1 Computer Institute, Nanjing University of the Posts & Communications, NO.66. Xin Mofan Road
Nanjing, Jiangsu Province, 210003, China
Email: zy_njupt@hotmail.com

2 Information Network Institute, Nanjing University of the Posts & Communications, NO.66. Xin Mofan Road
Nanjing, Jiangsu Province, 210003, China
Email: Luhh@njupt.edu.cn, Sunyanfei@njupt.edu.cn

Abstract

Telecom billing system is one of the core systems of telecom operators. Nowadays, customers' demand for Real-time Processing becomes higher and higher. Parallel processing is usually used to process complex issues, promote efficiency, save time and expenses. In this paper, a billing system model based on parallel processing is put forward, which can solve the problems and can be expanded, improved and be more widely used in the telecom industry. Many parallel methods are discussed in the model, such as synchronization, loading balance, pipelining, etc.

Keywords: Billing System, Real-time Processing, Parallel, Multi-processor, Telecom Industry

1 Introduction

With the approaching of the Next Generation Network (NGN) and the growing of Communication Services, there has been an increasing demand for Real-time Processing. Customers wish to acquire their daily or immediate consumption. Taking into account the system performance and the server capacity, parallel processing must be used in order to meet the system requirements and reduce the response time of multi-processor.

As one of the hottest technology, parallel processing has developed a lot recent years, especially after the appearance of the Multi-processor. More and more attention have been focused on that how to use Multi-processor more easily and more efficiently. Parallel Computation and Distributed Cluster have emerged in rapid development and wide use to reverse the situations.

Nowadays, the traditional billing model cannot adapt to the times change, and many solutions have been put forward to achieve the system requirement better.

This paper made a synthesis of several parallel processing methods, and developed a feasible solution based on a billing system model.

2 Description

2.1 Description of the billing system

Jiangsu Province has a total 13 cites or prefectures. Assumed that billing center of Jiangsu Province must collect all billing files from the 13 cites or prefectures, and then need to be in order to do a series of options, such as pretreatment, rating and warehouse entry, which must be executed followed a strict sequence. General speaking, warehouse entry always costs a tripled time compared with the other two procedures. All the data will be delivered to the 13 cites or prefectures after been processed.

* Supported by National High-Tech Research and Development Plan (863) of China (No.2005AA121620,2006AA01Z232) and Natural Science Foundation of Jiangsu Province (No.2007BK603),Talent Introduction Foundation of Nanjing University of Posts & Telecommunications (No.NY2007044)

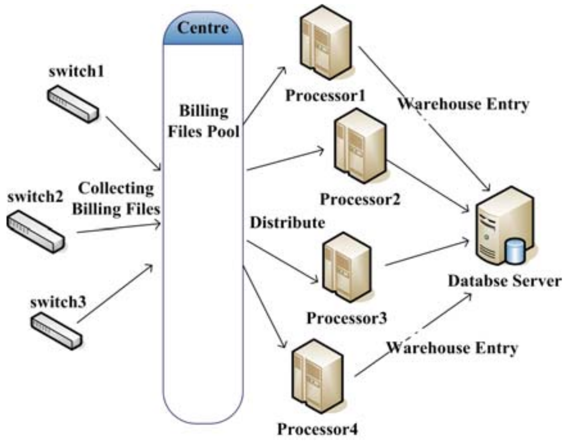


Figure 1 Sketch Map of the Billing Business Process. Note how the Billing System works

2.2 Issues of the system

How to use multi-processors more efficiently to process a huge numbers of billing files followed a strict sequence? How to design a model to meet the system requirements and reduce the response time of multi-processor?

2.3 Specifications

The specification of all the five key processes is as follows:

1) Collecting billing files

Billing Center will collect a lot of billing files from different switches, and then put them into Billing Files Pool. A billing file consists of several billing records from different cites or prefectures.

2) Task distribution

All the tasks will be distributed among multi-processors. Which are always finite.

The following three processes must be executed by a strict synchronization order.

3) Pretreatment

At this stage, processors will do a series of pretreatment preparing for the following processions.

4) Rating

Rating aims to read and price according to the billing records. In particular, we should note that different cites or prefectures may have a different rate when pricing. So a universal method is to process the billing records corresponding to different cites or prefectures.

5) Warehouse entry

The final results of the whole procession will be put into a data warehouse. This stage normally relates to options of connecting, searching and writing, so warehouse entry always costs a tripled time compared with the pretreatment and rating.

Neglected the stage of collecting billing files, we will focus our attention on the rest four stages and discuss a billing system model based on parallel processing.

3 Analyze and Design

3.1 Task distribution

After collecting billing files from switches, billing centre will read files and assemble records corresponding to different cites or prefectures in order to process rating more conveniently. Then the system will distribute and transfer records to processors.

Generally speaking, one of the effective measures is to partition the records into groups with a certain number. Suppose that there are a total of 100,000 records and 10 processors, so each processor may be distributed 10,000 records to be processed. But in fact, the number of processors may be less than 10, and different processors may have different performance. Taking into account this bottle-neck, the system can induct a method of dynamic loading balance, that is, choosing an integer named ‘n’ (0<n<10,000), so each processor may be distributed ‘n’ records to be processed, if one processor has finished its task, it can request the billing files pool for new records which is left in the pool to process. The number of ‘n’ should be chosen scientifically according to statistics or the experiments upon computer simulation.

Concrete method of the mind above can be realized by message transferring. The sketch map and pseudo-code are as follows:

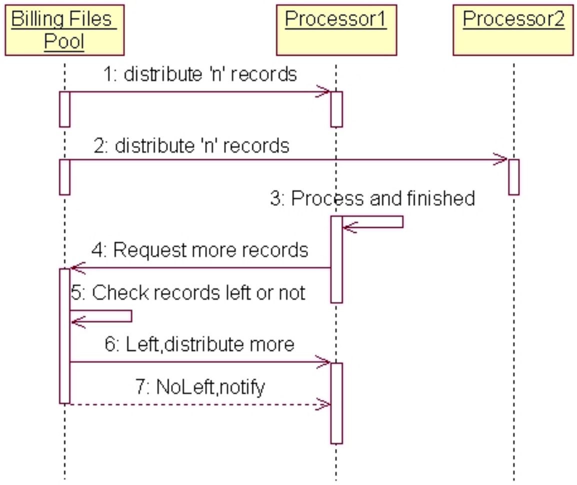


Figure 2　Sequence Diagram of message transferring in distribution.

The pseudo-code is described as follows:

```
Master Process:
//n is the certain number
int s=n;
//send 's' records to Slave Process
For(i=0;x=0;i<p;++i,x+=s)
    send(&Record[x],s,Pi);
//receive message from Slave Process
For(i=0;i<p;++i)
    Recv(&part,Pslave);
Slave Process:
Recv(numbers,s,Pmaster);
For(i=0;i<s;++i)
    doProcess();
send(&request,Pmaster);
```

3.2　Pretreatment, Rating and Warehouse entry

These three processions must be executed by a strict synchronization order. And the procession of warehouse entry may cost a tripled time compared with the pretreatment and rating. Both the method of Synchronization and the method of Pipelining can be used to meet the system requirements. Next we will discuss the availability of the two methods in order to choose a better one for our system model.

The method of Synchronization may need enough numbers of processors, because the procession is actually concurrent when executing. So enhancing the parallelism relies on the number and the performance of processors.

The method of Pipelining can solve the bottle-neck conveniently and efficiently. The model is as follows:

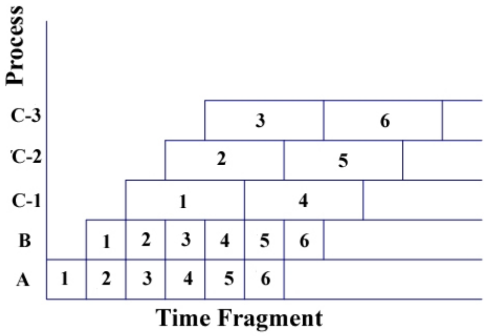


Figure 3　Sketch Map of The method of Pipelining

In Figure 3, 'A' stands for the procession of pretreatment, and 'B' stands for the procession of rating. The procession of warehouse entry is divided into three processions, which are separately marked by 'C-1', 'C-2' and 'C-3' in the figure. The whole processions can be executed more efficiently by this way.

4　Improvement

Based on the model above, the billing system can work very well from a coarse-grained perspective. But we can also improve the model from a fine-grained angle, such as database optimization, data structure, and so on.

Next we will discuss two improvements referring to warehouse and middleware. With the help of the two improvements, we can build a more wholesome and perfect billing system model.

4.1　Warehouse entry

The procession of warehouse entry always costs a tripled time compared with the pretreatment and rating. And the system enhances the parallelism by the method of Pipelining. Meanwhile, we can also improve performance more by fine-grained methods, such as the optimization of database.

The database can be designed with a lot of data sheets. Each data sheets stands for a city or a prefecture and store the records of the city or the prefecture.

Furthermore, there would be some imbalances among the data sheets. Takes Jiangsu Province for example, Jiangsu Province has a total 13 cites or prefectures, but there may have an imbalance among the amount of business, which leads to a different number of records in the data sheets. So some data sheets may have a normal number of records, and some other may have a huge number of records, such as the city of Nanjing, Wuxi and Suzhou. In these circumstances, the database can create subordinate data sheets for these cities, whose records are in a huge number. For example, Nanjing controls 11 prefectures and 2 countries, there are many people and a great amount of business in the district of Gulou, so we can create a data sheet named "Nanjing_Gulou" in order to share the pressure of warehouse entry. If the district of Gulou always has an imbalance, we can subdivide the data sheet once more, etc.

If needed query the records, we can use joint enquiries, or we can merge the data sheet and its subordinate data sheet so that we can query more easily and more quickly.

4.2 Using Middleware

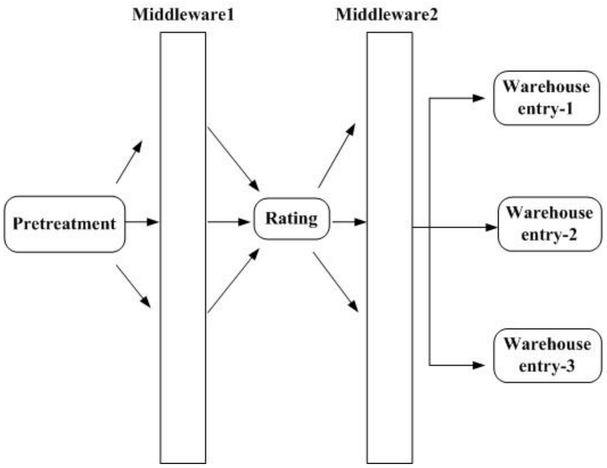


Figure 4 Using middleware to process more efficiently

Middleware is always known as a kind of computer software which connects software components or applications. It commonly consists of a set of useful and prepared services which allow multiple processes running on many machines to interact and process across a network. Several middleware models have been brought forward and been applied abroad, such as CORBA, EJB, COM+, and so on.

The middleware here means a kind of middleware with the responsibility to receive, classify and distribute records. It always contains a record pool and meanwhile it should do a lot of logic services.

The "Middleware1" in Fig. 4 will receive records after they're pretreated. And then the records would be classified according to cites or prefectures. The next procession, which is rating, may price the records with different rates, so distributed and processed by cites or prefectures will cut down the pressure of procession, and contribute to the efficiency and convenience.

The "Middleware2" in Fig. 4 is nearly similar to the "Middleware1", the records after rated would be classified and put into a record pool in the middleware. For example, the records of Nanjing would be put into a record list named "ListNanjing". The list would be distributed to the next procession only when the length of the list reaches a certain number. So a fixed-number list of records will be sent to be processed in order to reduce the pressure of warehouse, and this can promote the parallelism prominently.

If the length of a record list cannot reach the certain amount for a long time, the procession may fall into a deadlock. So we can set a timeout for the procession. The middleware will distribute the record list when waiting overtime as well. By this way, can the middleware work healthily and efficiently.

5 Conclusions

In this paper, we discussed about a billing system model based on parallel processing through using the parallel methods, such as divide and conquer, message transferring, loading balance, pipelining, etc. The model can meet the system requirements of Real-time Processing and reduce the response time of multi-processors. Meanwhile it can be expanded, improved and be more widely used in the telecom industry.

Acknowledgements

Thanks to the support of Fujian Fujitsu Communication Software Co., Ltd. (FFCS). With their support and help, this study works well and be processed in a smooth way.

References

[1] Prentice Hall and Pearson, Parallel Programming Techniques and Applications Using Networked Workstations and Parallel Computers, 2nd Edition, China Machine Press, 2005

[2] Jordan, H and Alaghband, G, Fundamentals of Parallel Processing, Tsinghua University Press, 2003

[3] Samuel P. Midkiff, Languages and compilers for parallel computing, Springer Press, 2002

[4] Hwang K, Xu Z, Scalable Parallel Computing: Technology, Architecture, Programming McGraw-Hill Companies, 1998

[5] Lin lin, Chen Weixin, Zhang Peng, Parallel TCP congestion control based on strength control, Journal of Computer Applications, Vol.28 ,No.4, 2008, pp.853~855

[6] Gong Mei, Wang Peng, Research and implementation of master-slave parallel file transfer system based on cluster of MPI, Application of Electronic Technique, 2007(11), pp.121~124

[7] Wang Wei, Yang Li and Liu Jianfeng, A New Billing System in the HPC Environments, Computer Engineering&Science, Vol.30, No.1, 2008, pp.148~150

[8] Huang Ling, Huang Haiming and Liu Jingang, The Investigation of Rating System Model Based on IP Data Business, Microcomputer Information, vol. 24, 2008, pp.218~220

[9] Liu Xiaojing and Liu Jianping, Telecommunication cost model design, China Water Transport, vol.5, No.1, 2007, pp.110~111

QoS based Distributed Multipath Routing and Admission Control Algorithm for IPv6

Muhammad Omer Farooq[1] Sadia Aziz[2]

1 Computer Engineering Department, Center For Advanced Studies in Engineering Islamabad, Pakistan
Email:1 amusecub_1983@yahoo.com; 2 sadia.aziz@case.edu.pk

Abstract

Multimedia applications running on Internet require minimum delay, minimum jitter and minimum packet loss. Various models such as IntServ and DiffServ try to satisfy QoS needs of an application. One of the common problems with these models is that they do not use any dynamic mechanism to find multiple paths to the destination network for routing QoS sensitive flows on a path that is the most appropriate for the QoS requirement of the flow. This can lead to a situation when we may need to deny admission to further QoS flows till some resources are made available. “Distributed QoS based Multipath Routing Algorithm for IPv6”, uses a QoS based Multipath Routing Algorithm (Qos-MPR) which can work alongside the traditional routing algorithm and it is able to construct multiple Label Switched Paths to a particular network. This enables us to inject more QoS sensitive flows in the network without deploying excess resources. QoS based Multipath Routing will not only allow to inject more traffic into the network but it will also provide efficient label switching mechanism using the “Flow Label” field of the IPv6 header. It can also serve as a simple load balancing mechanism. Simulations results have shown that QoS-MPR provides better packet delivery ratio and its QoS flows experience low delay.

Keywords: QoS routing, DiffServ, Admission Control, Label Switched Forwarding Table (LSFT)

1 Introduction

At present, majority of the Internet's infrastructure provides best effort services. Packets are processed as fast as possible but there is no guarantee about delay and packet loss. These issues did not represent any problem to elastic applications like Email, FTP and HTTP. But, in order to support real time traffic flows that require minimum delay, minimum jitter and packet loss we need to add extra functionalities to the Internet's infrastructure. Hence, providing QoS support in Internet is a challenging task. Last several years of work on providing QoS has resulted in development of two major models i.e. Integrated Services Model (IntServ)[1] and Differentiated Services Model (DiffServ)[2]. Multiprotocol Label Switching [3], which was initially intended for fast switching mechanism have now introduce traffic engineering features for providing QoS in IP networks.

Some Internet links are already presenting the signals of congestion. Reserving resources on such links can immediately result in denial of admission to real time flows. Such problem can be solved by deploying excessive infrastructure or by introducing traffic engineering solutions. First approach is contradictory to engineering perspective therefore, in this paper we present QoS based routing protocol that incorporates traffic engineering features.

We have devised a QoS based Multipath Routing (QoS-MPR) algorithm for QoS provisioning in the Internet. The routing protocol will be based on the DiffServ style for providing QoS and it will be able to meet QoS requirements of multiple real time flows. Since standard DiffServ does not provide any admission control mechanism therefore, we will incorporate the admission control procedure in QoS-MPR protocol which will also help to eliminate the signaling overhead. In traditional networks, QoS techniques depend on the underlying routing protocol for getting a path on which resources can be reserved. Such schemes tend to direct all flows for the same network on a single path. This can result in denial of admission control to some flows due to none availability of resources on that particular path. We present a scheme that can discover multiple paths to the destination and can direct flows with QoS guarantees on different (sub-optima but meeting minimum QoS specifications) paths when congestion blocks admission on the optimal path. The presented scheme will definitely alleviate the denial of admission problem; hence it will meet the QoS requirements of more flows as compared to the existing techniques. QoS-MPR maintains two forwarding table, one for forwarding best effort traffic and other one is the Label Switched forwarding Tables (LSFT) that will be used to forward traffic belonging to real time flows. Whenever, a source wants to initiate a QoS flow, it will generate a QoS-Route Request message regardless of the fact whether their exits a route to the destination in its forwarding table or not, in this model it is necessary that only destination generates a route reply message.

This paper has been structured in the following sections. We present the related work in section 2. QoS-MPR is presented in section 3. QoS-MPR algorithm is presented in section 4. Section 5 presents simulation results and we conclude this research in section 6.

2 Related Work

Last several years of work on the Internet Protocol for providing QoS have resulted in the two dominant models i.e. Integrated Services and the Differentiated Services. IntServ tries to provide hard QoS guarantees on per flow basis and for this purpose it uses the signaling protocol known as Resource Reservation Protocol (RSVP). RSVP uses the IP uni-cast and multicast routing algorithms, and it maintains the flow information i.e. flow label, minimum bandwidth, delay, and packet loss on each hop along the path. RSVP stores the state information in a soft state manner i.e. we periodically need to send keep-alive messages in order to keep reservations alive otherwise timeout will release the resources. Although IntServ reserves resources in a soft state but it can suffer from the scalability problem as the number of flows increase the state information that needs to be kept also increases and in this process the memories of routers/switches along the path may get exhausted. Secondly, since RSVP uses IP routing to reserve resources along the path, all the traffic to the same network will reserve the resources along a single path and the admission control of IntServ will deny admission to flows that are destined to the same network when no more free resources are available along the route. In most cases, there exist multiple paths to the same network and if we deploy a mechanism to dynamically discover these paths and then send traffic on the appropriate path, more traffic will be injected into the network. Hence, throughput of overall system will increase and applications will experience better QoS.

DiffServ[2] was invented to circumvent the scalability problem of IntServ by providing QoS provisioning on a class based granularity. DiffServ model has defined three main types of traffic classes i.e. Expedited Forwarding (EF) [4], Assured Forwarding (AF)[5] and best effort forwarding. Traffic requiring the guaranteed bandwidth, minimum loss, delay, and jitter are mapped to the EF class. The AF class is further divided into four classes and each class has three level of drop precedence, different flows are mapped to different categories of the AF classes depending upon their needs. The best effort class of DiffServ treats the traffic in the same way as IP does. Flows corresponding to different DiffServ classes are marked with a code point known as Differentiated Services Code Point (DSCP). In IPv4 "TOS" byte is used to mark the DSCP of a particular flow. In IPv6 this can be achieved by using the "Traffic Class" field. In the standard DiffServ model there is no admission control procedure defined in order to limit the traffic w.r.t. to the capabilities of the network. This limitation can degrade the performance of already established flows when excess QoS sensitive traffic is allowed into the network. There are many admission control procedures introduced for DiffServ in literature but these solutions have their own limitations. Secondly, similar to the IntServ model DiffServ relies on IP routing therefore, it inherits the same problem that we have already pointed out in our discussion of IntServ architecture.

IPv4 provides a rudimentary form of QoS by only providing the "TOS" byte in the IPv4 header, the limitation of the QoS provided by IPv4 can be removed by Internet Protocol version 6 (IPv6) in which two fields are exclusively reserved in the header for QoS i.e. Traffic class field and Flow Label field . Traffic Class field is primarily used to classify traffic related to various classes and afterwards each type of traffic is scheduled according to its priority, this field is typically used by the DiffServ model. Flow Label field is used to identify a particular flow among various other ongoing flows, such a field is primarily used by the IntServ style of providing QoS which is based on per flow granularity. We shall combine the use of both these field to come up with a very elegant solution for providing QoS and traffic engineering mechanisms.

Multiprotocol Label Switching (MPLS) [3] is another architecture that can provide some level of QoS though it was not the primary reason for its invention. In fact, it is a switching architecture that can route the flow in network with enhanced speed compared to the traditional longest prefix match lookup performed by the routers. MPLS builds a label switching table knows as Label Forwarding Information Base (LFIB), whenever a packet arrives that contains a label the MPLS enabled router will look for the label in its LFIB and swap the incoming label with the outgoing label. MPLS standard header contains a three bit field for QoS, this means MPLS can differentiate among eight service classes these eight classes of traffic are not sufficient to support varying QoS needs for different types of real time applications.

The constraint based Routing with the Label Distribution Protocol (CR-LDP) is an enhancement introduced to MPLS that performs traffic engineering functionality. (CR-LDP) treats QoS in more like an IntServ manner, therefore, it suffer from the scalability problem. For MPLS, new version of RSVP is defined which is known as RSVP-TE (TE stands for Traffic Engineer). In RSVP-TE a new object of EXPLICIT-ROUTE has been introduced. By using this option we can pin point the route that a flow should traverse in-order to reach its destination. The shortcomings of RSVP-TE is that for Explicit Routing we should first manually configure the labels on our known path and secondly, the problems that were reported for RSVP still exists with RSVP-TE.

In [8] Admission Control over Assured Forwarding PHB has been presented. This scheme does not seem to be a promising solution. If AF_2 packets (signaling packets) from two different sources manages to find a path and they have at least one common node in between, if that node can only accommodate one more flow since, none of the two flows have started their transmission their signaling packets will get through but when both of the sources begin to transmit congestion can occur hence performance of all the admitted flows will be effected.

In [9, 10] admissions schemes have been presented for DiffServ architecture. The admission control scheme presented in [9] suffers from the same problem that has been mentioned for [8].

3 QoS Based Multipath Routing

Our QoS-MPR is primarily composed of following three messages.

1. Route Request and Admission Control Phase.

2. Router Reply message 3. QoS lost message

In the following sub-sections we will elaborate on above listed three main components of QoS-MPR.

3.1 route request and admission control phase

In QoS-MPR enabled networks admission control is performed by maintaining the information on each node in a network about the number of flows of different DiffServ classes that the node can accommodate. Each node will allocate some bandwidth for different DiffServ classes depending upon the type of service a particular class can offer. Every node uses the following formula to calculate number of flows it can accommodate.

$$N = \text{Celing}(A_B/M_B) + C \qquad (1)$$

A_B = Total Allocated Bandwidth to a specific class

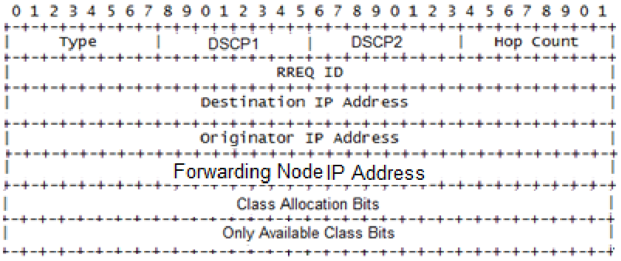
M_B = Maximum Bandwidth that can be allocated to a single flow in a particular class.

C = constant

We have added a constant number 'C' in the above expression in anticipating that it is possible that some of the flows will not utilize their full allocated bandwidth, this will allow some extra flows to enter into the network thus, by accepting more flows we can increase the total throughput of the network.

Every node will maintain a table that will contain the Differentiated Services Code Point (DSCP) and the corresponding numbers of flows that a node can accommodate for that particular class, every time a new flow is established the remaining flows entry will be decremented and when node receives no packet for already established flow for specified amount of time, time out will occur and the reservation will be deleted by incrementing the remaining flows field corresponding to the DSCP that a flow was using and an entry from QoS label Switch Forwarding table will be deleted.

The admission control phase starts whenever any node in the network wants to establish a QoS session with some other node. Source node will always send a route request message regardless of the fact whether there exist an entry into its forwarding table for that particular node or not. Following is the route request message of QoS-MPR which also incorporates the information necessary to perform admission control.



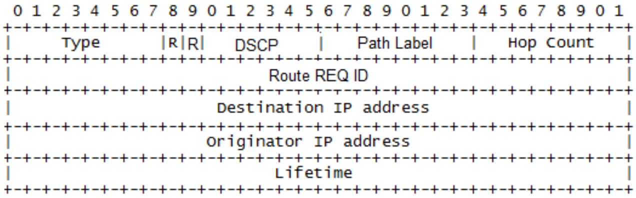
```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |     DSCP1     |     DSCP2     |   Hop Count   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            RREQ ID                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Destination IP Address                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Originator IP Address                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                  Forwarding Node IP Address                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                     Class Allocation Bits                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                   Only Available Class Bits                   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 1 QoS Route Request Message

For “Type” field the Route Request Message will have value of 1. We will elaborate the reason for DSCP1 and DSCP2 fields after elaborating the other fields of this route request message. The Hop Count field indicates the number of nodes that can forward this Route Request message. Every node will decrement this field by 1 at each visited node and when this field reaches to 0 the route request message will be discarded. The Request ID field will be a locally unique number to a node. The Request ID along with the originator IP address uniquely identifies the route request message of any node within the network. Every node that forwards this route request message will save the request ID, originators IP address and Forwarding node IP address so that it can match the route reply message with the actual route request message. The Destination IP address field gives the IP address of the node to which we want to find a route. Every node that will forward the route request message will insert its own IP address in the Forwarding Node IP address field so that route reply message gets back on the reverse route. The originating node will demand service for a particular DiffServ class, in DSCP1 field the node will set the DSCP of the DiffServ class that can easily accommodate the QoS requirements of the flow. It is possible that a node may not be allowed to establish the flow because no node in the path can accommodate a flow pertaining to DSCP1. In this case it is possible that nodes flow can serve its purpose if it gets forwarding treatment of some other DiffServ class that does not provide as stringent forwarding behavior as DSCP1. In DSCP2 field the DSCP of that particular class will be mentioned. We have introduced this feature to enhance the throughput of the overall network. The Class Allocation Bits are 32 in number, one class allocation bit corresponds to each DSCP, same is the case with Only Available Class Bits field. When an intermediate node receives route request message, it will check whether it can accommodate a single flow in all the classes ranges from DSCP1 and DSCP2, if yes then the node will mark the bit corresponding to DSCP2. If node can only accommodate the flow in a single class ranging from DSCP1 and DSCP2 then node will mark the corresponding bit in the Class Allocation Bits as well as in Only Available Class Bits field. If there is already a bit marked in the only available Class bits field then the node will discard the route request message. There is a possibility that on a particular node their exits some DiffServ classes that can accommodate the flow but these classes do not include classes corresponding to DSCP1 and DSCP2, in this case node will check whether any bit in Only Available Class Bits field is marked which is within the range of its available DSCP's if yes the node will mark the class allocation bit corresponding to that particular DSCP otherwise node will mark the bit corresponding to highest priority DSCP that it can accommodate in class allocation field and Only Class Available bits field.

3.2 Route reply message

Following is the format of the route reply message.



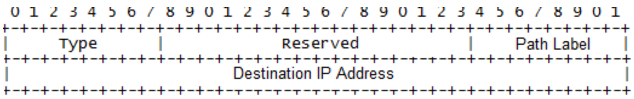
```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |R|R|   DSCP    |  Path Label   |   Hop Count   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         Route REQ ID                          |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Destination IP address                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Originator IP address                      |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                           Lifetime                            |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure 2 QoS Route Reply Message

In the above route reply message the type field will have a value of 2 to indicate that it is a route reply message. The next two bits are reserved for future use. The DSCP field will inform about selected service class for this particular flow. The label given in the Path Label field will be used by the upstream node to forward packets belonging to this particular flow to its downstream node. Every downstream node in the path will select locally unique label for the flow and inform about its selected label to its upstream node using route reply message. The Route REQ ID will contain the Request ID that was mentioned in the route reply message; the Route REQ ID along with the originator IP address will enable the receiver of the route reply message to match the reply with the route request message. The Destination IP address will contain the IP address of the node for which a route was required. The lifetime field will inform the receiver about the interval between two successive packets of the flow. If no packet is received within this interval the routing table entry pertaining to this particular flow will be deleted.

3.3 QoS lost message

The QoS route lost message has the following format.



```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |            Reserved           |  Path Label   |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    Destination IP Address                     |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Figure # 3 QoS Route Lost Message

In case of QoS Route Lost message the type field will have value 3. The path label field will have the label that an upstream node is using to send data to this downstream node. Using this Path Label field every intermediate node will send QoS route lost message to its upstream node until this message reaches the source.

3.4 Entries is QoS label switched forwarding path

Following entries will be maintained inside the QoS Label Switch Forwarding Table.

(i) Incoming Label (ii) Outgoing Label (iii) Next Hop

(iv) Destination IP Address (v) Destination Sequence no.

(vi) IP address of Upstream Node (vii) DSCP for this forwarding table entry.

3.5 Dual forwarding tables with load balancing

We have mentioned that QoS-MPR maintains two forwarding tables i.e. best effort forwarding table and Label Switched Forwarding Table. The challenge was to identify the mechanism to make the forwarding node aware that when to use the best effort forwarding table and when to use the Label Switched Forwarding Table. QoS-MPR uses a simple and elegant solution to this problem; in DiffServ we use IPv6 Class type field to identify the forwarding behavior of the packet. Whenever a packet arrives at any intermediate node, the node will check the Class Type field of the IPv6 header and if the value in the Class type field is other than the value that has been reserved for best effort forwarding then node will use the Label Switched Forwarding Table (LSFT) otherwise it will use the Best Effort Forwarding Table.

Using QoS-MPR when packet needs to be

forwarded using LSFT, we need to add the label that has been assigned by the downstream node for that particular flow, we decided to use Flow Label field of IPv6 header. Whenever, a packet will arrive at a node, the receiving node will check Class Type field of the IPv6 header and if the Class Type field does not correspond to the best effort forwarding, the node will extract the label from the flow label field of the IPv6 header, search for the outgoing label corresponding to the incoming label and switch the labels before forwarding the packets.

Since, we are maintaining two different tables whenever a QoS Reply is received, QoS-MPR tries to select the path that is not in use in the Best Effort Forwarding table and vice versa. We found this technique very handy for load balancing when there exists multiple paths to the destination.

4 QoS-MPR Algorithm

Whenever the route request arrives at any node, it will perform the following operations. Node will check the DSCP1 and DSCP2 fields of the route request message.

Node will check data structure that holds the information about remaining flows ranging from DSCP1 and DSCP2.

If every class ranging from DSCP1 and DSCP2 can allow a single flow, the node will mark the bit corresponding to DSCP2 in Class Allocation Bits field of the route request message. Node will update the remaining flows information for all these classes and it will also save this information in separate data structure that will be used to release the extra reservation when route reply message will be received. If no route reply message will be received timeout will occur and remaining flows filed corresponding to each class will be updated.

If node cannot accommodate single flows in all the classes ranging from DSCP1 to DSCP2 but it can accommodate a flow in some classes in between DSCP1 and DSCP2. In this case the node will check the Only Available Class Bits field of route request message and if any bit is marked and this node can accommodate a flow corresponding to that marked bit it will multicast the route request message to its direct neighbors, if the node cannot accommodate a flow in a class corresponding to which a bit is marked in the Only Available Class bits of the route request message then this node will drop the route request message. If none of the bit is marked in Only Available Class Bits field of the route request message then this node will mark the bits in both Class Allocation Bits and only available class bits field of the route request message.

If a node can only accommodate a flow in only a single class ranging in between DSCP1 and DSCP2, then the node will check the Only Available Class filed of the route request message and if any of the bit is marked and this bit does not correspond to a class in which this particular node can allow a flow then route request message will be dropped otherwise node will mark both Class Allocations Bits and Only Available Class Bits fields of the route request message corresponding to the available class and multicast the route request message to its direct neighbors.

When a destination receives this route request message it will first check the Only Available Class Bits fields of this message and if any bit is set the destination will reply to this route request message with DSCP field set to DSCP of the corresponding DiffServ class and if no bit is set in Only Available Class Bits field it means that every node can accommodate a flow in every class ranging from DSCP1 and DSCP2 in this case DSCP field of route reply message will be set to DSCP1. Destination will assign a label to this flows which will be locally unique to the destination and send this label back to the node from which it had received this route request message.

Every node on reception of route request message will assign a label to this flow and make an entry into their label switched forwarding table and they will also release the excessive resources using the DSCP field of the route reply message and data structure that has been maintained during the route request phase.

The entry into the label switched forwarding table will automatically get deleted when no packet for a particular flow arrives before time out occur in this case remaining flow field corresponding to this particular flow will also be incremented.

If duplicate route request message is received node will drop this duplicate route request message. Duplicate route request messages will be detected using Route Request ID and originator's node IP address.

4.1 QoS-MPR example

Let us consider the following network.

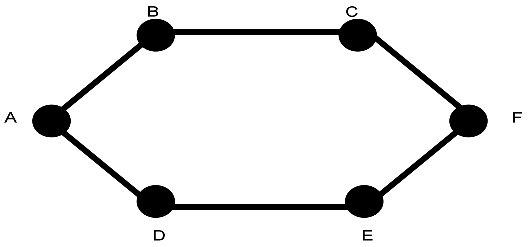


Figure 4 An Example Network

Suppose node 'A' wants to establish a QoS session with node F, with DSCP1 field set to Differentiated Services Code Point of EF and DSCP2 field set to Differentiated Code Point of AF_2, '2' refers to AF class with dropping level precedence of 2.

Following table shows the reservation status of different DiffServ classes at intermediate node B, C, D, and E.

Table 1 Reservation Status on Different Nodes

B		C		D		E	
DS	ID	DS	ID	DS	ID	DS	ID
EF	2	EF	1	EF	0	EF	1
AF1	1	AF1	1	AF1	1	AF1	0
AF2	1	AF2	1	AF2	0	AF2	0

When node 'A' multicast the QoS Route Request message with the stated parameters the request will reach to node 'B' and node 'D'. Node 'B' will check the DSCP1 and DSCP2 fields of the route request message and afterwards it will check its reservation status table. After checking its reservation status table node 'B' will conclude that it can accommodate request in any of the three classes EF, AF1 and AF2. According to QoS-MPR admission control procedure node 'B' will mark the bit corresponding to the AF2 class in "Class Allocation Bits" field and it will not mark any bit in the "Only Available Class bits" fields, because node 'B' can accommodate this flow in any of the three classes and these three classes fall within the range of DSCP1 and DSCP2. When node 'B' will multicast this route request message it will get to node 'C', node 'C' will perform the similar operation and it will mark bit corresponding to EF due to the reasons which we have described for node 'B' from node 'C' route request will reach 'F' which is actually the destination in this case. When node 'B' and 'C' processed the route request message they reserved one flow for EF, one for AF1 and one for AF2 and stored this information in a separate data structure, the access resources will be released when these nodes receive the DSCP of the class for which they actually need to reserve the resources in a Route Reply message and if no route reply is received within a predefined period of time all the resources will be released.

When Node 'A' multicast the route request packet it also reached to node 'D', when node 'D' consulted its resource reservation table it realized that it can only accommodate the request in AF1 class therefore; it will mark the corresponding bit in "Class Allocation Bits" field since this is the only class in which node 'D' can accommodate this flow it will also mark the corresponding field in the "Only Available Class Bits" field of the route request message. When this route request messages gets to 'E' through 'D', node 'E' will consult its reservation status table and it will see that it can accommodate this flow only in EF class, when node 'E' will go to mark the corresponding fields in the route request message it will notice that bit corresponding to AF1 in "Only Available Class Bits" is set this means that some prior node can only accommodate this flow in AF1 and node 'E' can only accommodate this flow in EF class hence, node E will drop the route request message. No Route Reply message will get back to node 'D' hence it will release its resources after some time.

Therefore, node 'F' will receive one Route Request from B → C → F and in this route request message no bit in the "Only Available Class Bits" field is set hence node 'F' will generate a reply with DSCP field in the route reply message set to DSCP1 field of route request message. When this route reply message traverses back to node 'C' and 'E'; they will not release resources corresponding to class EF and release resources corresponding to classes AF1 and AF2.

5 Simulation

In order to compare the performance of DiffServ based QoS-MPR with DiffServ model that use traditional routing algorithms we have performed two set of simulations. In these simulations we compared packet delivery ratio i.e. the number of data packets delivered to the destination and we have also compared the average delay of data packets.

5.1 Simulation environment

The simulations are carried out using NS-2 version 2.29.3. Figure 5 shows the topology of our simulated network. Each of our traffic sources are being mapped to Expedited Forwarding class. UDP is used as the transport layer protocol and the Constant bit Rate (CBR) traffic is being generated. Each data packet consists of 512 bytes and the interval between two consecutive data packets is configured is 0.008 seconds. Hence the bandwidth requirement for each flow is 500 Kilobits per second. Each link in the network has a bandwidth of 2 megabits per second. Random Early Detection (RED) with two colors marking is used as the queuing mechanism

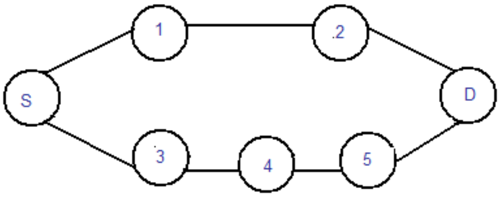


Figure 5 Simulated Network Topology

Node labeled 'S' is the source and node labeled 'D' is the destination of all the flows. All other nodes act as the core of the network.

5.2 Simulation results

From figure 5 it is evident that the packet delivery ratio of DiffServ based QoS-MPR is much better compared to the performance of the DiffServ using the traditional link state routing. We know that standard routing algorithms maintains only one route to the particular destination therefore; all the traffic to that network will be routed through that optimal path. When the offered load on the link increases packet delivery ratio will decrease thus, in such scenarios it is difficult to meet QoS guarantees of a flow. This problem decreases the packet delivery ratio of DiffServ when we are using traditional link state routing algorithm. QoS-MPR tries to find the multiple paths to the destination and then route the traffic of a particular flow on a path that best suits the requirements of a flow. Therefore, if multiple paths exist to the destination the performance of QoS-MPR is much better as depicted in figure 5.

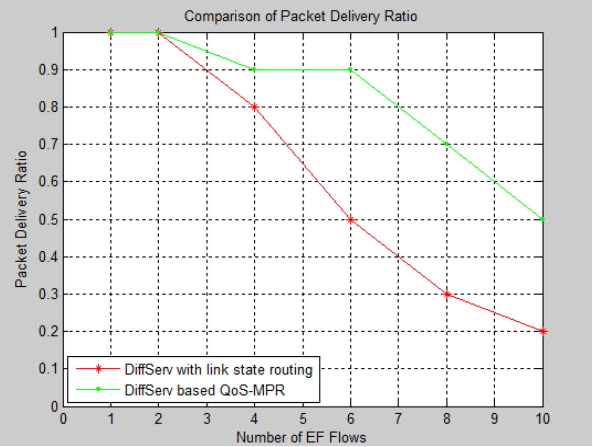


Figure 6 Packet Delivery Ratio

Figure 7 show the plot of the delay experienced by the Expedited Forwarding (EF) class traffic using link state routing and the DiffServ based QoS-MPR. When we simulated DiffServ using Link State routing, increase of single flow causes traffic in the network to experience large amount of delay. Because all the flows are directed on a single path this causes buffers along the paths to get full hence, delay increases. Since, QoS-MPR direct flows on multiple paths thus flows experienced low delay.

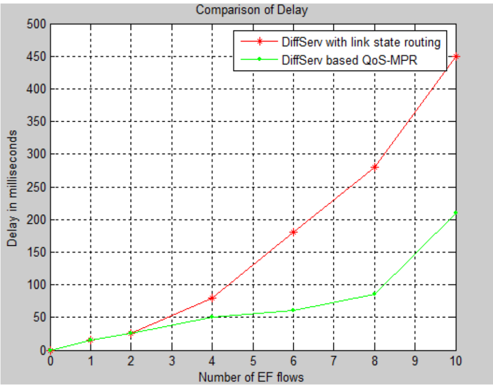


Figure 7 Delay Comparison

Figure 8 depicts a graph of "Number of Flows vs. Compliance with SLA. This graph is plotted by incorporating an admission control mechanism in QoS-MPR. We can observe that in the simulated network topology QoS-MPR has restricted number of EF flows to 8 and each admitted flow compliance with its SLA is above 80%. While DiffServ with Link state routing does not employ any distributed admission control mechanism hence, when number of QoS gets above to a certain threshold level there is an abrupt degradation in the quality of the admitted flows.

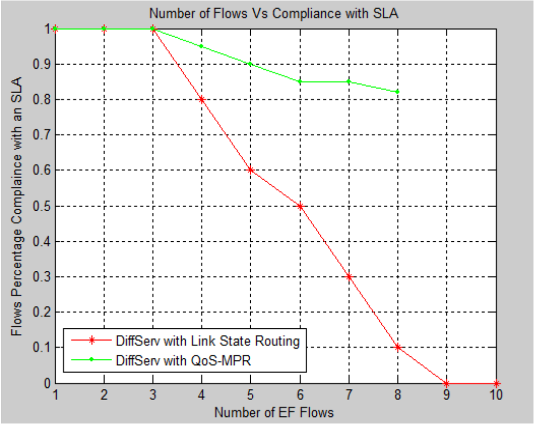


Figure 8 Number of Flows Vs Compliance with SLA

6 Conclusion

In this paper we have presented QoS based Multipath routing. QoS-MPR not only provides multiple paths that satisfy the QoS requirements of flows but it also provide an admission control mechanism for a DiffServ based network. Simulation results have shown that QoS-MPR meets the QoS requirements of all the admitted flows in an effective manner. It can also help to limit the traffic into the network so that QoS of already admitted flows remains intact.

References

[1] R.Barden, D.Clark, S.Shenker, "IntServ in the Internet Architecture an Overview" RFC 1633, June 1994

[2] K. Chan, J. babiarz, F. Baker "Aggregation of DiffServ Service Classes", RFC 5127 Feburary, 2007

[3] E. Rosen, A. Viswanathan, R. Callon "Multiprotocol Label Switching Architecture" RFC 3031, January 2001

[4] J. Bennet, Kent Benson, Angela Chiu, Sharama Davari, C.Kalmanek, D.Stiliadis, B.Davie, A. Charny, F.Baker, J. L.Boudec, W. Courtney, V. Firioiu K.K. Ramakrishnam, "An Expedited Forwarding PHB", RFC 2598, March 2002

[5] J. Heinanen, f. Baker, W. Weiss, J.Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999

[6] B. Jamoussi, R. Callon, R. Dandu, L. Wu, P. Doolan, N. Feldman, T. Worster, A. Fredette, M. Girish, E. Gray, J. Heinanen, T. Kilty, A. Malis "Constraint based LSP set using LDP", RFC 3212, January 2002

[7] D. Awduche, L. Berger, D. Gan, T. Li, V. Srinivasan, G. Swallow, RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001

[8] G. Bianchi, N. Blefari-Melazzi, "Admission Control over Assured Forwarding PHBs: a way to provide Service Accuracy in a DiffServ Framework", Global Telecommunications Conference San Antonio 29 Nov, 2001vol.4 pp. 2561-2565

[9] M. Li, B. Hoang, A. Simmonds, "Fair Intelligent Admission Control over DiffServ Network", 11th IEEE International Conference on Networks, pp. 501-506, oct, 2003

[10] E. Mykoniati, C. Charalampos, P. Geogatsos, T. Damilatis, Algonet, D. Goderis, P. Trimintzios, G. Pavlou, "Admission Control for Providing QoS in DiffServ IP Networks: The TEQUILA Approach",IEEE Communication Magzine vol.4 2003

Research on Communication Mechanism of Cooperated Multi-robot

Dong Zhao[1] Shixiong Zheng[2]

1 School of Mechanical Engineering, South China University of Technology, Guangzhou, Guangdong, 510640, China

Email:1 scutzd@sina.com ;2 mexzheng@scut.edu.cn

Abstract

How to effectively construct communication mechanism to solve the communicating problem of cooperated multi-robot's work has become the chief problem of the research on cooperated multi-robot. This paper deeply studies the communication system construction of multi-robot system, puts forward completely construction levels of communication system, and gives effective solution to the form level and transmission level of communication system.

Keywords：multi-robot; communication system; KQML

1 Introduction

In recent years, Cooperated Multi-robot System that based on multi-intelligence agent theory under distributed artificial intelligence has become the focus in the field of robot studying. It is also widely concerned by the scholars all over the world. Organizing and controlling multi-robot depending on the characteristic of Agent, it can complete complex works that single robot can't.

In the system of multi-robot, robots are in an unknown environment. How to obtain unknown environment information, draft the assignment and work cooperated to complete scheduled task are the core of study. Among this process, perception to the environment, planning and distribution of the assignments and its implementation need every intelligent Agent of the system to participate in communicating and negotiating. The communication and negotiation among multi intelligent Agents are the key point when the cooperated robots are dynamically running.

During the working process of cooperated multi-robot, however, the communication of information is frequent and massive no matter in planning and implementing of the task or in cognition to the environment. At the same time, the communication of information is often involved in more than one intelligent Agent, therefore it is necessary for them all to participate in together, and it will make the process of communication of information very complex. Especially for self-mobile multi-robot system, its communication usually uses wireless network, so the resources of bandwidth are limited. How to effectively construct communication mechanism to solve the communicating problem of cooperated multi-robot's work has become the chief problem of the research on cooperated multi-robot.

2 The communication system frame of cooperated multi-robot

The realization of the communication of multi-robot system is a complicated process. Generally speaking, the processing of information will experience the following steps when robots release that.

1) Generation of information: robots participate in the process of system cooperating, deal with related information according to the strategy and protocol of system cooperating task, and generate corresponding data to interact.

2) Formalization of information: when robots are interacting information, firstly they analyze information to make them to be a completed format that can be comprehended by all robots. Commonly the system will use CCL(Computation and Control Language) to formalize data.

3) Transmission control: the interaction of data is

based on the network communication system of multi-robot system. The key problem should be solved in this step is planning the communication priority and avoiding the clash during communicating in accordance with the requirements of cooperated task.

For robots at reception end, the processing of information is a converse process. The steps are like that, firstly receiving and processing information according to its priority, then analyzing the formalized data in accordance with grammar to make them to be information that can be cognized and processed by robots themselves, lastly doing cooperated action according to the strategy and rule of cooperated task. The communication process among Agents of multi-robot system is as following figure.

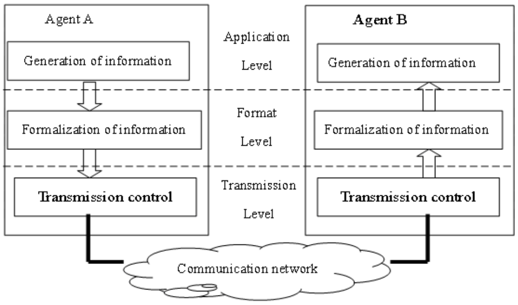


Figure 1

Figure1 The communication process of multi-robot system

The communication of cooperated multi-robot system can be divided into three level, they are application level, format level and transmission level. Every robot Agent communicates in this three level. The application level mainly involves in cooperating method and strategy of cooperated multi-robot system,

3 Format level design of expended KQML based

The informational formalization of multi-robot communication system is usually using multi-Agent communication language. Among languages used in multi-Agent communication, language Actor put forward by Hewitt is the typical application in information transmission. Geogeff, another scientist, uses "original language for communication" to avoid conflict between plans. Recently, some researchers bring forward some non-natural languages based on Speech-act, and make it as communication language among Agents. Among them, KQML （Knowledge Query and Manipulation Language ）is one used the most popular.

KQML is a communication language and protocol for realizing self-government, knowledge sharing and solution finding of cooperated problems of asynchronous Agents. It is the base that every Agent can work cooperatively in multi-Agent system. As knowledge query and operation language, KQML provides information format and its processing protocol for knowledge share when Agents are running, and defines Agent's possible operation to knowledge and storage objects through original extensible operation language collection.

To share the knowledge and information, KQML defines a set of reserved, basic, original operation language. Conceptually, a KQML message consists of three parts: a semantic action, content expression and a set of message description parameter. Through the sub expression of "content language" included in the message, and according to actual demand, developing high level interacted model among Agents, such as contract network and negotiation system to realize the irrelevance of KQML and the format of message itself.

The communication requirements of multi-robot are simultaneous, dynamic and Real-time. The message description parameters of KQML, however, only include basic description to receiver and sender, noumenon symbol and interactive requirements, etc. So it can't support high level, complicated and multi-round interaction on tasks and objects among robots. To meet this demand, new description parameters should be added to KQML.

Associate: describing the work of interactive message. Its basic format is:: associate<word>

In <word>, the standard of work is given. Through the setting of associate, Agent is able to know the host of the message when simultaneous communicating, consequently making the queuing of messages possible.

```
<performative>
:content <expression>
:aspect <expression>
:language <word>
:ontology <word>
:reply-with <expression>
:associate<word>
:priority<expression>
:sender <word>
:receiver <word>
```

Priority: describing its importance of the message in the work. Its basic format is::priority<expression>

The priority of the message is given in <expression>. Agent, according to the priority of work and message, will synthetically analyze to get the actual priority of every message and deal with the message queue depending on their priority.

Thus, the format of the basic information of an expanded KQML is as the following:

```
<performative>
:content <expression>
:aspect <expression>
:language <word>
:ontology <word>
:reply-with <expression>
:associate<word>
:priority<expression>
:sender <word>
:receiver <word>
```

It can be seen that through expanding KQML, the formalization problem of multi-robot communication can be solved well.

4 The design of transmission level

The communication among Agents in transmission level has three main modes: direct communication, intense signal communication and blackboard communication. Direct communication means that the sender Agent directly transmits message to receiver Agent, i.e. point to point communication. Other Agents of the system don't receive this message. Intense signal communication is that when an Agent sends message, it actually send the message to the whole system or surroundings around. Blackboard communication exchanges message through accessing a public data area, i.e. when any Agent sends message, it sends it to a public place, and the messages in this place are visible to all Agents.

For multi-robot system, due to its complication complexity and real time, communication is usually simultaneous in several modes. Using a good transmission strategy avoiding the conflict and effectively processing simultaneous messages is the core in transmission level research.

4.1 Determination of Priority

For simultaneous interactive process of the messages, an important precondition for the effective transmission of them is orderly processing message queue depending on the priority of them. For multi-robot system, determination of the priority of every message package is very difficult. Not only does the priority of the message itself (the priority within work, given by the parameters of the priority) be considered, but also the priority of work that message affiliated to and of Agents participated in be considered synthetically. In transmission level, the priority of the message can be obtained through Formula 1.

$$P_T = P_i C_i + P_a C_a + P_i C_i$$

$P_\mathrm{r}, P_i, P_c, P_1$ are transmission priority, message priority, priority of the Agent that message affiliated to, and priority of the work that message affiliated to, respectively.

C_r, C_i, C_c are the weighting coefficients of message priority, Agent priority and work priority respectively, given by the system dynamically.

4.2 Conflict Control

Wireless LAN (local area network) based criteria IEEE802.11 regulates the protocol standard of MAC (media access control) level, using CSMA/CD technology in MAC level. This standard regulates before the sending of network connection, every node needed sending should wait for vacancy on the channel. When the vacancy on the channel is detected, the node will send the data, then, it will wait for respondent confirmation. If the signal received within a certain time is not the respondent one, it is considered that conflict happened. Conflict means another node is transmitting message at the same time. Once conflict occurred, it probably mixes two signals, and the message transmitted to receiving node could easily be lost. And then every node should wait for channel vacancy, and delay the time period produced at random. And this delay should avoid another conflict between two nodes that are transmitting data.

Combined with the priority management of message transmission, the basic transmission control process of the message in transmission level can be illustrated like figure 2 (with the sending of message to be an example)

5 Conclusion

This paper deeply researches communication frame of the multi-robot system, brings forward a completed frame level and gives an effective solution to the realization of format and transmission level. However, as an application instance of multi-Agent system, multi-robot system involves the process to complicated distributing data, knowledge and controlling. Therefore, its stability, robustness and reliability when commu-nicating in an uncertain environment still need to be perfected.

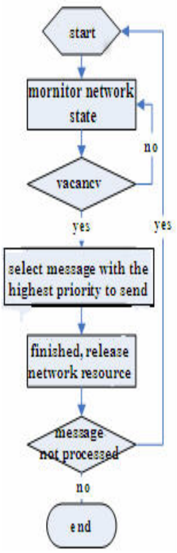

start
mornitor network state
no
vacancy
yes
yes
select message with the highest priority to send
finished, release network resource
message not processed
no
end

Figure2 The basic transmission control process of the message in transmission level

References

[1] R. Alur, R. Grosu, Y. Hur, V. Kumar, and I. Lee. Modular specifications of hybrid systems in CHARON. In Hybrid Systems: Computation and Control, LNCS 1790, pages 6{19, Pittsburgh, PA, 2000. Springer-Verlag

[2] A. Carruth. Real-time UNITY. Technical Report TR-94-10, University of Texas at Austin, 1994

[3] K. M. Chandy and J. Misra. Parallel Program Design: A Foundation. Addision-Wesley, 1988

[4] L. Cremean, B. Dunbar, D. van Gogh, J. Hickey, E. Klavins, J. Meltzer, and R. M. Murray. The Caltech multi-vehicle wireless testbed. In Conference on Decision and Control, Las Vegas, NV, 2002

[5] R. D'Andrea, R. M. Murray, J. A. Adams, A. T. Hayes, M. Campbell, and A. Chaudry. The RoboFlag Game. In American Controls Conference, 2003. Submitted for review

[6] E. W. Dijkstra. Self-stabilizing systems in spite of distributed control. Communications of the ACM, 17(11):643~644, November 1974

[7] T. A. Henzinger, B. Horowitz, and C. M. Kirsch. Giotto: a time-triggered language for embedded programming. In Proceedings of the First International Workshop on Embedded Software, LNCS 2211, pages 166~184. Springer-Verlag, 2001

[8] E. Klavins. Automatic compilation of concurrent hybrid factories from product assembly specifications. In Hybrid Systems: Computation and Control, LNCS 1790, pages 174~187, Pittsburgh, PA, 2000. Springer-Verlag

[9] E. Klavins. Automatic synthesis of controllers for distributed assembly and formation forming. In Proceedings of the IEEE Conference on Robotics and Automation, Washington DC, 2002

[10] E. Klavins. Communication complexity of multi-robot systems. In Workshop on the Algorithmic Foundations of Robotics, December 2002

[11] E. Klavins and U. Saranli. Object oriented state machines. Embedded Systems Magazine, pages 30~42, May2002

Design and Analysis of Multi-core Parallel Model for BP Inversion of Gravity and Magnetic Anomalies*

Junbao Xia Tao Li Qun Wang

School of Information Engineering, China University of Geosciences Beijing 100083, China

Email: xiajb@cugb.edu.cn, litaopier@163.com, qunw@cugb.edu.cn.

Abstract

During the procedure of the gravity and magnetic 3-D inversion with large scale data, BP inversion algorithm will cause large command of computation and storage. This paper gives an efficient solution based on the method of the multi-core parallel model and equivalent storage. That is, equivalent storage principle is used upon the position function of physical property units to lower its storage cost, while a parallel model of BP inversion algorithm is designed to increase its computation efficiency. Experimental results shows that this method's storage reduces from $O(\mathrm{k}n^4)$ to $O(kn^2)$ on an eight-core sever, and gets a speedup of 7.628, which is very close to the theoretical value of 8.

Keywords：L3-D inversion; Equivalent storage; Multi-core; Parallel

1 Introduction

It is becoming a hotspot for research and industry to maximize the processing power of multi-core processors. Thus, the software developers focus more and more on the parallel programming based on multi-core architecture.

In the gravity and magnetic anomalies processing domain, BP (Back Propagation) inversion of gravity and magnetic anomalies for 3-D physical properties is similar to neural network system, but it needn't know the training sample collections. And BP inversion neural network is composed of the input layer, the concealing layer and the output layer. Mostly it subdivides the underground magnetic properties into regular rectangular physical property units (magnetic intensity), which is considered as the concealing layer neurons. Besides, it gets the total measuring points from the number of neurons in the input layer and output layer. It makes those initial parameters act on those neurons of the input layer (the neurons of the concealing layer being forward calculation), and then the output layer is composed of output theoretic field values of measuring points. Meanwhile it makes the output layer feed back the input layer and then amends the magnetic property parameters of every physical property unit, which finally forms the inversion iteration. During the process of the BP inversion iteration, it will produce huge cost of storage and computing.

This paper briefly analyses the multi-core developing environment, the parallel design model and BP inversion, then gives an efficient magnetic anomalies BP inversion solution based on multi-core parallel design and the storage equivalence. Finally, the experimental results are given.

2 Parallel programming environment and design model

On multi-core computer, by changing some configuration options of the common integrated development environment-Visual studio 2005 Enterprise Edition for example, it will support the OpenMP parallel design model based on sharing storage.

* Beijing Municipal Education Commission Co-established Project

OpenMP is a multi-threaded parallel programming language oriented to shared memory and distributed shared memory multi-processor. It is an interface that explicitly guides parallel application programming with multi-thread and shared memory. And its programming model bases on thread-level. It can offer explicit paralleled through compiling guided statement, which supplies paralleling with full control. The architecture of shared memory multi-processor is shown as Fig.1. The implementation mode of OpenMP uses Fork-Join Model, in which there only exists a thread called "main thread" at the beginning and it will derive threads to do the paralleled task whenever needed. During the paralleled implementing, main thread and derived threads will work together. Then after implementing parallel code, those derived threads will quit or hang up in the mode of no longer working, thus the control process is back to the separate main thread. The Fork-Join Model is described as Fig.2.

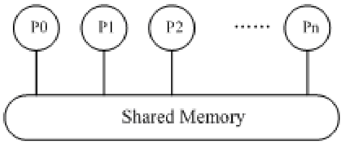


Figure1 Architecture of Shared memory multi-processor

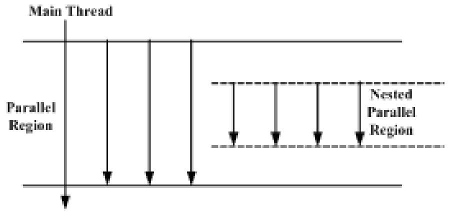


Figure2 Fork-Join Model in OpenMP

3 BP inversion model

3.1 Forward and inversion formula

The forward formula for the physical property units of the concealing layer is as below.

$$T_i = \sum_{j=1}^{m} J_i O_{ij} \qquad i=1,\ 2,\ 3\cdots\cdots M \qquad (1)$$

In Eq.(1), T_i stands for the magnetic anomaly output of the measuring point i. And M is the total measuring points. m is the number of underground physical property units. Thus J_i represents the magnetic parameters (magnetic intensity) of physical property units. For specific measuring areas, with the assumption of the magnetic azimuth being the same, O_{ij} is only related with the relevant position of the physical property unit j and the measuring point i. Thus we calls O_{ij} function position, calculated by Eq.(2).

$$\begin{aligned}
O_{ij} = &-k_1 \Bigg\|\Bigg| tg^{-1}\frac{(\eta-y)(\xi-x)}{R(\zeta-z)}\Bigg|_{x_j-b}^{x_j+b}\Bigg|_{y_j-L}^{y_j+L}\Bigg|_{z_j-h}^{z_j+h} \\
&+k_2 \Bigg\|\Bigg| tg^{-1}\frac{(z-\zeta)(\eta-y)}{R(\xi-x)}\Bigg|_{x_j-b}^{x_j+b}\Bigg|_{y_j-L}^{y_j+L}\Bigg|_{z_j-h}^{z_j+h} \\
&-k_3 \Bigg\|\Bigg| tg^{-1}\frac{(\xi-x)(\zeta-z)}{R(\eta-y)}\Bigg|_{x_j-b}^{x_j+b}\Bigg|_{y_j-L}^{y_j+L}\Bigg|_{z_j-h}^{z_j+h} \\
&+k_4 \Big\|\, l\,\mathrm{n}\big[R+(\xi-x)\big]_{x_j-b}^{x_j+b}\Big|_{y_j-L}^{y_j+L}\Big|_{z_j-h}^{z_j+h} \\
&+k_5 \Big\|\, l\,\mathrm{n}\big[R+(\eta-y)\big]_{x_j-b}^{x_j+b}\Big|_{y_j-L}^{y_j+L}\Big|_{z_j-h}^{z_j+h} \qquad (2)
\end{aligned}$$

In Eq.(2), k_1, k_2, k_3, k_4, k_5, k_6 are correlation coefficients of magnetic inclination of measuring district and magnetic declination. (x_j,y_j,z_j) is the center coordinate of the physical property unit j, b, L, h stands respectively for the half of the cuboids' length, height and width. And R represents the distance between the centre of the physical property unit and the measuring point i.

Amendment for the magnetic property parameters of every physical property unit in the concealing layer in every iteration process is the key part in the BP inversion. Assuming in the iteration k, the magnetic property parameter of unit j is $J_j(k)$, thus in the iteration $k+1$ the corresponding magnetic parameter will be $J_j(k+1)$, and the calculating formula of magnetic parameter amendment is as follows.

$$J_j(k+1)=J_j(k)+\Delta J_j(k)\ \ \Delta J_j(k)=\eta\sum_{i=1}^{M}(f_i-T_i)O_{ij} \quad （3）$$

In the Eq.(3), η is the learning step, *M* is the total number of measuring points, *k* is the serial number of the iteration time, f_i is the measuring value of measuring point *i*, and T_i is the theoretic output value of the measuring point *i*.

3.2 Equivalent Storage Model

With data scale increases, using Eq.(1) and Eq.(3) to calculate iterative inversion will conduct huge computing quantity. In order to make analysis facilitate, it assumes the ground measuring net being regular square grid with *n* rows and *n* columns and makes the division of physical property units in the concealing layer correspond with the measuring point grids (subdivision is k layers, each layer is with *n* rows and *n* columns). From Eq.(1) and Eq.(3), it can be known that every time to calculate the forward field value or amend physical property parameters will have to calculate O_{ij} kn^4 times. Thus the algorithm complexity will be $O(kn^4)$. At the same time, to calculate O_{ij} will take many times of complex computing such as logarithm, arc-trigonometric function, evolution and square, which will cause the computing quantity extremely gorgeous.

Since O_{ij} only relates with the relevant position of the physical property unit and the measuring point, it will greatly reduce the calculation cost of O_{ij} brought by the Eq.(1) and Eq.(3) by beforehand calculating all the O_{ij} involved once and then storing, but on the other hand, it will also cause huge storage cost, which will achieve $O(kn^4)$. Assuming that *n* is 100, *k* is 10, and O_{ij} is stored by single precision float point type, the total storage cost will achieve $4\times 100^4\times 10$B, almost 4GB.

Further analyzing O_{ij}, we found there is a lot of redundant data, defining $O_{k,l,m,p,q}$ is the position function of the physical property unit of row *l* and column *m* in the underground layer *k* to the measuring point of row *p* and column *q*, it will have translation equivalence.

Magnify the actual measuring area 4 times; calculate the position function O' of each measuring point in the virtual measuring area and each first physical property unit in the each underground layer. Assuming the actual measuring area has *x* rows and *y* columns, the position function contains the following equivalent relationship:

$$O_{k,l,m,p,q}=O'_{k,1,1,p+x-l,q+y-m} \quad （4）$$

According to this principle of equivalence, the storage requirement of position function changes into $O(kn^2)$. Thus the Eq.(1) and Eq.(3) turn into simple table-consulting operation, which makes large-scale data inversion possible.

4 Parallel implementation of BP magnetic anomalies inversion algorithm

After applying the storage equivalence to the BP inversion algorithm, its time complexity still remains O(kn^4), which is the main bottleneck to constrain the inversion data scale. The experimental data is with 500 rows and 500 columns. And the underground physical property unit is divided into 1 layer, each of which is also divided into 500 rows and 500 columns. One iteration will cost approximately 970 seconds running on the PC with Celeron 1.8G and memory of 512M. With the scale of measuring area increasing, the iteration time will grow to quartic. Parallel computing technology based on multi-core provides a fundamental solution to this problem, which is also an efficient solution with low cost. In this section, it will describe the Parallel BP inversion algorithm model implemented through OpenMP under the environment of Visual Studio 2005.

In the BP magnetic anomalies inversion serial iterative algorithm, three tasks execute in serial order: (1) make current physical property parameters act on physical property units in concealing layer, and calculate the theoretical output field value; (2) feedback theoretical field value and measuring field value to the parameter amendment of physical property units in the concealing layer; (3) calculate the iterative error and determine whether it reaches to the convergent limit.

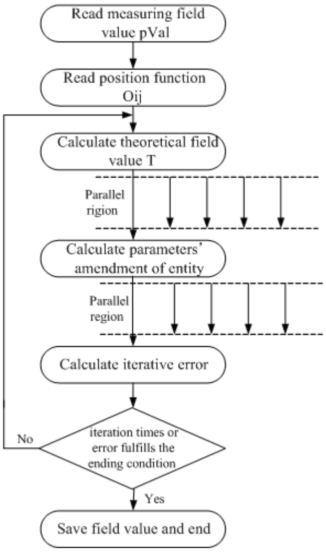

Read measuring field value pVal
Read position function Oij
Calculate theoretical field value T
Parallel rigion
Calculate parameters' amendment of entity
Parallel region
Calculate iterative error
iteration times or error fulfills the ending condition
No
Yes
Save field value and end

Figure3 Algorithm Implementation

Through searching for parallel region in BP magnetic anomalies inversion algorithm, several possible parallel regions of this algorithm could lay in the following aspects: read measuring field value into memory; read position function into memory; calculate theoretical field value; calculate parameters' amendment of entity and calculate iterative error. By analyzing the running time of those possible parallel regions listed above, calculating theoretical field value and parameters' amendment of entity are the most important two aspects.

After that, we adjust the scheduling strategy to dynamic scheduling for our algorithm to avoid the load imbalance. That is, in those parallel regions, divide the cycle section into several parallel cycle blocks, each of which is executed by one thread. And after one cycle block is finished, the thread will execute another cycle block through scheduling. By testing, alter the number of cycle times of each cycle block to 1 will further shorten the running time of this procedure.

Our paralleled algorithm is implemented as shown in Figure3.

5 Experimental Results

We have implemented the parallel algorithm described in Section 4. Using the following data (Measuring data 500*500, concealing layer 1, physical property unit 500*500, the physical property unit parameter 100), we tests the time cost of the serial and parallel inversion algorithm in different environment with different configuration. The test environment mainly includes a microcomputer (single core, Celeron 1.8G, 512M memory, Operating System: Windows 2003) and a multi-core server (8*Xeon 5310@1.6G, 2G memory, Operating System: Windows Server 2003 Compute Cluster Edition Service Pack 2). The test result is shown in table 1, while the time cost stands for running iteration once with different number of cores.

Table1 Test result

	Micro-computer	Multi-core Server			
		1 core	2 cores	4 cores	8 cores
Time cost of iteration 1 (s)	970	717	360	184	94
Speedup	----	1	1.992	3.897	7.628
efficiency	----	1	0.996	0.974	0.953

On the server with two dual-core CPUs, the iteration using 1 core, 2cores, 4 cores and 8 cores will get the speedup diagram listed below as Fig.4. Through observation, the speedup of this algorithm is very close to the theoretical speedup with high level parallelism. Through using the corresponding number of cores, the performance of this algorithm is almost with linear growth. The algorithm can achieve such result mainly due to itself has a high level parallelism. And the reason it can not reach the theoretical value is that making code paralleled will also bring some system cost.

Treat the forward theoretical magnetic anomalies as observation data, which is produced with all the physical property units' parameter is 100. Set the initial magnetic parameter to 50 of all the physical property units and learning step to 0.8. Then after 200 iterations on the server with two quad-core CPUs, it will produce inversion result contour map shown as Fig.5. From the figure we can see the parameter value of vast majority physical property units is very approximate the true value, which also shows that the correctness and effectiveness of the parallel inversion algorithm. 200 iterations' time cost is about 5 hours, while it will need more than 50 hours running on the microcomputer with one core to finish 200 iterations.

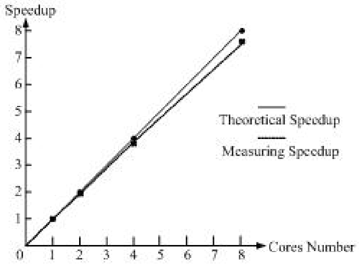


Figure4 Theoretical and Measuring speedup

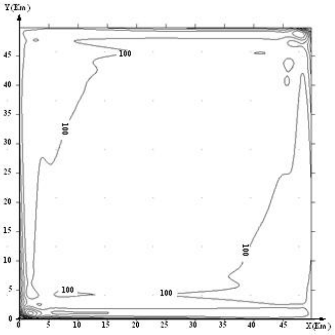


Figure5 Inversion result on Server with two quad-core CPUs

Using 8 cores on the server with two quad-core CPUs, the speedup reaches 7.628. Compared with the situation using one core, the solving time will be shortened to 13.1% if given the same data scale; while the problem scale that can be solved will be increased by approximate 68.1% if given a limited time to solve the problem.

6 Conclusion

This paper presents an efficient solution based on the method equivalent storage and the multi-core parallel model to do the BP inversion. The result is very approximate the theoretical parallel speedup. This solution will be the fundamental of fast inversion of large scale data. Experimental results show that the solution can effectively increase the inversion speed or enlarge the data scale that can be processed. With the popularity of multi-core server with lower cost, this solution will have a board application prospect in the fields such as nonlinear 3-D physical property inversion and so on.

References

[1] Zhining Guan, Junsheng Hou, Linping Huang. Inversion of gravity and magnetic anomalies using pseduo-BP neural network method and its application[J]. Chinese Journal of Geophysics, 1998, 41(2):242-251

[2] Changli Yao, Tianyao Hao, Zhining Guan. High-speed computation and efficient storage in 3-D gravity and magnetic inversion on genetic algorithms[J]. Chinese Journal of Geophysics, 2003, 46(2):252-258

[3] Zhining Guan, Tianyao Hao. Prospect of gravity and magnetic exploration in the 21st century[J]. Progress in Geophysics, 2002, 17(2):237-244

[4] Changli Yao, Yuanman Zheng, Yuwen Zhang. 3-D gravity and magnetic inversion for physical properties using stochastic subspaces[J]. Chinese Journal of Geophysics, 2007, 50 (5): 1576-1583

[5] Xizhe Geng, Yanhong Ding. The prospect of applying wavelet neural network to the inversion of gravitational and magnetic data[J]. Geophysical and Geochemical Exploration, 2001, 25(2):102-108

[6] Zhining Guan, Tianyao Hao, Changli Yao. Restrictions in gravity and magnetic inversions and technical strategy of 3d properties inversion[J]. Geophysical and Geochemical Exploration, 2002, 26(4):253-257

[7] Shameem Akhter, Jason Roberts. Multi-core programming: Increasing performance through software multi-threading[M]. Beijing: Publishing house of electronic industry, 2007

[8] Intel Corporation, Intel® 64 and IA-32 Architectures Software Developer's Manual[M], Volume1: Basic Architecture[M] 2006. http://www.intel.com/design/processor/manuals/253665.pdf

[9] Intel Corporation, Intel® 64 and IA-32 Architectures Software Developer's Manual[M], Volume2: System Architecture[M]. 2006. http://www.intel.com/design/processor/applnots/317080.pdf

[10] Intel Corporation. Intel® 64 and IA-32 Architectures Optimization Reference Manual[M]. May, 2007. http://www.ntel.com/design/processor /manuals/248966. pdf

[11] Ananth Grama, Anshul Gupta, George Karypis et al. Introduction to parallel computing (second edition) [M]. Beijing: China Machine Press, 2004

A Parallel Solver for Diagonally Dominant Tridiagonal Linear Systems With Constant Synchronization Requirements*

Xiping Gong Junqiang Song Lilun Zhang Wentao Zhao Jianping Wu

School of Computer Science, National University of Defense Technology, Changsha, China, 410073
Email: gongxpjia@163.com

Abstract

We propose a parallel solver which only requires one synchronization for diagonally dominant tridiagonal linear systems of equations. The amount of data transmitted in the only one communication round is related to the number of processors and independent of problem size (n). The parallel solver only needs about $23n$ float operations and one communication synchronization totally. The computation procedure can be dived into two parts: one we called pretreatment needs about $13n$ float operations and the other one we called kernel solver needs about $10n$ float operations. In addition to showing its theoretical complexity, we have implemented this algorithm on a real distributed memory parallel machine. The results are very promising and show an almost linear speedup for large n indicating the efficiency and scalability of the proposed solver. In theory it can be used to solve arbitrary diagonally dominant tridiagonal linear systems correctly and efficiently.

Keywords: parallel solver; tridiagonal linear systems

1 Introduction

Solving tridiagonal linear systems is a fundamental problem useful throughout computational science and engineering. Due to the large-scale nature of many problems in computational science and engineering, parallel tridiagonal linear solvers should, and have been investigated. Designing and analyzing tridiagonal solvers is clearly an important area of research. Nowadays, many direct factorization methods fitting for parallelization have been developed, including the Stone's “scan-based” algorithm [1], ”odd-even cyclic reduction” [2] and “partitioning” [3][4]. At the first glance, these methods seem to be efficient. But for distributed memory parallel computing, they do not scale well for needing more synchronization points. Therefore these methods often give disappointing speedups when implemented on real parallel machines just because of the high communication requirements. The original motivation for this paper are some intriguing questions arising in the parallel solution of tridiagonal linear equations one of which is to reduce communication. Xiping Gong el. [5] given a parallel algorithm which needs one global communication with a pretreatment. Although the pretreatment can be paralleled efficiently, it needs $O(135n)$ (n is the size of the problem) float operations to get Np (number of the processors) directions and the solver needs $O(13n)$ float operations. Therefore the parallel method can't suitable the situation which solve the tridiagonal equation only once.

In fact, the effort to reduce communication is centered on reducing the number of communication rounds. In this paper, we take account of the special characteristic of the tridiagonal matrix, and present an efficient parallel solver for solving diagonally dominant tridiagonal systems. The solver needs almost twice of float operations to LU factorization method. We also demonstrate how to implement the method on a parallel computer to obtain high efficiency. It should be

* This work was supported by National Nature Science Foundation under grant number 40505023.

emphasized that in this paper we adopt a purely algebraic viewpoint for the reason that our intention is to illuminate the algebraic structure that enables parallelism. Other numerical behavior of the new method, such as stability, will be discussed in future work.

The paper is organized as follows: In section 2 we will present the parallel algorithm. Algorithm analysis and computational complexity will be discussed in Section 3. Numerical experiments are given in section 4. In Section 5, we finally get a conclusion and some remarks for the parallel solver**.**

2 Derivation Of The Parallel Solver

2.1 Problem

A diagonally dominant tridiagonal linear system of equations is described by the tridiagonal matrix A with coefficients shown below,

$$A = \begin{bmatrix} a_1 & b_1 & & & \\ c_2 & a_2 & b_2 & & \\ & c_3 & a_3 & \ddots & \\ & & \ddots & \ddots & b_{n-1} \\ & & & c_n & a_n \end{bmatrix}.$$

The system of equations associated a tridiagonal matrix A is $Ax = r$ (1) $x, r \in \Re^n$ are column vectors and matrix A is diagonally dominant.

2.2 Data partition

Assuming that the tridiagonal linear system is distributed on a number of processors, so that each processor owns a contiguous part of rows. The processors have topological line structure and between every two neighboring processors there is one point we denote as black circle, other points are denoted as white circle and the points allocated to one processor are included a red circle just as shown in Figure 1. All data can be arranged just like Figure 1 and there have $N_p - 1$ black circles and $n - N_p + 1$ white circles.

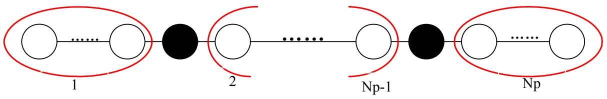


Figure 1 Data partition

The parallelism of the partition is N_p which is the number of the processors. The points can be arranged two sets, one is in the red circles and the other is not or one is white points and the other is black points. Every red circle is allocated to one process and every processor have a part of the tridiagonal matrix A from equation k_1 to equation k_2. The local coefficient matrix allocated to it can be denoted as $\left[\bar{A}_1, \bar{A}_2, \bar{A}_3\right]$ where

$$\bar{A}_1 = A^{k_1:k_2, k_1-1}, \bar{A}_2 = A^{k_1:k_2, k_1:k_2}, \bar{A}_3 = A^{k_1:k_2, k_2+1}$$

and superscript index the range of the matrix.

The right term correspondence of the equation k_1 to equation k_2 can be denoted as $\bar{r} = \begin{bmatrix} r_{k_1} & \cdots & r_{k_2} \end{bmatrix}^T$.

Obviously, if we can get a formulation that the values of the white points be described by the values of the black points, after got the values of the black points, the values of the white points can be computed directly just using the formulation. How to get the formulation and compute the values of the black points are the keys of the parallel solver. Next in section 2.3 we will introduce how to deal with the original linear equations in order to obtain the formulation and the black points' values.

2.3 Describe the Parallel solver

In fact from equation k_1 to equation k_2, we can get

$$\bar{A}_2 x^{k_1:k_2} = \bar{r} - x_{k_1-1}\bar{A}_1 - x_{k_2+1}\bar{A}_3 \qquad (2)$$

in which x's subscript denotes a element of x.

Let matrix A is diagonally dominant so $\bar{A}_2$ is nonsingular, so everyone linear system of the multi-right terms linear systems

$$\bar{A}_2 Y = \begin{bmatrix} \bar{r} & \bar{A}_1 & \bar{A}_3 \end{bmatrix} \qquad (3)$$

has only one solutions. In order to describe parallel solver simply we can denote $Y = Y^{k_1:k_2,3}$.

After solved multi-right terms tridiagonal linear systems Eq.(3), just using the Eq.(2), we obtain

$$x^{k_1:k_2} = Y^{k_1:k_2,1} - x_{k_1-1}Y^{k_1:k_2,2} - x_{k_2+1}Y^{k_1:k_2,3} \qquad (4)$$

Substitute x_{k_1} in Eq.(4) into the original equation k_1-1 and x_{k_2} in Eq.(4) into the original equation k_2+1 there are a new tridiagonal linear systems with the unknown variables x_{k_1-1} and x_{k_2+1} which are denoted as black points in Figure 1. In the new tridiagonal linear systems there are $Np-1$ unknown variables and the new tridiagonal linear systems could be denoted as

$$A_{new}x_{new}=r_{new} \tag{5}$$

2.4 Description of the parallel solver

Suppose we only need to solve the diagonally dominant tridiagonal linear system once, the parallel algorithm consists of the following five steps:

1) Each processor solves the multi-right terms tridiagonal linear systems Eq.(3).

2) Every processor computes the every terms of tridiagonal linear systems Eq.(5) according parts. For every processor if there has equation k_1-1 according black point then substitute x_{k_1} into it and if there has equation k_2+1 according black point then substitute x_{k_2} into it. At the end we can get the according parts.

3) Communicate to each other in order to get all the right term of the linear system Eq.(5) on every processor.

4) Each processor uses suitable method to solve linear system Eq.(5).

5) Each processor uses Eq.(4) to compute these white points' values in the red circle allocated to it.

By now we have obtained the parallel solver correctly.

3 Analysis of the Parallel Solver

From 2.4, we needn't care about 2)-4) because of always $Np<<N$. We can see the first step 1) only needs O(18n) float operations just using LU factorization method to solve linear systems Eq.(3) totally and the step 5) only needs O(5n) float operations. So the parallel solver needs about $23n$ float operations. This result is more greater than $21n$ [3] and 17n [4]. Usually we need to solve the diagonally dominant tridiagonal linear system Eq.(1) repeatedly where the coefficient matrix is fixed and only the right term is changed from time to time. So we can get the common part as a pretreatment which include LU factorization $\overline{A}_2=LU$, solving $LU\overline{Y}=\left[\overline{A}_1 \quad \overline{A}_3\right]$ and compute coefficient matrix A_{new} and its' LU factorization $A_{new}=L_{new}U_{new}$ in Eq.(5). At different time every processor only needs to solve $LU\overline{Y}=\overline{r}$, to get the right term r_{new} of Eq.(5) and solve $L_{new}U_{new}x_{new}=r_{new}$ and lastly using Eq.(4) to compute these white points' values in the red circle allocated to it. We call the common part as pretreatment and the residual part as kernel solver. Obviously the pretreatment only needs about $13n$ float operation and the kernel solver only needs about $10n$ float operations. The parallel solver is more acceptable for it only needs little float operations and one communication. Because we are willing to see the parallel efficiency, so we firstly implement the parallel solver just like 2.4 completely and the numerical experiments are based on it.

4 Numerical Experiment

In order to show the efficiency of the parallel algorithm, we take a tridiagonal linear system Eq.(1) that the coefficients are $a_i=5, b_i=1, c_i=1, i=1,2,\cdots,n$, the true solutions are $x_i^*=1, i=1,2\cdots,n$ and all real numbers are double precision. The numerical experiment is implements on 7 nodes HP Cluster. Every node be connected with Ethernet and every node has 4 CPU. In the experiment $n=2^{20}$, the number of the processors changed from 2 to 11. Table 1 is the execute time (10^{-2} second).

Table 1 The execute time of n=2^{20}

N_p	time(10^{-2}s)	N_p	time(10^{-2}s)
2	13.077	7	3.8347
3	9.1936	8	3.3375
4	6.7533	9	2.9873
5	5.3457	10	2.7166
6	4.4139	11	2.4804

Table 2 The speedup of $n=2^{20}$

N_p	time(10^{-2}s)	N_p	time(10^{-2}s)
2	1.0000	7	3.a4102
3	1.4244	8	3.9182
4	1.9364	9	4.3775
5	2.4463	10	4.8137
6	2.9626	11	5.2721

We defined the speedup is $speedup = T(2)/T(Np)$, table 2 gives the speedup when $n = 2^{20}$ at different number processors. From table 2 we can see the parallel solver almost have complete speedup that is the radio between the number of processors used and 2. There are some factors what affect the performance of this parallel algorithm. Using more processors we need to compute bigger tridiagonal linear systems Eq.(5), so obtain Eq.(5) and solve it needing more time. Although the parallel solver needs one communication, but as the number of processors increasing, it needs more time to communicate.

Till now, we have implemented the parallel solver correctly just as 2.4, getting some computing part as pretreatment needs more care to implement until now we haven't realize it correctly. We will continue to investigate the performance of this parallel solver carefully.

5 Conclusion

In this paper, we presented a parallel solver for diagonally dominant tridiagonal tridiagonal linear systems. Required the coefficient matrix is diagonally dominant is just for the multi-right terms linear systems Eq.(3) have determinate solutions. The parallel solver can be used for the direct solution of an arbitrary diagonally dominant tridiagonal linear system in theory. We have analyzed in detail the implementation of our parallel solver, from which it can get obvious advantage for solving the tridiagonal linear system with the same tridiagonal coefficient matrix repeatability. From the analysis, the parallel algorithm at most needs $23n$ float operations and one global communication round that gain an advantage over most parallel algorithms for solving tridiagonal linear systems on large parallel machines. If we get some computation as pretreatment just as described in section 3, the pretreatment only needs $13n$ float operations and every time the leaving computation is only about $10n$ float operations. It will be more worthwhile to do when we need to solve tridiagonal linear systems repeatedly and they have the same coefficient matrix. The case is usually in practice. In the future, we will continue to improve the performance of the parallel solver. Our aim is to adopt our algorithm, to implementation on large scale parallel computers, by increasing the adaptability, dependability, and scalability of the solution methods and to extend the parallel solver to analogy parallel solver for diagonally dominant and general narrow-banded linear systems.

References

[1] H.S. Stone, An Efficient Parallel Algorithm for the Solution of a Tridiagonal Linear System of Equations, JACM, 20, 1973, pp.27-38

[2] R.W. Hockney, A Fast Direct Solution of Poisson's Equation Using Fourier Analysis, JACM, 12, 1965, pp.95-113

[3] H.H. Wang, A Parallel Method for Tridiagonal Equations, ACM Trans. Math. Software, 7, 1981, pp.170-183

[4] Michelse P H, Van der Vorst H A. Data transport in Wang's partition method, Parallel Computing, 7, 1988, pp.87-95

[5] Gong Xiping, Song Junqiang el., A Parallel Algorithm for Solving Tridigonal Linear Systems, DCABES 2007 PROCEEDING, Hubei Science and Technology Press, Wuhan China,pp.19-22

Research of Application for Distributed OLAP System Based on Hybrid Scheduling Scheme*

Lijun Wang

Economics and Management Department, North China Electric Power University, Baoding, Hebei, P.R.China
Email:wljazg2008@yahoo.cn

Abstract

The technology of Data Warehouse and OLAP provide well analysis condition and present data means for enterprises [1]. But the OLAP system with only one server is inefficient and expensive; it doesn't meet the needs of most middle and small-sized enterprises. In this paper, we introduce distribute technology into OLAP system, so the system can assign a mount of query and analysis tasks into different servers. In order to improve the operating efficiency of the distributed system and avoid complex design, we propose a hybrid scheduling scheme that uses centralized scheme and stable sender-initiated algorithm. At last we analyses the performance results of the distributed OLAP system.

Keywords: OLAP, Distributed system, scheduling scheme

1 Introduction

On-Line Analytical Processing (OLAP) is a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user[2]. The OLAP technology has already been applied commonly in large company, it supports enterprise decision effectively. At Present, almost all the OLAP systems are based on only one server. They are excessively expensive to deploy for most middle and small-sized enterprises. By introducing distribute technology into OLAP the system can assign tasks into different servers, then the OLAP application can run on servers with low configuration or microcomputers. Thus it can make use of available equipment to deploy OLAP system, reduces business cost effectively for company.

In distributed systems, commonly used scheduling scheme includes centralized scheduling and distributed scheduling scheme [3]. In the centralized model, the scheduler maintains surface information about all servers. All jobs are submitted to the scheduler, the scheduler makes scheduling decisions. The centralized scheme is easier for implementation, but it is not very robust [4]. In the distributed model, each server maintain a scheduler and a job queue, they are self-managed. So it is more robust, but is complex for implementation. In accordance with the initiator we divide the current common Distributed algorithms into three classes: sender-initiated algorithm, receiver-initiated algorithm and hybrid algorithm [5]. The sender-initiated algorithm is easier and used generally in distributed systems. In the distributed OLAP system, there may be a number of enquiries tasks at the same time, when a task is complicated; system should decompose it into several subtasks. So, in the paper, we propose a hybrid distributed scheduling scheme with centralized scheme and stable sender-initiated algorithm based on common models. In this scheme, when the local server gets a query task, firstly the system uses centralized algorithm to allocate the task, and then the other sites use the stable sender-initiated algorithm to adjust load, so the high load sites can assign their task to others.

2 Distributed OLAP System Structure

Objects of the distributed OLAP system include enhancing the system's query and analysis speed, improving the system's reliability, enhancing the system's scalability and reducing the cost of deploying OLAP system.

The distributed system is composed of multi-OLAP servers, multi-data input module, and data querying module and application severs. These OLAP servers run on microcomputers that use TCP/IP protocol to communicate with local area network.

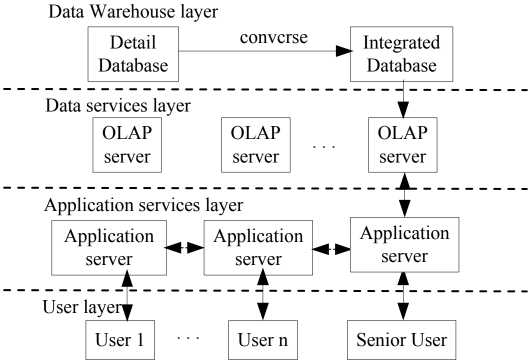


Figure 1 the distributed OLAP system structure

In order to maintain consistency for all servers, same server management procedures are installed on every server that uses the same storage cube structure and partition storage.

According to logic structure the system is divided into four layers: Data warehouse layer, Data services layer, Application services layer and User layer. Firstly, we create data warehouse for the system. In the detail database there are history data and various departments data. Then data processing is performed by data conversion tools, according to the subject collate, extract, and transform data and load it into the integrated database. And then OLAP servers establish data cube with the type of data warehouse. On application services layer, servers compose a dynamic distributed system. When the system receives the first external customer request, the job scheduler on the server assigns the task to a low-load server.

3 The Scheduling Scheme of Distributed OLAP System

3.1 Scheduling scheme design

In a distributed system, how to design the scheduling scheme depends on the actual state of the system. The purpose of the scheme is how to meet the need of the system while reduce the complexity of the system. in the Distributed OLAP system, the design of scheduling scheme is mainly considered with the following aspects[6]:

In a distributed system, how to design the scheduling scheme depends on the actual state of the system. The purpose of the scheme is how to meet the need of the system and reduce the complexity of the system. In the Distributed OLAP system, the design of scheduling scheme is mainly considered with the following aspects:

1) Certainty and enlightening algorithm

Certainty algorithm is used when all acts of the process can be forecasted. But we cannot forecast the system load, the load conditions change all the time [7]. So enlightening Algorithm is adopted for the distributed OLAP system.

2) Centralized and distributed algorithm

Centralized algorithm is easy for implementation, but it is not very robust. With the distributed algorithm, each site is self-managed. It can dynamically adjust the load of system. In the distributed OLAP system, when a query task appears, firstly system uses centralized algorithm on local server, while each server can adjust the load, when a server is overloaded it can assign the new task to others.

3) Sender-initiated and receiver-initiated algorithm

When a site wants to assign a task, it must certain that site it send the task to. It needs the load information of other sites to decide which server should be sent the task to. There are two paths to deal with the question: one is initializing by sender, the other is initializing by receiver. Sender is overload, it will send the task to others, receiver is the low load site, it can receive tasks from others. In the sender-initiated algorithm, to obtain load balance, the sender always tries to send the task when it is overload [8]. The sender-initiated algorithm is easy to be understood and to be designed, so it is adopted for the distributed OLAP system.

3.2 Job scheduling principle

In the distributed OLAP system, at a moment, there is only one coordinator. When the server receives a new query task, it assigns it as centralized algorithm, firstly, it will check its own load condition, if the load of the server is little then the scheduler will assign the task to itself, else it will assign the task to the lowest load server according. Because the system is changed dynamically with performance of tasks, the load condition of servers is also changed, while the server updates its global load table per a certain period time, so the load of server that receives the task from other servers may not be lowest. In the worst condition its load may be the highest. Thus, when a server whose load is higher receives a new task, it should assign the task to a low load server. In the system, the scheduler use stable sender-initiated algorithm to assign tasks.

On each server there are four queues: Sender queue, Ok queue, Receiver queue and Mission queue. The first three queues are used to identify the status of each server; the Mission queue is used to store tasks. The system defined unified threshold load LT and UT, LT means lower limit and UT means upper limit. The server state depends on its load conditions, In the system, we can get the load condition from the four queue length.

Receiver queue: load of the server < LT

Ok queue: LT<=.load of the server <=UT

Sender queue: load of the server > UT

Servers in the Receiver queue have low load; they can continue to receive task. Servers in the Ok queue have moderate load; they can just deal with the received tasks. Servers in the Sender queue have high load; they cannot receive new tasks. If these servers receive new task, they should assign the task to other severs. When a server receives a new task, it should firstly check the site state after receiving the task. If the state is “Receiver” or “Ok”, the server should receive the task and insert it into its Mission queue. If the state is “Sender”, the server should assign the task to other servers. For assigning the task the server need to certain a server that can receive new task. Firstly, scheduler randomly finds a server from Receiver queue and check if it can receive the new task, if the server is not busy and it can receive task then assigns the task to it, else it will find next server from the Receiver queue to check, just do the same checking until finding a moderate server. If there are no moderate server in the Receiver queue, scheduler then insert the task into its own Mission queue.

There is a Mission queue on each server of distributed OLAP system. On the server all received query tasks are treated equally, a new task will be inserted into the end of Mission queue.

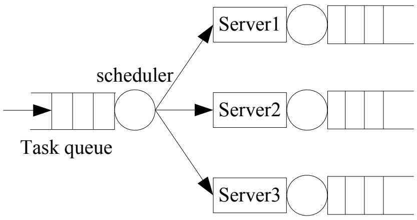


Figure 2 the schedule scheme of the system

3.3 Job scheduling implementation

According to the real system design, the distributed job scheduler is divided into three modules: initialization, server status updating and scheduling module [9].

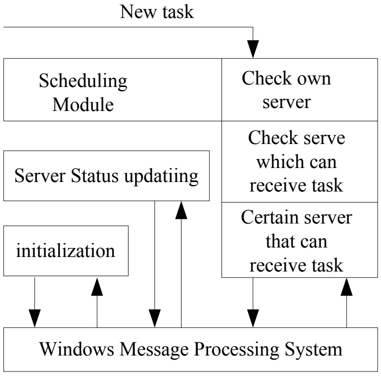

New task
Scheduling Module
Check own server
Check serve which can receive task
Server Status updatiing
Certain server that can receive task
initialization
Windows Message Processing System

Figure 3 System models structure and relationship

When the distributed system restarts, all servers begin to initialize the scheduler. First it numbers the servers from node_1 to node_n, and initializes some global variables, and then it waits for a new task. At a moment there is only one scheduler enabled. When the server receives a new task, it inserts the task into its Mission queue, and schedules the other tasks. The scheduler updates the servers' information through the status updating module. The communication among servers completes through the Windows Message Processing System. One scheduling process is shown as Figure 4.

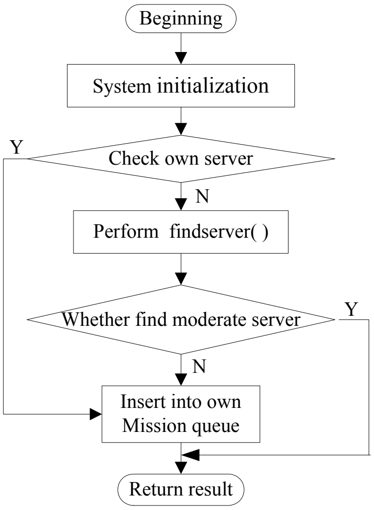


Figure 4 Flowchart of a scheduling course

When the scheduler receives a new task, it checks itself firstly, if it can receive the task, then the task will be performed on the local server. If the server will be a member of the Sender queue after receiving the task, it should send message to the servers of the Receiver queue, then assign the task according to the responses of these servers. If all the servers in Receiver queue cannot receive the new task, then the task will be inserted into the local Mission queue. Figure 4 shows how to assign the task if the local server is overload.

When a server receives a new task, if its load is low the server will insert the task into Mission queue. Then the load of the server has changed, and the status may also change, so it should update its own status. If its status changed from low to high load or from high to low, it should send messages to other servers.

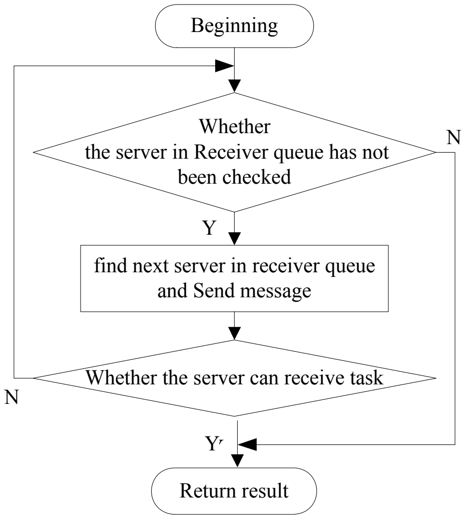

Beginning
Whether the server in Receiver queue has not been checked
N
Y
find next server in receiver queue and Send message
Whether the server can receive task
N
Y
Return result

Figure 5 Processing of finding next server

4 Experiment Result Analysis

The development environment of distributed OLAP system is Microsoft Visual C++ 6.0, the data warehouse is established with SQL Server 2000, and OLAP servers are built through SQL Server 2000 Analysis Services [10]. The operating environment of whole system is Windows 2000 Server. OLAP servers are connected with LAN and they communicate through TCP/IP protocol. Each server has the same data warehouse and OLAP analysis services and job scheduling procedure.

Figure 6 shows the query performance comparison of single, double and three OLAP server system with a single kind task. Thin broken curve shows the performance result that all leaf data are queried through a single server. Thick broken curve is the performance characteristic curve with double servers and full curve is characteristic curve with three servers. We can know from Figure 6, while the system has only one server; the query processing time increases almost linearly with the increase of the query objects' number. While the system has double or three servers, the relationship between processing time and the number of query objects is non-linear, it becomes to a smooth growth curve. With the same increased number of query objects, the time increase speed of single server system is fastest, second is the double servers system, and it is slowest for the three-server. So we can know that with the server increase the whole system's performance has been improving, but the improvement of performance is nonlinear because of communication costs.

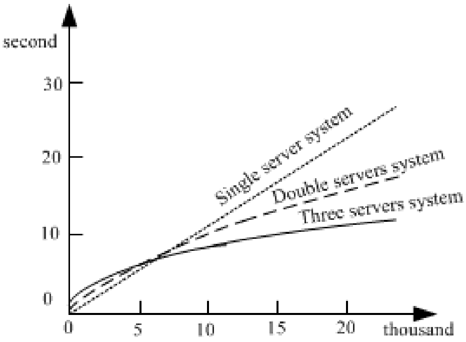


Figure 6 query performance of system

5 Conclusion

The OLAP system with one server is inefficient and expensive. The introduction distribute technology into OLAP system can assign tasks into different servers. So the OLAP application can run on servers with low configuration or microcomputers. In this paper we propose a hybrid scheduling scheme to deal with the tasks assigning. And then we propose the system structure, scheduling principle, implementing method. At last we analyze the test results of the system. Through the experiments, we know that this distributed system deal with general query tasks well.

Refrences

[1] W.H.Inmon. Building the Data Warehouse. John Wiley & Sons, Inc., New York,1993

[2] Sen Arun. Metadata management: past, present and future. Decision Support System, 37(4), 2004, pp. 151-173

[3] George Coulouris, Jean Dollimore, Tim Kindberg. Distributed system Concept and Design. Jin Beihong. China Machine Press, Beijing,2004

[4] Soleimany Cyrus, Dandamudi Sivarama P. Performance of a distributed architecture for query processing on workstation clusters. Future Generation Computer Systems, 19(5), 2003, pp.463-478

[5] Zhou Yongluan, Ooi Beng Chin, Tan Kian-Lee. An adaptabale distributed query processing architecture. Data and Knowledge Engineering, 53(6), 2004, pp. 283-309

[6] Jianjun Han and Qinghua Li. A novel static task scheduling algorithm in distributed computing environments. 18th International Parallel and Distributed Processing Symposium, 2004, pp.3-12

[7] COMINOS P, MUNRON. PID controllers: recent turning methods and design to specification [j].IEEE Proceeding of Controll Theory and Applications, 2002,149(1):46-53

[8] He K, Zhao Y. Modeling and analyzing resource schedules in grid environments[J]. Journal of Huazhong University of Science and Technology (Nature Science Edition), 2006,34(3):35-38

[9] Butala P, Sluga A. Dynamic Structuring of Distributed Manufacturing Systems[J]. Advanced Engineering Informatics, 2002, 16(2):127-133

[10] Reed Jacobson. SQL Server 2000 Annalysis Services[M]. Microsoft Press, 2000

The Design and Implementation of Distributed File Access Middleware*

Buzhong Zhang[1] Haidong Jin[2]

1 Department of Computer and information, Anqing Teachers College, Anqing, Anhui 246011, P. R. China
Email: zhbzhong@aqtc.edu.cn

2 School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu215006, P. R. China
Email: haindong@suda.edu.cn

Abstract

Traditionally, remote access of distributed file system is implemented by remote procedure call, RPC. But service provided by RPC is based on procedure. It doesn't support message transaction processing, and the implementation of RPC in Operating System kernel is difficult as well. In order to resolve these problems, a distributed file server accessing middleware (DFAM) was designed and implemented. DFAM provides not only original file access methods, but also some new interface, such as content-based access. An application of DFAM on distributed file system supporting content management is also introduced in the paper.

Keywords: Middleware, RPC, File Server, DFAM, content management

1 Introduction

In the initial stage, computer was mostly a simple, single system and online terminal was used to access it. Along with the birth of Internet, resource sharing between different computers becomes possible. Great convenience is given to common users to access remote computer files after the emergence of DFS (distributed file system). Storage inadequacy of native file system is resolved by DFS. File resources, which are scattered throughout the whole network, are reunified in a same operating semantics to users. And convenience of mutually operation on different systems is also improved.

Because of the complexity and efficiency of operating system, operating systems are mostly developed by procedure language. Remote file system access is mainly via the Remote Procedure Call[1], and it's commonly implemented in operating system kernel.

Technology of RPC is primitive and with the development of related technology, it's gradually being substituted by the technology of object-oriented middleware (CORBA, DCOM, and so on). Service of RPC provided is based on procedure. The client connects with the server directly, and there is no intermediary to deal with the request. Fault-tolerance of RPC technology is inferior, and it does not support message and transaction processing. Usually, in order to locate the server, it needs to deal with some network details. When a client delivers a request, the server must be active. Besides, it is hard to program in OS kernel and extend its interface.

In order to resolve problems above, a schema is discussed in this paper. By using distributed object middleware technology, a Distributed File Access Middleware (DFAM) is designed and implemented, in which RPC interface is substituted. Not only original interface, but also some new interface is implemented, such as content-based access. And this paper also describes the implementation processing and characteristic of DFAM. At the same time, an application of distributed file system supporting

* Supported by college young teacher funds, Anhui Province, China under Grant 2006jql210.

content-based management is introduced as well.

2 Distributed File Systems and Middleware Technology

2.1 RPC and distributed file system

Mechanism of RPC was often acting as a synchronous request-response protocol at the early period when distributed applications were being developed. With the help of this agreement, the programmer may program Client/Server applications. When client is necessary, remote server method can be called. Result will be returned to client when server's processing finished. XDR (eXternal Data Representation, RFC1014 [1]) is used as data format when data transferors between client and server.

The RPC program is distinguished by a 32-bit number, the size of TCP or UDP port number is 16 bits. For the client program can evoke the RPC program, a mechanism, that RPC port mapping should be used to map RPC identifier to Internet protocol port. When server program runs, a local protocol port is allocated. Then RPC program identifier, protocol port number and version number are registered to RPC port mapper. When a client program calls RPC, request is sent to port mapper and a corresponding port will be assigned. Then the application can be achieved.

Typically, distributed file systems such as NFS, AFS, SMB, etc. are generally C/S structure. Exported files provided by one or more file servers are offered for client using. Supported by operating system, users can mount/map exported files to local machine. Then we can use these files similar to local files. Remote communication is mostly implemented by RPC in distributed file systems. For example, parts of NFS v3 access interface are as follows:

```
    {Program NFS_PROGRAM
    {Version NFS_V3
    Void NFSPROC3_NULL (void) = 0;
    GETATTR3res NFSPROC3_GETATTR
(GETATTR3args) =1;
    SETATTR3res NFSPROC3_SETATTR
(SETATTR3args) = 2;
    LOOKUP3res NFSPROC3_LOOKUP
(LOOKUP3args) = 3;
    ACCESS3res NFSPROC3_ACCESS
(ACCESS3args) = 4;
    READLINK3res NFSPROC3_READLINK
(READLINK3args) = 5;
    READ3res NFSPROC3_READ (READ3args) = 6;
    WRITE3res NFSPROC3_WRITE (WRITE3args) = 7;
    CREATE3res NFSPROC3_CREATE
 (CREATE3args)= 8;
    MKDIR3res NFSPROC3_MKDIR (MKDIR3args) = 9;
    SYMLINK3res NFSPROC3_SYMLINK
(SYMLINK3args) = 10;
    ……
        }= 3;
    } = 100003;
```

2.2 Middleware

Middleware is the independent system software or service program in which distributed applications can use it to share resources between different platforms. Middleware services are sets of distributed software that exist between the application and the operating system in the network and manage computing resources, network communication. Goal of Middleware is to resolve the problem of interconnection and interoperability between different systems. And its core service is to solve the problem of name services, security controlling, concurrency controlling, reliability and efficiency. In order to implement interconnection, middleware should support kinds of network communication protocols, varieties of communication patterns, transmission of many data content, data formula translation, data flow controlling, data encryption, data compression and so on.

According to IDC's classification, middleware can take on the following different forms: Terminal Emulation/Screen Transaction Middleware, Data Access Middleware, Remote Procedure Call Middleware, Message-Oriented Middleware, Transaction Middleware and Distributed Object Middleware.

Because of advantages of object-oriented technology and distributed object middleware standardization, functions of object middleware become more powerful and it gradually becomes the main product of middleware. Currently, the main classification of distributed object middleware technology is: CORBA, DCOM and EJB technology.

CORBA (Common Object Request Broker Architecture) [2] was developed by Object Management Group (OMG). And it's a standard architecture for distributed object computing developed by hundreds of companies, which is based on interoperability. It allows users to develop applications by using different languages, operating systems, and hardware. It provides the portability and interoperability as well. CORBA components include ORB Core, ORB Interface, IDL Stub, DII (Dynamic Invocation Interface), Object Adaptor, IDL Skeletons, and DSI (Dynamic Skeleton Interface). CORBA interfaces and data types are defined using Interface Definition Language (IDL). In order to work on more heterogeneous environment, CORBA distributed object technology is used in our schema.

3 DFAM Design and Implementation

3.1 Environment configuration

IDL provides mapping of data types between CORBA and programming languages. It's an object-oriented interface definition language which is similar to C++, but not a programming language. When interfaces are defined in IDL, CORBA specifications allow for client and objects to be written in different languages while keeping the details transparent from each other.

Since CORBA is only a specification and not an implementation, a specific product is required. Our CORBA project uses mico2.3.11, open-source software [3]. For better combination with operating system, free software, Linux platform is chosen.

3.2 Distributed object interface design

File in DFAM, is defined as a document Object: document attributes, data and collection of operations, and version identifier is added too. On the one hand, UNIX semantics is preserved in DFAM, so files, directory and links can also be accessed in DFAM. On the other hand, interfaces supporting content-based research are provided too. It facilitates the extension of new applications.

DFAM interfaces are defined as follows:

```
Interface Document{
Attribute long version;
Attribute fileAttribute fAttribute;
Typedef sequence <octet> rwBuffer;
Long create (in string path, in long mode);
Long open (in string path, in long mode);
Long read (in string filePath, in long offset, in long nCount, inout rwBuffer buf);
Long write (in string filePath, in long offset, in long nCount, in rwBuffer buf);
Long remove (in string path); // delete a file
Long fileState (in string filePath, out fileAttribute fattr);
Long fileChmod (in string filePath, in long mode);
Long fileSetTime (in string filePath, in boolean atime, in boolean mtime, in unsigned long long time);
Long rename (in string oldName, in string newName);
……
};
};
```

If the two types of interface above can not satisfy some users' request, interfaces can be further expanded. To facilitate the access, DFAM provides some content-based access interface:

```
Typedef sequence <string> strArrays;
Long searchByContents (in strArrays keywords, in string startFolder, in long maxDepth, out strArrays filesPath);
Long searchByAttribute (in string className, in string startFolder, in boolean isIncludeSubFolder, in string attributeName, in string operator, in string attributeValue, in string dataType, out strArrays filesPath);
```

3.3 Interface implementation

Server-side program of DFAM is implemented by C++, which can be more easily compiled the program with expanded file system codes. Under CORBA, a servant is responsible for handling service requests, so servant object which is implemented the compiled skeleton interfaces, need to be implemented. CORBA relies on ORB for communicating between servers and clients. ORB acts as an object that connects remote objects transparently. ORB is responsible for finding the object implementation (servant) for the request and delivering request to the CORBA server. When server's program runs, monitor process accomplishes a series of initialization work, then waits for requests from the client. Parts of the programs are as follows:

```
CORBA:: ORB_var orb = CORBA:: ORB_init (argc, argv);
CORBA:: Object_var poaobj = orb-> resolve_initial_references ( "RootPOA");
PortableServer:: POA_var poa = PortableServer:: POA:: _narrow (poaobj);
PortableServer:: POAManager_var mgr = poa-> the_POAManager ();
Document_impl * doc = new Document_impl;
PortableServer:: ObjectId_var oid = poa-> activate_object (doc);
CORBA:: Object_var ref = poa-> id_to_reference (oid.in ());
CORBA:: Object_var nsobj = orb-> resolve_initial _references ( "NameService");
CosNaming:: NamingContext_var nc = CosNaming:: NamingContext:: _narrow (nsobj);
CosNaming:: Name name;
Name.length (1);
Name [0]. Id = CORBA:: string_dup (the "Document");
Name [0]. Kind = CORBA:: string_dup ("");
Nc-> rebind (name, ref);
Mgr-> activate ();
Orb-> run ();
……
```

Compared with server-side, client just needs call remote object, so the developing process is simple. After initializing client's ORB, Naming Service is quoted. By Naming Service, object reference is adopted. Then the client can call the interface provided by the distributed object.

When design and implementation of DFAM is finished, the third-party users may use DFAM middleware. And two methods can be selected for users: DFAM was compiled with user program together and its interface is used when needs it. The other method is: DFAM is running at the background as a daemon process. Message is sent to this process when necessary and data is returned from the sharing buffer or from returned message.

3.4 Background process compatibility with the existing file system

To preserve compatibility with current file system and originally interface can be used as well, a VFS file system in kernel is implemented that stackable layers structure [5] was used in its design process. DFAM is used to access remote server. Through a character device (or /proc file system), communication between user and kernel is achieved. Its structure is illustrated by Figure 1.

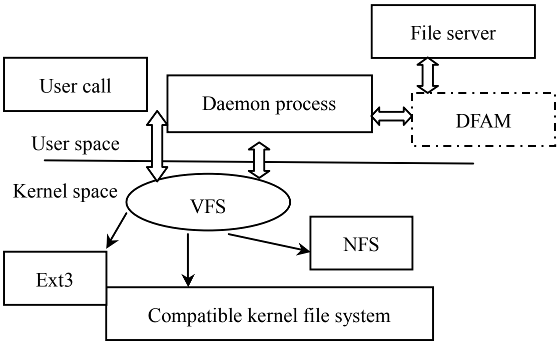


Figure 1 compatible solution with traditional file system

File system can be installed each other in Stackable layers architecture. A vnode in stacks denotes a file system. When a file operation is betaken from a user, OS kernel delivers the operation to the top of vnodes. This vnode maybe execute the following operations: the operation is executed entirely and result is traced back to the sponsor; the operation is just simple handled, and then delivers to next vnode in stack. In this way, the operation may traverse all layers. Of cause, the result can traverse all layers from bottom to top as well. Thus every vnode has the opportunity to perform additional execution. This method can be used to develop file system gradually. VFS in Linux operating system is UCLA architecture and file system in VFS layers is a vnode, so it gives us more convenience to develop file system.

4 DFAM Application in Content Based Management DFS

“File path +name” is generally used to access file in traditional file system. Tree structure is the mainly architecture that file system used to manage and organize files. But with the accelerated increase of information, efficiency of this data management became worse and worse. Many reforming schemes is proposed and researched. Among them, a scheme of content-based management is successful applied to manage business information in many companies. CFS (Content-based File System) [6][7] has the merit of file system and content-based management. Besides original interface, some new interface supporting content-based management is offered too, such as document version management, content-based access, content-based document management[8], and other powerful interface. But these functions are incompatible with traditional file system interface. So the client file system in Linux kernel is implemented using stackable layers and DFAM is used to access remote server. In order to be compatible with Linux interface, we expands Linux interface semantic of sys_fcntl( ). When the user calls, kernel analyses it, then deliver the request to the file system in VFS layer. Part program is as follows:

```
static long do_fcntl(unsigned int fd, unsigned int cmd,
                unsigned long arg, struct file * filp)
{
……
    if(cmd>14&& cmd<1024)/*We add */
        return ex_do_fcntl(filp,cmd,arg);
……}
```

Then call is analyzed:

```
static long ex_do_fcntl(struct file * filp, unsigned int cmd,
                unsigned long arg)
{
    int tmp;
    long err = -EINVAL;
    char *ch_data;
    struct un_chars *set_attr,*get_attr;
    struct i_un_ch *del_data;

    switch (cmd) {
            case 15:/*read version count*/
        err=k_fs_call(filp,9,&tmp);
        copy_to_user((int *)arg,&tmp,sizeof(int));
                break;
……}
```

In our schema, DFAM is implemented in user space, which runs as a daemon process. In order to transfer kernel request to it, a virtual device driver is programmed and it is used to communicate between user space and kernel:

```
static int cmd_dev_open(struct inode *inode,struct file *file ){
    if(cmd_device_open){
        printk("device in using now\n");
        return -EBUSY;
    }
    cmd_device_open++;
    cmd_bytes_read      =   0;
        //cmd_bytes_stored   =   0;
        cmd_bytes_written =   0;
    cmd_buffer_read=cmd_dev_buf;
    cmd_buffer_write=cmd_dev_buf;
    printk("communication device opened!\n");
        MOD_INC_USE_COUNT;
    return 0;
}
……
struct file_operations cmd_dev_fops = {
    owner:      THIS_MODULE,
        read:cmd_dev_read,
        write:cmd_dev_write,
        open:cmd_dev_open,
        release:cmd_dev_release ,
    ……
    };
```

For example, when file version count is required, system call delivered to do_fcntl->ex_do_fcntl->kfs _call, then sys_fs_rdvscount( ). At last following program executed:

```
    int sys_fs_rdvscount (struct cfs_operations *cfs_op, struct dbfs_server *server,char *path,int *count)
    {
    ……
    mm_segment_t old_fs;
    strcpy(user,server->user_name);
    strcpy(svr_address,server->svr_address);
    ret=cfs_op->read_versions_count(user,svr_address,path);
    dfam_t=find_task_by_pid(DFAM_pid);
            //search DFAM process
    wake_up_process(dfam_t);// wake up the process
    interruptible_sleep_on( &wq ); //wait DFAM result
    schedule();
    error = PTR_ERR(file); //wake up, data is returned.
    if(IS_ERR(file))     goto out_error;
    old_fs = get_fs();
    set_fs(get_ds());
    if(file->f_op)
    ret=file->f_op->read(file,buf,64,&file->f_pos);
    set_fs(old_fs);
    }
    return chtoi(buf,count);
    ……
    }
```

Then DFAM process is awoken. After accomplishing distributed task, data will be written to kernel by virtual char device. In write( ) function of the device driver, DFAM process will be asleep because of efficiency:

```
    dfam_t=find_task_by_pid(DFAM_pid);
    dfam_t->state=TASK_INTERRUPTIBLE;
```

When important problem is resolved, the remainder is programming. In order to increase the speed of implementation, FUSE file system [9] is used and server is based on Oracle iFS [6].

5 Conclusions

DFAM middleware has the following characteristics:

1. Platform independence.

For using CORBA technology, DFAM is irrelevant with system platform and program language.

2. Needless operating system support and scalability.

Traditional distributed file system needs the support of operating system and accepts its management. In this new architecture, service provided by distributed file system is released from operating system kernel and it's no longer constrained by it. Because of independence from operating system, more interfaces can be provided if the user needs.

3. Content-based access interface provided.

Characteristic of content-based access is that file location is no longer necessary for file access. Files can be searched by file attribute or other information. For demonstration, our CFS provides some interfaces, such as content-based retrieval, attribute-based retrieval, version controlling, etc. Users can expand it if necessary.

Technology of RPC is commonly used in traditional DFS. A middleware, DFAM is introduced in this paper. Remote file access is implemented by distributed object technology. It has the merits of platform independence, content-based management and expansibility. Because of technology and implementation in user space, it efficiency may be inferior to RPC while its implementation is simpler.

References

[1] RFC1057, RFC1014, RFC1094, RFC1813. http://www.faqs.org/rfcs

[2] Object Management Group, the Common Object Request Broker: Architecture and Specification-Version 2.2,

Framingham, Massachusetts: QED Pub Co, July 1998

[3] Arno Puder and Kay Römer, MICO: An Open Source CORBA Implementation, San Francisco: Morgan Kaufmann Inc. Pub., Mar 2000

[4] DAVID GIFFORD, et al, "Semantic File System," ACM Operating Systems Review, Oct. 1991, pp 16-25

[5] J. S. HEIDEMANN, G. J. POPEK, "File-System Development with Stackable Layers," ACM Transactions on Computer Systems, 12(1), Feb. 1994, pp58-89

[6] Oracle. Oracle Content Management Software Development Kit Developer Reference，http://www.oracle.com

[7] Mallik Mahalingam, Chunqiang T., Zhichen X. "Towards a Semantic, Deep Archival File System," Proceedings of the Ninth IEEE Workshop on FTDCS.May, 2003

[8] Gopal, B. and U. Manber, "Integrating Content-based Access Mechanisms with Hierarchical File Systems," Proceedings of 3rd SOSDI, Usenix, Feb. 1999, pp. 265-278

[9] FUSE (File system in User Space). http://fuse.sourceforge.net/

[10] Alessandro Rubini and Jonathan Corbet, Linux Device Drivers(2nd Edition), Beiing:O'Reilly , June 2001

[11] Ben Martin AID. Formal Concept Analysis and Semantic File Systems, 2nd International Conference on Formal Concept Analysis.Feb,2004

Learning Decision Trees from Distributed Datasets

Xie Hongxia[1,2]　Shi Liping[2]　Meng Fanrong[1]　Wang Chun[3]

1 School of Computer Science and Technology, China University of Mining and Technology, Xuzhou
Jiangsu, 221000，China

2 School of Information and Electrical Engineering, China University of Mining and Technology
Xuzhou, Jiangsu, 221000, China

3 SINOPEC Pipeline Storage & Transportation Corporation, Xuzhou, Jiangsu, 221000, China
Email: cumtcs2002@gmail.com, slpbbbb@263.net Tel: 15952260558

Abstract

Decision trees are an important data mining tool with many applications. Like many classification techniques, decision trees process the entire database in order to produce a generalization of the data that can be used subsequently for classification. Distributed databases are not amenable to such a global approach to generalization. This paper describes architecture of decision trees induction from distributed datasets which includes configuration manager retrieval data from distributed data, pruning data, and partial decision trees and data integration. In retrieval data, we explore a general strategy for explores a general strategy transforming traditional machine learning algorithms into algorithms for learning from distributed data; then we devise a pruning algorithms to optimal the data retrieval; finally we integrate the distributed sub-result data into final decision trees.

Keywords: Decision Trees; Data Retrieval; Pruning; Data Integration

1 Introduction

Data mining refers to a particular step in the KDD process. According to the most recent and broad definition [1,2], "data mining consists of particular algorithms (methods) that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (models) over the data".

With the growth in the use of networks has come the need for pattern discovery in distributed databases. In these cases, data may be inherently distributed and cannot be localized on any one machine (even by a trusted third party) for a variety of practical reasons including security and fault tolerant distribution of data and services, competitive (business) reasons, statutory constraints imposed by law as well as physically dispersed databases or mobile platforms like an armada of ships [1, 6]. In such situations, it may not be feasible to inspect all of the data at one processing site to compute one primary "global" concept or model. In such applications, the data sources of interest are typically physically distributed and are often autonomous. Given the large size of these data sets, gathering all of the data in a centralized location is generally neither desirable nor feasible because of bandwidth and storage requirements. In such domains, there is a need for knowledge acquisition systems that can perform the necessary analysis of data at the locations where the data and the computational resources are available and transmit the results of analysis (knowledge acquired from the data) to the locations where they are needed [3, 4]. In other domains, the ability of autonomous organizations to share raw data may be limited due to a variety of reasons (e.g., privacy considerations). In such cases, there is a need for knowledge acquisition algorithms that can learn from statistical summaries of data (e.g., counts of instances that match certain criteria) that are made available as needed from the distributed data sources in the absence of access to raw data.

This paper addresses the problem of data mining in distributed databases, in particular learning models for classification or prediction. Certain machine learning methods, referred by the terms integrating multiple models and ensemble methods can potentially be used with distributed databases. For the purpose of this discussion, this paper describes architecture of Decision Trees Induction from Distributed Datasets which includes retrieval data from distributed data, pruning data, and partial decision trees and data integration. In retrieval data, we explore a general strategy for explores a general strategy transforming traditional machine learning algorithms into algorithms for learning from distributed data; then we devise a pruning algorithms to optimal the data retrieval; finally we integrate the distributed sub-result data into final decision trees.

2 The Architecture of Decision Trees Induction from Distributed Datasets

The scalability of a data mining system refers to the ability of the system to operate as the number of data sites increases without a substantial or discernable reduction in performance. Efficiency, on the other hand, refers to the effective use of the available system resources. The former depends on the protocols that transfer and manage the intelligent agents to support the collaboration of the data sites while the latter depends upon the appropriate evaluation and filtering of the available agents to minimize redundancy [7, 8]. Combining scalability and efficiency without sacrificing predictive performance is, however, an intricate problem. To understand the issues and better tackle the complexity of the problem, we explore architecture of decision trees induction from distributed datasets (see Figure 1).

Let's focus on the components of the system and the overall architecture. Assuming that the data mining system comprises of several data sites, each with its own resources, databases, machine learning agents, and learning capabilities, we used a protocol[4,5,6] that allows the data sites to collaborate efficiently without hindering their progress. First, local data retrieval and pruning algorithms execute on the local database to compute the data site's local classifiers. Then, each data site may import (remote) classifiers from its peer data sites and combine these with its own local classifier using data integration component. Again, the local data retrieval and pruning algorithms can be either local or imported from other data sites. Finally, once the base and decision trees are computed, the system manages the execution of these modules to classify data sets of interest. These actions may take place at all data sites simultaneously and independently. The system can thus be viewed as a coarse-grain parallel application where the constituent sites function autonomously and (occasionally) exchange classifiers with each other.

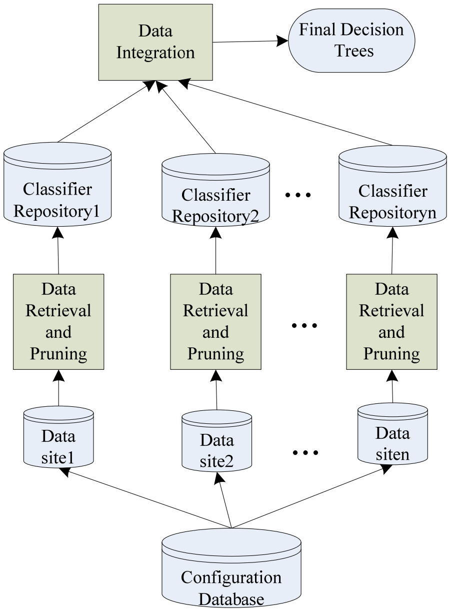


Figure 1 the Architecture of Decision Trees Induction from Distributed Datasets

The configuration of the distributed system is maintained by the Configuration Manager (CM), an independent server that is responsible for keeping the state of the system up-to-date. In this example, all sites exchange their base classifiers to share their local view of the learning task. The user of the data site controls the learning task by setting the parameters of the user configuration file, e.g. the algorithms to be used, the database to learn, the images to be used by the animation facility, the folding parameters, etc.

3 Data Retrieval from Distributed Datasets

We say that a distributed learning algorithm Ld (e.g., for decision tree induction from distributed data sets) is exact with respect to the hypothesis inferred by a batch learning algorithm L (e.g., for decision tree induction from a centralized data set) if the hypothesis produced by Ld using distributed data sets D1, … ,Dn stored at sites 1,…, n (respectively), is the same as that obtained by L from the complete data set D obtained by appropriately combining the data sets D1,…, Dn. Similarly, we can define exact distributed learning with respect to other criteria of interest (e.g., expected accuracy of the learned hypothesis).

In this setting, the problem of learning from distributed data sets can be summarized as follows: The data is distributed across multiple autonomous sites and the learner's task is to acquire useful knowledge from this data. For instance, such knowledge might take the form of a decision tree or a set of rules for pattern classification. In such a setting learning can be accomplished by an agent that visits the different sites to gather the information needed to generate a suitable model (e.g., a decision tree) from the data (serial distributed learning). Alternatively, the different sites can transmit the information necessary for constructing the decision tree to the learning agent situated at a central location (parallel distributed learning).

Our general strategy for transforming a batch learning algorithm (e.g., a traditional decision tree induction algorithm) into an exact distributed learning algorithm involves identifying the information requirements of the algorithm and designing efficient means for providing the needed information to the learning agent while avoiding the need to transmit large amounts of data. This yields a natural decomposition of a learning algorithm into two components [9,10]: (1) an information extraction component that formulates and sends a statistical query to a data source and (2) a hypothesis generation component that uses the resulting statistic to modify a partially constructed hypothesis (and further invokes the information extraction component if needed).

Suppose we define a distributed information retrieval operator Id that generates from each data set Di, the corresponding information Id(Di), and an operator C that combines this information to produce I(D). That is, the information retrieved from the distributed data sets is the same as that used by L to infer a hypothesis from the complete dataset D. That is, C[Id(D1), Id(D2),…, Id(Dn)] =I(D). Thus, we can guarantee that Ld(D1,…, Dn) =H(C[Id(D1),…,Id(Dn)]) will be exact with respect to L(D) = H(I(D)).

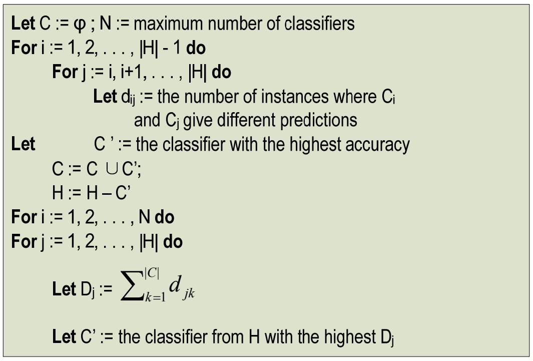
Now we introduce a new algorithm for building decision tree, called DRDD (see Figure 2). This algorithm is a modification of practical impurity-based decision tree building algorithms to use exact values of purity gain.

```
DRDD (s, X, R,φ){
    //  The parameter s identifies current tree's node,
    //   X is the set of all attributes available
    //   for testing and R is the set of function's
    //  restrictions corresponding to s and either
    //  $p(s1(i)) \ge p(s) \ge p(s0(i))$ or $p(s0(i)) \ge p(s) \ge p(s1(i))$[3].
    if  all instances arriving at s
          have the same classification then
        Set s as a leaf.
        Set classification of s as a classification of
          any example arriving to it.
    else
        Let $x_i = \arg\max_{x_i \in X}\{PG(f_R, x_i, \varphi)$[3]$\}$ be the
          variable with the greatest purity gain at $f_R(x)$.
        Choose a variable $x_i$ as a splitting variable.
        Let $s_0$ and $s_1$ be the left and right sons of s.
        Run DRDD ($s_1$, X-{$x_i$}, R∪{$x_i$ = 1},φ).
        Run DRDD ($s_0$, X-{$x_i$}, R∪{$x_i$ = 0},φ).
    end if
    }
```

Figure 2 DRDD algorithm

4 PRuning of Data Retrieval

Pruning refers to the evaluation and selection of classifiers before they are used for the training of the decision trees. A pruning algorithm is provided with a set of pre-computed classifiers H (obtained from one or more databases by one or more machine learning algorithms) and a validation set V (a separate subset of data, different from the training and test sets) [11]. The result is a set of classifiers C⊆H to be combined in a higher level meta-classifier. Determining the optimal meta- classifier is a combinatorial problem, so we employ the accuracy, diversity, and coverage and specialty metrics to guide the greedy search. More specifically, we implemented a diversity-based pruning algorithm and three instances of the combined coverage/specialty-based pruning algorithm, described in Figure 3.The algorithm works iteratively selecting one classifier each time starting with the most accurate base classifier. Initially it computes the diversity matrix d where each cell dij contains the number of instances of the validation set for which classifiers Ci and Cj give different predictions. In each round, the algorithm adds to the list of selected classifiers C the classifier Ck that is most diverse to the classifiers chosen so far, i.e. the Ck that maximizes D over C∪Ck, ∀k in 1, 2, …, |H|. The selection process ends when the N most diverse classifiers are selected. N is a parameter that depends on factors such as minimum system throughput, memory constraints or diversity thresholds. The algorithm is independent of the number of attributes of the data set.



```
Let C := φ ; N := maximum number of classifiers
For i := 1, 2, . . . , |H| - 1 do
        For j := i, i+1, . . . , |H| do
                Let $d_{ij}$ := the number of instances where $C_i$
                        and $C_j$ give different predictions
Let             C ' := the classifier with the highest accuracy
        C := C ∪C';
        H := H – C'
For i := 1, 2, . . . , N do
For j := 1, 2, . . . , |H| do
        Let $D_j$ := $\sum_{k=1}^{|C|} d_{jk}$
        Let C' := the classifier from H with the highest $D_j$
```

Figure 3 Pruning Algorithm

Even though the algorithm performs a greedy search, it combines classifiers that are diverse (they classify correctly different subsets of data), accurate (with the best performance on the data set used for evaluation with respect to the class specialty) and with high coverage.

5 Data Integration from DIQ

The discussion of distributed learning in the preceding section assumed that it is possible to extract the information needed by the learning algorithm from the distributed data sources. This is rather straightforward in the case of data sources that have a homogeneous structure and semantics.

Combining multiple models has been receiving increased attention in the literature [11, 12，13]. In much of the prior work on combining multiple models, it is assumed that all models originate from different subsets (not necessarily distinct) of a single data set as a means to increase accuracy, (e.g. by imposing probability distributions over the instances of the training set, or by stratified sampling, sub-sampling, etc.) and not as a means to integrate distributed information.

In all cases considered so far, all classification models are assumed to originate from databases of identical schemas. Since classifiers depend directly on the format of the underlying data, minor differences in the schemas between databases derive incompatible classifiers, i.e. a classifier cannot be applied on data of different formats. Yet these classifiers may target the same concept. We seek to bridge these disparate classifiers in some principled fashion.

In this session we formulate DIQ that bridge the schema differences to allow data mining systems to share incompatible and otherwise useless classifiers. DIQ is designed to provide a unified query interface over a set of distributed, heterogeneous and autonomous data sources which enables us to view each data source as if it were a table. The heterogeneity of the data sources implies that each data source may use different technology for supporting its own operation (relational databases, flat files, web pages, etc.), store its information using different syntactic representation, and use different ontological structure (semantic differences) [5, 9]. Thus, we need mechanisms for transforming the original data explicitly or implicitly into a common form (e.g., tables). Furthermore, the autonomy of the data source limits the range of operations that can be performed on the data source (e.g., the types of queries allowed), and the precise mode of allowed interactions can be quite diverse. Hence, strategies for obtaining the required information within the operational constraints imposed by the data source are needed. The complexity associated with accessing the data answering queries must be hidden from the users. Furthermore, a user must be able to view the data sources from multiple ontological perspectives as appropriate in different contexts. The data integration component of DIQ provides an ontology-based solution to the data integration problem [11,13].

6 Summary

Like many classification techniques, decision trees process the entire database in order to produce a generalization of the data that can be used subsequently for classification. Distributed databases are not amenable to such a global approach to generalization. This paper addresses the problem of data mining in distributed databases, in particular learning models for classification or prediction. Certain machine learning methods, referred by the terms integrating multiple models and ensemble methods can potentially be used with distributed databases. For the purpose of this discussion, this paper describes architecture of decision trees induction from distributed datasets which includes retrieval data from distributed data, pruning data, and partial decision trees and data integration. In retrieval data, we explore a general strategy for explores a general strategy transforming traditional machine learning algorithms into algorithms for learning from distributed data; then we devise a pruning algorithms to optimal the data retrieval; finally we integrate the distributed sub-result data into final decision trees.

References

[1] K. Ali and M. Pazzani. Error reduction through learning multiple descriptions. Machine Learning, ,1996

[2] L. Breiman. Stacked regressions. Machine Learning, , 1996

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees. Wadsworth,Belmont, CA, 1984

[4] C. Brodley and T. Lane. Creating and exploiting coverage and diversity. In Work. Notes AAAI-96 Workshop Integrating Multiple Learned Models, 1996

[5] M. Ben-Or and N. Linial. Collective Coin Flipping. In Randomness and Computation (S. Micali ed.), Academic Press, New York, 1990

[6] A. Bernasconi. Mathematical Techniques for the Analysis of Boolean Functions. PhD Thesis,Universita di Pisa, 1998

[7] A. Blum. Rank-r decision trees are a subclass of r-decision lists. Information Processing Letters, 1992

[8] P. Chan and S. Stolfo. Sharing learned models among remote database partitions by local meta-learning. In Proc. Second Intl. Conf. Knowledge Discovery and Data Mining, 1996

[9] W. Cohen. Fast effective rule induction. In Proc. 12th Intl. Conf. Machine Learning, 1995

[10] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone. Classi_cation and Regression Trees. Wadsworth International Group, 1984

[11] N.H. Bshouty. Exact Learning Boolean Functions via the Monotone Theory.Information and Computation, 1995

[12] N.H. Bshouty and V. Feldman. On Using Extended Statistical Queries to Avoid Membership Queries. Journal of Machine Learning Research, 2002

[13] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; and Uthurusamy, R. 1996. Advances in Knowledge Discovery and Data Mining. Menlo Park, California/Cambridge, Massachusetts/London, England: AAAI Press/MIT Press

Xie Hongxia is a instructor and a doctor in School of computer science and technology, China University of Mining and Technology. She graduated from China University of Mining and Technology in 2002. She has published several Journal papers. Her research interests are in distributed parallel processing, data mining, neural network.

Wireless Distributed Monitoring Network System Research Based on Embedded Control*

Shihong Qin[1] Zhou Ge[2] Junqiao Xiong [1] Bin Shen[1]

1 Electrical and information Engineer Department of Wuhan Institute of technology, Wuhan, 430023 P.R.China

2 Electrical and information Engineer Department of Wuhan Polytechnic University, Wuhan, 430023 P.R.China
E-mail: Qinsh@whpu.edu.cn, Shinegogo@gmail.com

Abstract

There is no mobile phone devices or wire net line equipments for real-time control in some industrial fields such as metallurgy, mine, transferring port. Therefore, the WLAN monitoring platform system is investigated for these industrial fields in the paper. Using Radio Frequency (RF) wireless distributed network and the embedded technology, the monitoring platform's module structure has been established. The hardware and software of monitoring system have been analyzed. The configure mode of the wireless monitoring devices have also been given.

Using the given control modules, RF wireless transmitting, data acquiring and processing of signals, as well as equipment states diagnosing can be implemented. The research shows that the system platform can solve the communication and monitoring problems in such no cable net line equipments and no mobile phone field-working environments and that the control system given has the characteristic of reliability and flexibility. More ever this system is economic, and is auspicious for further applications.

Keywords: embedded system, distributed monitoring control, WLAN

1 Introduction

At present, in metallurgy, mineral and transferring port industry, the real-time control system is needed, and this leads to the demand for control system devices monitoring and management. In such industrial field environment, real-time control network devices are difficult to be equipped and even so, it cost much. Wireless Local Area Network (WLAN) can control and manage the mobile devices immediately. By using RF and the WLAN the monitoring system can track out the device states; evaluate system control performance, error diagnosis.

Along with arrival of internet era, monitoring technology will be on the base of open network platform. That is a key of improving manufacturing productivity. The sensor and transducer mode system based on the wire network for some industrial field has limitations. This research gives an integrated -distributed RF monitoring system based on the WLAN, which is the core of the. It conquers the limit above, fits for the conditions of locomotive device, and makes device management and control intelligent by the RF WLAN. Because the control system is distributed, all nodes on the net have data process capability and the data transmitting is reduced enormously the main node only receives processed data from rest nodes. The RF WLAN monitoring system is useful for the innovation and effectiveness in the manufacture system.

2 The Embedded System Research about Hardware

2.1 Hardware Topological Layout

The hardware structure adopts center-remote node

* This paper format should be strictly adhered.

mode, which is made up of three parts. Those are center nodes, WLAN transmitting medium and remote nodes, shown as in Figure 1:

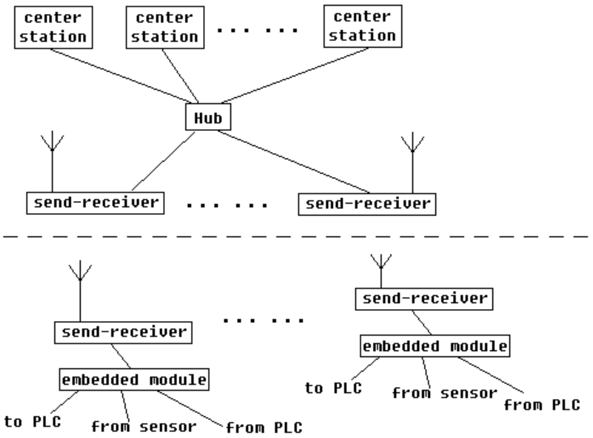


Figure 1 Hardware Topological Layout

The system is used Radio Frequency (RF) WLAN, instead of twisted-pair or optical cables, and adopts Breeze

NET DS.11 base station devices, which uses Direct Sequence Spread Spectrum (DSSS) technology. That can carry out both large-scale control and fast stabilization net connections.

The center nodes adopt AU-DS.11D, which is a base station production on IEEE 802.11 TGb. Those using DSSS technology sends and receives signals from remote nodes on the Industrial Standard Ethernet. The data transmission speed reaches up to 11 Mbps. And these devices have two styles, integrated antenna and external one, which offer choice for devices installing.

Wireless transmission medium adopts RBD-DS.11 and BUD-DS.11 relay devices. It works at 2.4GHz band, is well suited to connection layout. It is also easy to expand nodes and meet the standard of Ethernet interface. These relays agree with IEEE 802.11 TGb as well.

Remote nodes adopt RB-DS.11D devices. As nodes of send-receiver on network, these are connected to the embedded control modules given below.

The modules collect the data received from sensors installed at industrial devices or PLC (Programmable Logical Controller, an industrial standard controller), process the information, and diagnose devices' states and feedback to the center stations. The Topological structure is shown as Figure 2:

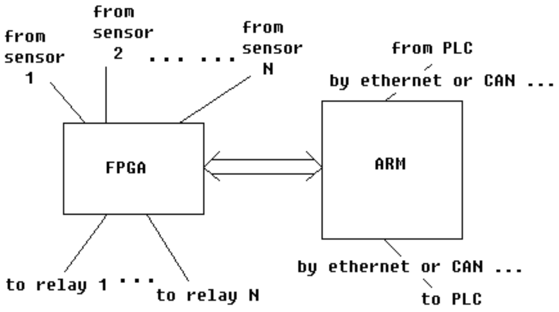


Figure 2 Embedded module system structure

As shown, FPGA (Field Programmable Gates Array, a programmable logical unit) is played at the front of ARM, which is mapped ARM's input/output port set and transmits the data to ARM by the ports. Because the data rate from sensors is low than ARM CPU computing speed, FPGA is also a data buffer between them. ARM unit is selected ARM7 designed by Texas Instrument corp. which has 200MHz CPU and abundance interfaces. And PLC sends the states data by the Ethernet port of ARM7. If necessary, the A/D ports supported by ARM7 can receive analog signals from sensors directly.

After processing the data, ARM sends the collected and diagnosed information to the center stations through Ethernet protocol by wireless network.

2.2 The Characteristic Analysis

Designed in the topology above, network nodes can be added, removed and modified easily for fulfilling to the requirements. At the same time, on this network construction, it can take low cost to carry out the redundancy systems by controlling software. When a certain node works failure, another in the topological structure can fill in.

Adopted this, the system power loss can be reduced and the signal transmission distance can be lengthened greatly.

2.3 The System Base Stations Configure Designing

Owing to the RF WLAN, the hardware install of wireless devices is simple and convenient than wire network. The one of merit is leaving out twisted-pair cable layout and the former is not needed to design and setup the cable ring. This design enhanced the characteristics of embedded technology, such as smaller and more ambulant, shown in Figure 3.

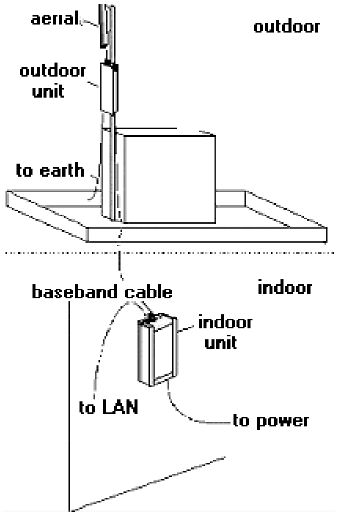


Figure 3 The install of the center station transmit medium

If adopted accustomed setup mode, the stations' wireless devices would have been built out of the center node control room. In that way, the data sent from the center control computers are transmitted to outdoor antenna by cables, and the digital signal would be damped on these twisted-pair cables. Besides, the distance from the center nodes to the antenna cannot be designed too long. The center stations' wireless devices, which are divided into indoor and outdoor parts, can solve the problems mentioned above. This design is not only to fit for installing easily, but also to reduce signal loss. By way of this, signals sent from the center stations have been conversed into the baseband cables transmitting. Thus, the center stations' wireless devices performance has been improved a lot such as stronger anti-interference, longer transmitting distance and reliability.

3 Software Systems Analysis

3.1 Operation System and System Software

The center station nodes are located at the center control room, which is far from execution fields. And PC with WINDOWS operation system is used as these control units usually. WINDOWS is used widely and has lower developing cost.

On the remote station, embedded ARM7 units adopt ucLinux operation. This operation system is open source code, and has some features, including true multitasking, real-time controllable, shared data base, proper memory management and TCP/IP networking. On the ucLinux platform, the remote station can be added into wireless network easily through its Ethernet driver and TCP/IP protocol package.

3.2 Application Software Design

The application software is made of two parts, which are the center station's software and the embedded module's. The former's function is to process the data from the remote stations then sending control signal to them, and offers human-computer interface. The latter is to analyze measuring data, to diagnose industrial device states and sending real time data to the center stations. Furthermore, the embedded application software controls the executive device, such as PLC or control relay.

4 An Application for Port Crane Devices

In the application of Nantong Tongsha Bulk Product Port, wire network system cannot manage the crane devices in real time. These devices are required feeding back status information immediately, however, whose positions are scattered. If the center stations could not receive the data from remote nodes in time, the whole control system should not work efficiently. Therefore, the embedded wireless control system had been applied.

This port is 80,000m^2 wide and has 8 cranes (more in building). The positions of devices are scattered in whole port area. The control system had been designed as showing Figure 1. In this topological layout, the distance from center station to remote can reach 1 km. The remote station topology shows as in Figure 4.

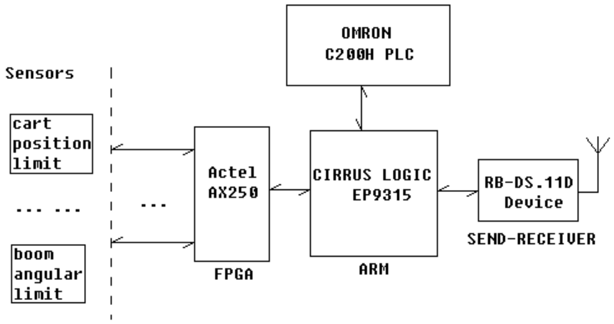


Figure 4 the remote station topology

ARM is adopted EP9315 by Cirrus Logic Co. FPAG is adopted AX250 by Actel, and PLC is adopted C200H by OMRON. The status data of devices has been acquiring by embedded process block through FPAG and PLC. The data has been processed by EP9315, and then, sent to the center station by the RB-DS.11D send-receiver.

The center station receives information from the remote, and the center nodes analyze and process the data. The data processed will be sent back to the remote embedded control block.

5 Conclusions

The wireless distributed monitoring network system based on embedded control has been investigated. The system can implement data acquiring, processing, and communication, devices monitoring for mine and transferring ports industrial fields. The research shows that the control system given has the characteristic of reliability and flexibility for controlling and monitoring, and that it carry out data communication in such no wire net line equipments and no mobile phone devices by using RF WLAN. This system is effective and economic, and has auspicious for further applications.

References

[1] Paulo P., FTT-Ethernet: A flexible real-time communication protocol that supports Dynamic QoS management on Ethernet-based systems. IEEE Transactions on industrial informatics, 2005, 1(3): 162-172

[2] Dolejs O., Smolik P., Hanzalek Z. On the Ethernet use for real-time publish-subscribe based applications. IEEE International Workshop on Factory Communication Systems -Proceedings, WFCS, 2004:39-44

[3] C. M. Krishna, Kang G. Shin. Real-time Systems. Tsinghua University Press, 2001

[4] G. Bianchi. Performance Analysis of the IEEE 802.11 Distributed Coordination Function. IEEE Journal on Selected Area Communications, 2000, 18(3):535-547

[5] Hong Y., Gregory C. W, Linda B. Wireless Local Area Networks in the Manufacturing Industry. Proceedings of the American Control Conference Chicago, Illinois,2000:2363-2367

[6] P. A. Wiberg, U. Bilstrup. Wireless technology in industry applications and user scenarios. Proc. IEEE Int. Conf. Emerging Technologies and Factory Automation (ETFA ’01), 2001:123-133

[7] Shihong Qin is full professor and the dean of electrical and information engineering department of WIT, graduated from Huazhong university of science and technology in January 1999, and awarded PH.D degree. His most interest fields are information processing, electrical intelligent instrument on-line test, etc

Isomorphic New Parallel Division Methods for Special Large Numerical Matrixes’ Transpose

Zhou Qihai Huang Tao Li Yan Wang Yonglei

Research Institute of Information Technology Application, Southwestern University Of Finance and Economics, Chengdu, Sichuan, 610074, China

School of Economic Information Engineering, Southwestern University Of Finance and Economics, Chengdu, Sichuan, 610074, China
Email: zhouqh@swufe.edu.cn

Abstract

In recent year parallel computing access to a relatively rapid development, provide a solution to solve complex problems. With the development of economy, social, science and technologies of which all things are done, seen, studied deeply by human, so the dimension of information has been rising. Thereby creating numbers of large matrixes, there is much computation in processing of the large (including giant) matrix. Parallel computing is an effective solution to solve complex problems, but it involves an important and basic question of how to divide the large matrix into some sub-matrixes or smaller parts of the large matrix. The traditional extant dividing methods (such as band division and chessboard division) have same deficiencies: communication intensity between the processors is high. In order to overcome this weakness and realize large matrix parallel transpose efficiently, based on decreasing the communication intensity among the processors, some isomorphic new parallel dividing methods for special large matrix transpose are studied and given in this paper.

Keywords: Isomorphic, Special Large Numerical Matrixes, Transpose, Parallel Division Methods, Oblique belt division of a matrix, Right angle belt division

1 Important Information

In recent year parallel computing access to a relatively rapid development, provide a solution to solve complex problems. With the faster pace of current economic and social development, with the development of today’s science and technologies of which all things are done, seen, studied deeply by human, so the dimension of information has been rising [1], a large number of large (even giant) matrixes are producing. Dealing with such information may be involved in the process of large-scale matrix computing, and matrix computing especially the large computing is very complex and important, which provides the space where parallel computing could play an important part. However, to solve large matrix problems (e.g. transpose) with parallel computing must face a question how divide a large matrix into some sub-matrixes or smaller parts of the large matrix. It is necessary to analysis the extant dividing methods and point out their deficiencies:

(1) The matrix transpose algorithm based on the strip division requires each processor to communicate with other processors. That requires high capacity of communication lines between processors. Some processors may sometimes suspend because the communication lines are blocking, so that affect the normal operation of processors, and processors can not fully play to their potential.

(2) The matrix transpose algorithm based on the chessboard division. In the early, because of the relatively small number of matrix the load of communication lines is few and low. But when the computing is going on, while the problems to be computed are more and more complicated, the number

of sub-matrix increases rapidly, so that the load of communication lines will be growing more rapidly.

In order to overcome the weakness of both, the two new parallel dividing methods are constructed in this paper which can decrease the communication intensity between processors, so as to enhance the efficiency of solving the problems. In this paper only discusses matrixes, of which the row number is equal to the column number.

2 Traditional Division Methods

The two typical Traditional division methods are given as following.

2.1 Band division method

The characteristics of band division method is that dividing the large matrix into some vertical (or lever) bands including some whole entire row or some whole entire columns, each of them assigned to one different processor [2-4]. These rows or columns can be continuous or isometric phases. The former called as continuous block strip, the latter called as cycle band division shown as Figure 1 (where P_x is described as a processor x, “i,j” is described as the row number and column number of an element a_{ij} in the given matrix; it is the same in followings).

P_0	P_1	P_2	P_3	P_4
0,1	2,3	4,5	6,7	8,9

A) Continuous vertical band division

P0	P1	P2	P3	P4
0,5	1,6	2,7	3,8	4,9

B) Cycle vertical band division

Figure 1 A case of band division of a matrix

2.2 Chessboard division method

The characteristics of chessboard division method are that the square divided into a number of sub-matrixes, each of them assigned to one different processor. Each of processors has non-complete rows or columns (shown as Figure 2).

0,0 0,1 P_0 1,0 1,1	0,2 0,3 P_1 1,2 1,3	0,4 0,5 P_2 1,4 1,5	0,6 0,7 P_3 1,6 1,7	0,8 0,9 P_4 1 8 1,9
2,0 2,1 P_5 3,0 3,1	2,2 2,3 P_6 3,2 3,3	2 4 2,5 P_7 3,4 3,5	2,6 2,7 P_8 3,6 3,7	2,8 2,9 P_9 3,8 3,9
4,0 4,1 P_{10} 5,0 5,1	4,2 4,3 P_{11} 5,2 5,3	4,4 4,5 P_{12} 5,4 5,5	4,6 4,7 P_{13} 5,6 5,7	4,8 4,9 P_{14} 5,8 5,9
6,0 6,1 P_{16} 7,0 7,1	6,2 6,3 P_{17} 7,2 7,3	6,4 6,5 P_{18} 7,4 7,5	6,6 6,7 P_{19} 7,6 7,7	6,8 6,9 P_{20} 7,8 7,9

Figure 2 A case of chessboard division of a matrix

3 New Way of Dividing Methods for Parallel Transposing Matrix

It was known that the result of exchanging elements a_{ij} and a_{ji} in a given matrix A is the transposed matrix AT of matrix A (where i=1,2,3, …, n; j=1,2,3, …, n). But it is only considered and located every single element couple (a_{ij},a_{ji}) in matrix A to realize the relationship between matrix A and its transposed matrix A^T. In fact, if people expend the field of vision to look at the relationship of elements in A and A^T, two important isomorphic natures should be found at least:

(1) the row i of A is just the column i of A^T (where i=1,2,3, …, n);

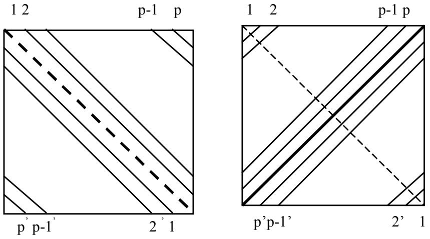
(2) the distributed forms of elements of matrix can be studied from different views.

So, using the isomorphic natures about the relationship between matrix A and its transposed matrix A^T, some isomorphic new methods of parallel computing division for special large numerical matrixes’ transposing are created in this paper as follows.

4 Oblique Belt Division Methods

All equations must be typed or written neatly in black. They should be numbered consecutively throughout the text. Equation numbers should be enclosed in parentheses and flushed right. Equations should be referred to as Eq. (X) in the text where X is the equation number. In multiple-line equations, the number should be given on the last line.

The elements of matrix A can be distributed as oblique belts which are parallel with the main (or vice) diagonals (shown as Figure 3). So people can get A^T by exchanging the elements distributed in the two sides of the main (or vice) diagonals which are symmetry about the main (or vice) diagonal. Based on this new idea, oblique belt divisions are given here.



A) Right oblique belt division

B) Left oblique belt division

Fig 3 Oblique belt division of a matrix

Oblique belt division includes right oblique belt division and left oblique belt division. A large matrix could be divided into some element belts by right oblique belt division of which all of the belts are parallel to the main (or vice) diagonal of the given matrix (shown as Figure 3, where left oblique belt division could be described similarly as right oblique belt division), in which the bandwidth (where $1 \le \text{bandwidth} \le k < n$) of both belt_i (which sets the right of the main diagonal line) and its symmetry belt'i (which sets the left of the main diagonal line) must be equal, but the bandwidth of both belt_i and other belt_j ($i \ne j$) is not necessary to equal according to specific circumstances. To keep the balance of the computed element number in every processor, the following different tactics for oblique belt division should used:

4.1 Oblique belt division with equal bandwidth

If the bandwidth of all oblique belts is same, then P_i (i.e. processor_i, where $1 \le i \le p$) could save and process one equal distance symmetry oblique belts group which is composed of four belts belt_i, belt'_i, belt_{p-i+1} and belt'_{p-i+1} (where belt_i and belt'_i are symmetric, belt_{p-i+1} and belt'_{p-i+1} are symmetric too, for which the distances from belt_i and belt'_i to the main diagonal line are equal to the distances from belt_{p-i+1} and belt'_{p-i+i} to the main diagonal line) at the same time. It is obvious that all of belt_i, belt'_i, belt_{p-i+1} and belt'_{p-i+1} are in the same processor when they are swapped in exchanging procession, and the exchanging of the four belts is directly internal exchange without communicating among the relational different processors. In other word, the cost of expensing about communication among these processors will be least.

4.2 Oblique belts division with unequal bandwidth

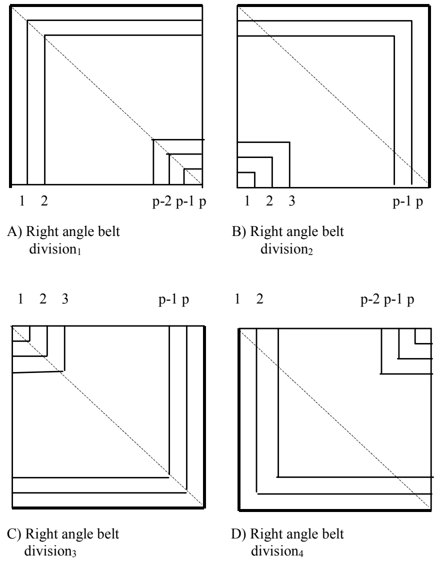
(1) If the bandwidth of the symmetry oblique belt (i.e. belt_i and belt'_i, where $1 \le i \le p$) is same while the bandwidth of some oblique belts is not same, but the element number of every symmetry oblique belts are about equal, then P_i could save and process two belts (i.e. belt_i and belt'_i) at the same time.

(2) Of course, if the bandwidth of all oblique belts are not same, but the element number of every equal distance-total (which is constant p+1) belt group consisted of belt_i and belt_{p-i+1} (or its equal distance-total symmetry belt group consisted of belt'_i and belt'_{p-i+1}) is about equal, then P_i could save and process four belts (i.e. belt_i, belt'_i, belt_{p-i+1} and belt'_{p-i+1}) without cost expense.

Therefore, the new method based on oblique belts division could raise the transposing efficiency to the four belts and whole large matrix.

5 Right Angle Belt Division Methods

Right angle belt division includes band-length- reduce right angle belt division (which means: longer the length of belt is, more the number of belt is) and band-length-raise right angle belt division (which means: shorter the length of belt is, fewer the number of belt is) of which the bandwidth satisfied $1 \le \text{bandwidth} \le k < n$. The method of reducing right angle belt division can be shown as Figure 4, where all elements standing on the two sides of the same right angle should compose of a special reducing right angle belt used to save and process the elements by the same P_i for transposing large matrix. In order to realize the balance of the computed element number in every processor, the following different tactics should be adopted for right angle belt division:



A) Right angle belt division$_1$

B) Right angle belt division$_2$

C) Right angle belt division$_3$

D) Right angle belt division$_4$

Fig 4 Right angle belt division of a matrix

5.1 Right angle belt division with equal bandwidth

If the bandwidth of all right angle belts is same, then P_i could save and process one equal distance-total right angle belts group which is composed of two belts belt$_i$ and belt$_{p-i+1}$ (for which the total of the distances from belt$_i$ and belt$_{p-i+1}$ to the high end of the main diagonal line is equal to constant p+1) at the same time. It is clear that both belt$_i$ and belt$_{p-i+1}$ are in the same processor when the elements in them are exchanging in swapping procession, and the swapping of the these elements is directly internal exchange without communicating among the relational different processors. It should tell us that the cost of taking the communication among these processors will be least.

5.2 Right angle belts division with unequal bandwidth

(1) If the bandwidth of the bandwidth of some right angle belts is not same, but the element number of every right angle belt is about equal, then P_i could save and process one belt (i.e. belt$_i$) at the same time.

(2) If the bandwidth of all right angle belts are not same, but the element number of every equal distance-total belt group consisted of belt$_i$ and belt$_{p-i+1}$ is about equal, then P_i could save and process the elements in the two belts (i.e. belt$_i$ and belt'$_{p-i+1}$) without cost expensed.

So, the new method based on could raise the transposing efficiency to the four belts and whole large matrix.

6 Lap Belt Division Method

Consider the special and interesting fact: it is the most natural that all elements which are placed on the same lap form the outermost to the most internal of a given matrix should be divided into the same group, shown as Figure 5.

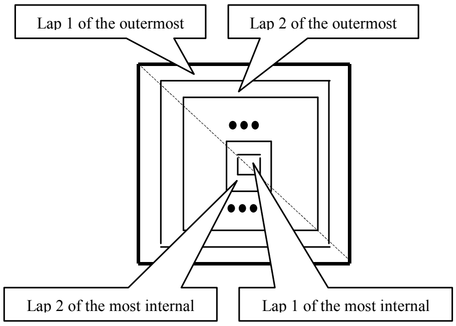


Fig 5 Lap belt division of a matrix

In general speaking, all the laps of a given matrix could be named as lap 1, lap 2, lap 3, …, lap m (where m=[(n+1)/2] ≤p) form outside to inside.

6.1 Lap belt division with equal bandwidth

If the bandwidth of all lap belts is same, then P_i could save and process one equal distance-total lap belts group which is composed of two lap belts lap belt$_i$ and lap belt$_{p-i+1}$ (for which the total of the distances from lap belt$_i$ and lap belt$_{p-i+1}$ to the most internal of the center lap belt (Notice: the center lap belt is just the most internal lap belt) is equal to constant [(n+1)/2]) at the same time. It is clear that both lap belt$_i$ and lap belt$_{p-i+1}$ are in the same processor when the elements in them are exchanging in swapping procession, and the swapping of the these elements is directly internal exchange without communicating among the relational different processors. It tells us that the cost of taking the communication among these processors will be least.

6.2 Lap belts division with unequal bandwidth

(1) If the bandwidth of the bandwidth of some lap belts is not same, but the element number of every lap belt is about equal, then P_i could save and process one lap belt (i.e. lap $belt_i$) at the same time.

(2) If the bandwidth of all lap belts are not same, but the element number of every equal distance-total lap belt group consisted of lap $belt_i$ and lap $belt_{p-i+1}$ is about equal, then P_i could save and process the elements in the two lap belts (i.e. $belt_i$ and $belt'_{p-i+1}$) without cost expensed.

Therefore, the new method based on lap belts could raise the transposing efficiency to the four belts and whole large matrix.

7 Conclusions

Firstly, both oblique belt division and right angle belts division could be used for raising the transposing efficiency of large matrix better than traditional division.

Second, use parallel computing to deal with the complex problem must take into account the following two questions:

(1) How to construct a parallel computer system;(2) How to divide the complex issues into some sub-issues. Faced with the reality of the complex and volatile world, the first response capacity is limited, but the second have great potential. The excellent division of a complex issue reflects the understanding of the nature of the problem.

Third, the continuing study works are how to design the algorithms based on the new methods of both oblique belt division and right angle belts division, how to realize these new algorithms in computer..

References

[1] He Ling, "High-dimensional spatial data similarity metric," Math practice and understanding, 2006,36(9)

[2] Cheng Guoliang, Parallel computing. Beijing: Higher Education Press, 2002

[3] Zhang Herui, Advanced Algebra. Beijing: Higher Education Press, 2002

[4] D. Takahash, Implementation of multiple-precision parallel division and squareroot on distributed-memory parallel computers. 2000 International Workshops on Parallel Processing, Issue , 2000, pp. 229-235

[5] Xie Ying-Ke, Zhang Tao, Han Cheng-De, "Design and Implementation of Matrix Transposition Unit for Real-Time SAR Image Systems," Journal of Computer Research and Development,2003, 40(1),pp:6-11

[6] Jiang Fan, ,Liu Guang Tao, Zhou Zhi Ming, "Distributed Matrix Transpose on Multicomputer," Microprocessors, 2002(2),pp:34-37

[7] Meng Xiangjie, Zhang Lilun, Zeng Yonghong, "Research of Parallel Matrix Transposition Algorithms in Distributed Memory Computing," COMPUTER ENGINEERING & SCIENCE, 1999,21(5) ,pp:67-71

[8] Wu Fei, "Sparse matrix transposition operation," COMPUTER ENGINEERING & SCIENCE,1989(03),pp:85-92

[9] Zou Qingyun, Huang Xinmin, Zhao Ling, "A Fast Computer Algorithm for Matrix Transposing," Journal of National University of Defense Technology,1990,12(3),pp:76-78

[10] Wen Qing Tao, "Utilizes the great replace to realize the data matrix transposition," .Forest Inventory and Planning,1998,23(2),pp:25-26

Zhou Qihai is a Full Professor, Doctor's (and Master's) tutor and a head of Research Institute of Information Technology Application in School of Economic Information Engineering, Southwestern University of Finance and Economics, China. He graduated from Lanzhou University, China in 1982; and is one of the "One hundred academic and managerial leading heads of China informationalization". He has published 43 books and 183 papers; and is Chair or Organizing Chair of some important international conferences. His research interests are in computational geometry, algorithm study, economics & management computation, isomorphic information processing, and so on. More (in Chinese) about Prof. Zhou is here:
http://graduate.swufe.edu.cn/files/zhouqihai.doc
http://www.cs.swufe.edu.cn/web/teacherinfo.jsp?teacherid=zhouqh

Modelling and Simulation of Distributed Communication Agents

Bippin Makoond[1,2] Souheil Khaddaj[1] Radouane Oudrhiri[2]

1 Faculty of Computing, Information Systems and Mathematics, Kingston University, London, KT1 2EE, UK.
2 Systonomy Ltd, Southbank House, Black Prince Road, London SE1 7SJ, UK.
Email: S.Khaddaj@kingston.ac.uk

Abstract

The task of building distributed application network is plagued by cost, ambiguous requirements and schedule overrun. A fundamental problem in building distributed applications is the presence of unreliable estimation to formulate the list of requirements in the initial stage of project life cycle. Subsequently, this often leads to reactive response from decision-makers against error occurrences, leaving a very small space for developers to establish forward planning. Unless estimations are enhanced to include more sophisticated verification and validation methodologies, the problem will remain. This paper looks at the Petri Net methodology as a modelling and simulation instrument to verify and validate models of distributed application networks.

1 Introduction

Historically, life critical software projects have routinely incurred significant checks, simulation and validation in order to reinforce the safety protocols and this approach does now apply also to critical business applications. In attempting to incorporate the simulation strategies in business critical systems, researchers in software engineering have developed several simulation tools with the intention of improving the effectiveness of estimation. An emerging area of research takes into account methodologies such as formal methods. Languages like B- method and the Z notation [2] provide mathematical formalism to illustrate and evaluate consistencies in software projects.

Building distributed multi-component systems introduces another layer of uncertainty, which very much depends on the underlying network and infrastructure. In this paper we particularly concerned with mobile communication and its infrastructure. The wireless networks with their mobile clients differ from traditional client server models and given that numerous wireless devices are likely to increase [1], the process of integrating those devices under a single server solution can become very complex and inefficient as the network grows. A more practical solution is to operate these mobile devices over a decentralised distributed application system.

In this work, the research questions the requirements that exist in building a highly distributed network applications and investigates if simulation can be resourcefully established to examine the model of a particular network. The work focuses on modeling communication agents based on the blueprint of the Virtual Mobile Redirector (VMR) [4] queue structure and subjects it to Petri Nets [3] simulation to carry out a number of experiments. The experiments include several simulation scenarios which when combined together yield an estimation of the agent’s performance and simulating different agents operating in tandem provides a fairly accurate approximation of the network performance.

The paper starts with a brief description of the problem domain, and Quality of Service (QOS) issues in a networking environment. Then, the simulation and modeling environment is presented. This is followed by describing the simulation process together with a number of experiments. Finally, we conclude with some suggestions for future work.

2 The Problem Domain and Quality of Service Issues

Simulation aims at observing a system enforced to react under a number of predefined circumstances to finally evaluate the behavioural outcome [6]. A key factor, however, prior to simulation, is to underline what features of the system one needs to simulate in order to maximise knowledge capture with a minimum amount of resources. Considering a distributed application network, at the highest level of observation, the features that a simulator should question are distinguished by five entities [7]. First, for a given network architecture, the task requires knowing the amount of data entering and leaving the system. This is termed as the traffic rate. For example, a network designer is required to decide on the size of a queue component over a network against the traffic rate. Second, the traffic inside the network may be congested and the designer would be required to know the congestion rate so as to alter the specification of the network for better performance. Third, the rate at which the data is flowing inside the network is also crucial to identify the latency that through simulation can be calibrated to an acceptable level. Fourth, is the throughput; the designer will need to know the delivery rate of the network under different condition. Throughput is usually the first measure that the customer investigates [1]. Finally, data loss is yet another important factor that illustrates how reliable the network system is. The above five entities form the programme's framework and are termed as the Quality of Service [7].

The Quality of Service of a network is described by its performance, i.e. how efficiently the system delivers its service to the world. Hence, acquiring knowledge on the quality of a system prior to product deployment would be competitive gain for any organisation building network application. We propose that to measure the Quality of Service of a distributed network application at the early stage of development imply the use of simulation tools since a real network system is not available at that level. This can be done using the five entities described earlier; the simulation engine should derive test scenarios to observe the network behaviour and reaction when any of the entities is questioned. The outcome from the simulation is analysed to develop contingencies and self-protecting mechanism when and where required [5]. In this paper we refer to these contingencies as handlers [7]. Each of the entities has their handlers. We hypothesise that a successful simulation procedure should be able to search and locate deficiencies related to the any of the five entities and next propose, with an acceptable degree of error, the appropriate handlers for recovery. The simulation can then test out such a procedure using four different measures to handle the five entities:

(i) Traffic metering to control and monitor traffic rate.

(ii) Resource allocation strategy to boost resource performance should the rate of throughput falls.

(iii) Data drop policy to regulate the rate of data loss.

(iv) Switching / Routing (Re-routing) strategy to alleviate congestion rate over the network.

(v) Resource allocation strategy to reduce network latency.

To evaluate the quality of service of a distributed network application, the experiments should look at three features of performance; namely the entities traffic rate, throughput and latency. The simulation scenarios position the Petri Net model of the network application under different situations whilst the experiments distinctively track the network behaviour and relate the relevant handlers for each situation.

3 The Simulation Environment

In the last three decades research in computer science strongly contributed to the generation of simulation tools. Several formalisms and related automatic tools have been developed, others extended to accommodate simulation engines and are currently available on the market. Within the wide spectrum of simulation techniques, we were required to focuses on a selected subset of simulation tools, which represents, either for historical reasons or recently achieved popularity, and subsequently choose a platform that best agrees with the basic entities needed to model and simulate a distributed network application., which is Petri Nets in our case.

Petri Nets are a formal modelling technique for the description of concurrent and distributed system behaviour [3]. Since their introduction in the 1960s they have impressively evolved. Currently, several versions of this graphical modelling language exist, which find widespread application in specification, verification, and performance analysis of distributed parallel systems, communication protocols included. Typical applications of Petri Net are in the field of communication protocols, telecommunication networks, and software engineering. Despite a modelling philosophy that is not easy to be acquired, Petri Nets, with their engaging graphical notation are a powerful modelling language to describe and investigate communicating and resource-sharing processes. Coloured Petri Net combines the basic formalism with programming language CPN ML [3] to enable the creation of accurate system models taking into account temporal and stochastic issues. Similar to other graphical tools, embedding user-defined functions into the specification is an interesting possibility to avoid heavy graphical descriptions of deterministic calculations. For undertaking the experiments, we considered CPN Tools, a freeware Coloured Petri Net tool available on the market. CPN Tools is a new platform for modelling and simulation of Petri Nets, which allows editing, simulating and analysing Coloured Petri Nets (CPN).

4 Simulation and Observations

The models as well as the procedures of the experiments are discussed in this section. The aim is to carry out an investigation as to how Petri Nets are utilised to simulate the VMR multi agent system (VMR MAS) at the highest level and the communication agent queue component at the lowest one. Thus, the information from the exercise can be statistically interpreted so that handlers (contingencies) can be formulated for the necessary QoS measures .

4.1 Modelling and Simulation

When making use of the hierarchical decomposition feature of CPN, we are able to describe the architecture of an agent-distributed system at different level of perceptions. At the highest level of expression, we modelled the incoming of data packets from the external networks to the VMR system which is illustrated in Figure 1.

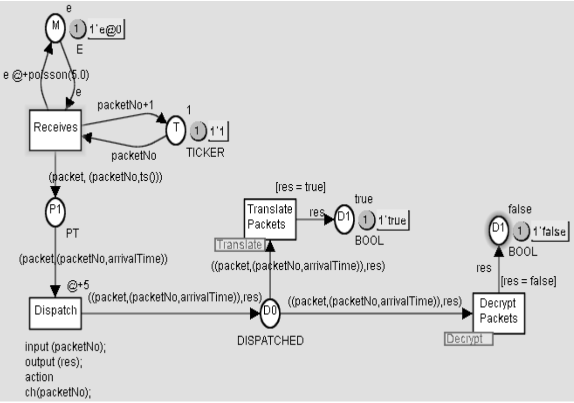


Figure 1 CPN Model of the VMR MAS

The CPN model starts with the transition Dispatch receiving one packet at a random time, which follows a Poisson distribution with a mean of 5.0 ms, from an external source. Since the model is a timed Petri Net, variable E at place M is timed in millisecond whilst each packet is numbered at the place TICKER, thus the place Dispatch, knows about the packet number (packetNo). Next, at transition Dispatch, a function ch() is triggered to randomly select between 2 transition operations, either Translate Packet or Decrypt Packet. These functions add more knowledge to the model; hence during simulations more complex processes can be investigated in an attempt to mimic reality as closely as possible. Upon exit of the transition Dispatch, the place DISPATCHED is fired. Inside the transitions Translate Packets and Decrypt Packets, there are a series of distinct software agents communicating to each other with a common objective. In the VMR system, operations like “translate the packet structure to a general format” are performed by a cluster of agents. In an attempt to simulate these functions, we constructed a hierarchical transition Translate Packet within which are a group of agents connected to each other, and the same design applies to the transition Decrypt Packet.

Figure 2, depicts the internal design of the transition Translate Packet, wherein, several agents are non-linearly assembled. Packets move concurrently across the agent system which is supported by the CPN infrastructure for such dynamic data movements. An agent receives the packetNo and the arrivalTime from the place DISPATCHED. As the packet travels from one agent to another, the task of translating a message to a generic format is distributed among various agents.

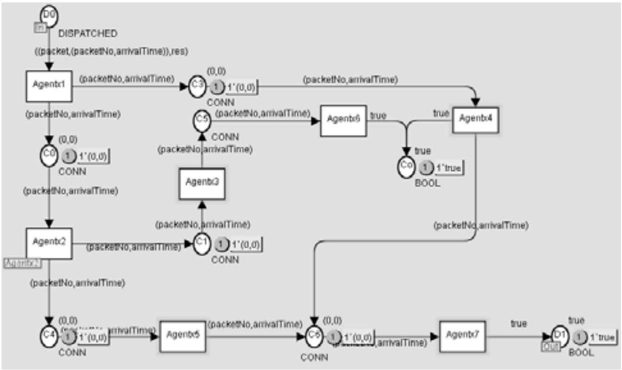


Figure 2 CPN Model of Agent Communication

Drilling further through the Petri Net hierarchy, we find the design of one agent which triggers several processes towards translating a packet into different formats. However, for simplicity we only present a model of the lowest level of the hierarchy, the FIFO queue structure (Figure 3), which is the communication medium of each agent, found in the Sensor class. In an agent we expressed the transition Queue that buffers all the packets that reaches the agent. The queue component resides within the sensor class of an agent with the value vh, being dynamic and adjustable during simulation, indicating the top stop mark capacity of the queue’s buffer and when hit, it blocks the packet entry. Unless the buffer is freed, outgoing packets have to wait for their acceptance. Another watermark is the high watermark denoted by h, and when reached, the rate of incoming packet is decreased until the number of packets inside the queue’s buffer reaches the low water mark which indicates that the queue is within normal parameters of operation and no protective measures are required.

When the Transition Packet Arrives is fired, the queue receives a new packet, hence a token is added to the place arrived. Upon arrival, each packet is paired with a number, (see place packet stamped), which contains a time stamp and it is equal to the current model time, created by means of the function timeStamp(), on the arc between Packet Arrives and arrived. The place arrived characterises the packets that reached the queue and are waiting for clearance. Next, the queue pushes the packet into its buffer and the waiting line is shown by the occurrence of the transition PushIntoBuffer. The arcs between Next Packet and PushIntoBuffer ensure that packets are loaded into the queue in the same order in which they joined which enforces the FIFO structure. When packets are loaded onto the buffer, shown by the place buffered, they are counted (see the inscriptions on the arcs between PushIntoBuffer and buffered, noOfPackets). At this point, packets are inside the buffer and waiting for the next run where the queue might perform some simple processing to the packets before releasing them. Once one packet completes its processing cycle, a token is added to the place leaves. Any claims from the place, causes the transition Claims Packets to be fired which finally state that a packet has left the queue.

The model is an integer-timed Petri Nets, i.e. the queue is processed with relation to time. The Poisson function, at the entry point is used to model the inter-arrival time for packets, and it can be found in the code segment associated with Packet Arrives. By changing the value in the function’s argument, the mean of the Poisson is altered and thus providing different spectrum of simulation conditions. The function CalcAccTime() in the code segment for the transition PushIntoBuffer to calculate how much time is needed for loading a given packet. The function NormDistr() is used to calculate how much time is used for a packet to be processed, as shown in the code segment for Process.

This is particularly useful when one needs to make a number of simulations using different parameters.

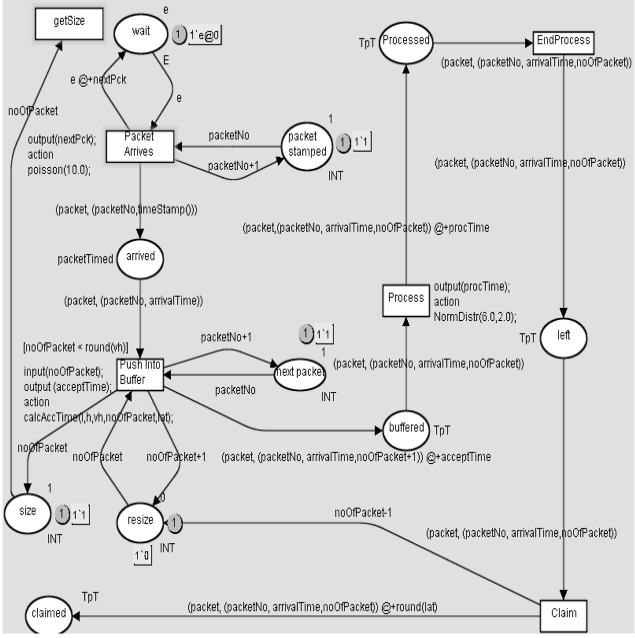

getSize
wait
e
1
1`e@0
E
e @+nextPck
e
noOfPacket
TpT
Processed
EndProcess
(packet, (packetNo, arrivalTime,noOfPacket))
Packet Arrives
output(nextPck);
action
poisson(10.0);
packetNo
packetNo+1
packet stamped
1
1`1
INT
(packet, (packetNo, arrivalTime,noOfPacket))
(packet, (packetNo,timeStamp()))
(packet,(packetNo, arrivalTime,noOfPacket)) @+procTime
packetTimed
arrived
(packet, (packetNo, arrivalTime))
Process
output(procTime);
action
NormDistr(6.0,2.0);
TpT
left
[noOfPacket < round(vh)]
packetNo+1
1`1
next packet
(packet, (packetNo, arrivalTime,noOfPacket))
Push Into Buffer
input(noOfPacket);
output (acceptTime);
action
calcAccTime(l,h,vh,noOfPacket,lat);
packetNo
INT
buffered
TpT
noOfPacket
noOfPacket
noOfPacket+1
(packet, (packetNo, arrivalTime,noOfPacket+1)) @+acceptTime
0
size
1`1
INT
resize
INT
1`0
noOfPacket-1
(packet, (packetNo, arrivalTime,noOfPacket))
TpT
claimed
(packet, (packetNo, arrivalTime,noOfPacket)) @+round(lat)
Claim

Figure 3 CPN Model of a Queue system within an Agent

In summary the objectives of the simulation of the VMR CPN model are as follows:

- To observe the percentage time a unit (packet) leaves an agent given a high input rate. This is to probe the information on the throughput of the queue during high input to the VMR system.
- To observe the average time a unit (packet) spends in the VMR System. This is to provide information on the likelihood of out-of-time packet(s), should the latter wait excessively in any queue buffer of any agent.

4.2 Results and Observations

The simulation exercises investigate how packets move amongst the VMR Agents, with variable parameters at different level of hierarchy. The model was fed with packets at random intervals and the output was observed and recorded. The experiments were broken down into two series (experiment 1 & 2).

- Experiment 1 explains that the queues inside the agents resorted to protect themselves whenever the time interval between incoming packets was smaller than the service time of the queue, which consequently resulted in some packets being rejected.
- Experiment 2 stresses on the likelihood of out-of-time packet in the VMR System. This is defined by a function of the agent's internal latency, queue's buffer size and the input rate at the receiving end. It confirms that while attempting to gracefully slow down the number of packets entering the buffer, should an overflow occur; the queue inside an agent consequently causes packets to wait longer in its buffer.

Experiment 1 starts by investigating the percentage unit time (ms) a packet leaves the multi agent system given a burst of input. The objective is to observe the throughput of the system in unconditional circumstances and evaluate its behaviour and efficiency to protect itself. We run through three simulations of 4000 steps and at step 1000 , we changed the Poisson mean from 1 packet per 10ms to 1 packet per 3ms, which provided the scenario for the burst and is the type of situation common to the domain of messaging. At each simulation we decreased the value of the watermarks, thus decreasing the waiting line in view of achieving conditions where packets have to wait longer to be serviced.

Figure 4 runs through the simulation as the buffer sizes of the queues inside the agents were increased. We sampled the data into batches of 100 ms interval and recorded the number of packets entering and leaving the queue. Given a burst in packet entry, the graph shows the difference between the input rate and the output rate. The close proximity between the input and output rate justifies that that the service rate of queues within the VMR system is adequately lower than the input rate given an input rate of 100 packets per second (1 pck/10ms) and a burst of approximately 350 packets per second (1 pck/3ms). The simulation validates both decisions on the buffer size of lower limit 10 and upper limit 20.

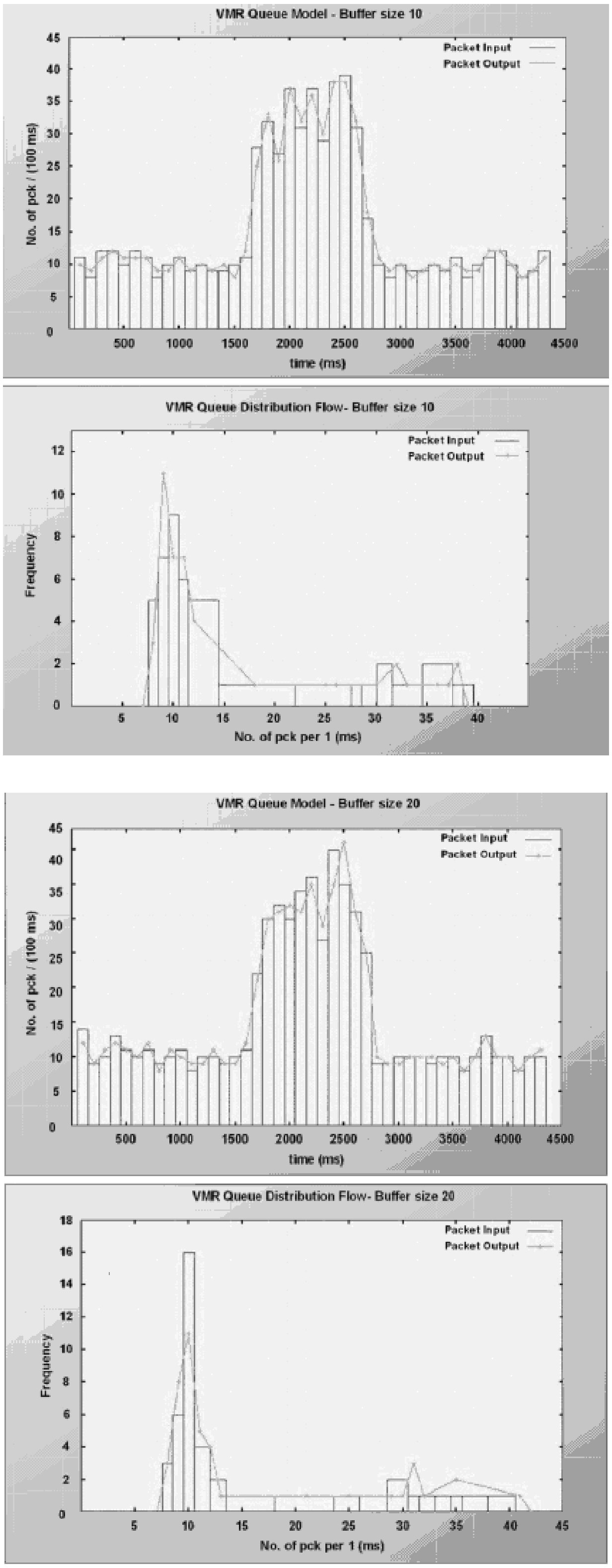


Figure 4 Observation of Queues performance

Experiment 2 investigates the period of time a packet stays in the VMR at a given queue of an agent and draws the distribution of the packet's lifetime as the Figure 5 depicts. The graph reports on the time taken for a large sample of packets to leave any queue of an agent. The objective was to estimate the number of out-of-time packet that exists given an input burst and a decrease in the buffer size. The graph shows the period of time a packet takes to leave an agent, e.g. over hundred packets take 12 ms to leave the queue of a VMR agent. Hence using such analysis, a threshold can be established to identify the packets that have a probability of being out-of-time when claimed. For instance, should the out-of-time threshold be marked at 15 ms, any packet beyond this mark is irrelevant for the system.

Given a series of requirements, at the early stage of the life cycle, we found out that by translating these requirements into Critical Parameters, and then to Petri Net simulation models, the different behavioural patterns of the models can be observed. These observations provided more knowledge to enforce better accuracy for quality estimations, in terms of QoS and CTQs.

The experiments also show how Petri Nets as a methodology can be exploited to mine appropriate data so that the potential behaviour of a model under a set of parameters can be known. The obtained results provide a very good idea about the performance of the model, hence the formulation of directives prior to product deployment. If we look back at experiment 1 where a discussion on burst of input rate is presented, this is often very true in the mobile telecommunication and messaging systems. For instance, if we consider the message entry to a SMS Centre (SMSC) for one day, there is a pattern, wherein bursts can be found during the morning when people wake up and starts sending text message and burst at lunch time during their break time. To deal with the surge, system designers over-estimate the buffer size of queues within their systems, which is very inefficient in terms of resource allocation and economic viability. However, using simulation based on historical data of typical message distribution, one can provide a range, a lower limit specification size for off peak period and upper limit specification size for peak period. This is type of design is called design for capacity, which is a competitive advantage if possessed.

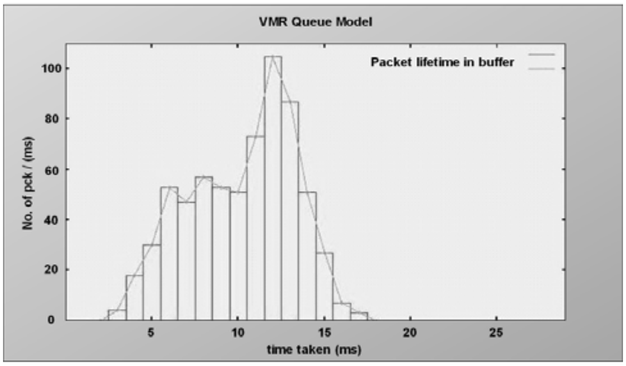


Figure 5 Queue Buffer Population

5 Conclusions

Given a series of requirements at the early stage of the software life cycle, we found out that by translating these requirements into Petri Net models and simulation, we could observe the different behavioural patterns of the models. These observations provided more knowledge to enforce better accuracy to estimations particularly at the early stages. In terms of QoS, it shows how Petri Nets as a methodology can be exploited to mine appropriate data so that the potential behaviour of a model under a set of parameters can be known. The obtained results provide a very good idea about the performance of the model prior to product deployment.

As it can be observed, Petri Net is an adaptable tool, i.e. it allows alterations to a model so as to retrieve the necessary information from it. This paper presented two experiments on how simulation can be executed to question a model. However, if properly handled, Petri Nets can accommodate more complex scenarios without losing the real description of a model. This is sometime not true for other methodologies. Through the base language CPN ML, Petri Net can be customised according to the project needs. This is essential, as some methodologies force the designer to adapt the project needs to their specification.

References

[1] Price Water House Cooper, “Technology Forecast: 2002 –2004, Vol 1: Navigating the future of Software”, 2002

[2] J. M. Wing, "A Specifier's Introduction to Formal Methods", IEEE Computer, vol. 23, no. 9, Sept. 1990

[3] Kurt Jensen, “Coloured Petri Net – Basic Concepts, Analysis Methods and Practical Use ”, Vol 1, 2nd Edition, Springer, 1997

[4] Empower Interactive Group Ltd, “Virtual Mobile Redirector Manual Specification Guide”, 2002

[5] Roger S. Pressman, “Software Engineering, A Practitoner’s Approach”, 5th Edition, McGrawHill, 2000

[6] Simulation concepts – http://simulation.artshot.com/economy/

[7] Cisco Systems, “ IP Quality of Service, – The complete resource for understanding and deploying IP quality of service for cisco networks”, Cisco Press, 2001

[8] J. Ellsberger, D. Hogrefe, and A. Sarma, SDL Formal Object Oriented Language for Communication Systems, Prentice Hall, 1997

[9] C-M. Huang and J-M. Hsu, "An Estelle-Based Probabilistic Partial Timed Protocol Verification System," Proc. 7th Int'l. Conf. Parallel and Distributed Systems, 2000

[10] T. Bolognesi and E. Brinskma, "Introduction to the ISO Specification Language LOTOS," Computer Networks and ISDN Systems, vol. 14, Apr. 1987

[11] L. Millett and T. Teitelbaum, "Slicing Promela and Its Applications to Protocol Understanding and Analysis," Proc. 4th SPIN Wksp., Paris, France, 1998. http://spinroot.com

[12] Nam,S.Y., Sung,D.K., “Measurement Based Delay Performance Estimation in ATM Networks”, Globecom, Nov. 2000, pp.1766-70

[13] Mouharam, A.A.K., “Monitoring Quality of Service on Broadband Networks”, PhD Thesis, Kingston University, 2002

Research of Single Source Application Layer Multicast Protocol

Kunhua Zhu

School of Information Engineering, Henan Institute of Science and Technology
Xinxiang, Henan Province 453003, China
Email: zwkh100@163.com

Abstract

After simply discussing the merits and demerits of application layer multicast, we propose a single source multicast protocol in this article, we call it SSMPAL. In this protocol tree topology is given the priority to construct multicast transmit tree. As to the maintenance of multicast tree, we use the PUP algorithm which selects a backup father node for every non-leaf node to set up a spare linker. This protocol, which inherits the merit of application layer multicast and to a certain extent overcomes the instability of application layer multicast, greatly improves the stability and reliability of multicast tree.

Keywords: Pre-Using-Parent Algorithm; Multicast Tree; Single Souse Multicast; Father Node; Application Layer Multicast

1 Introduction

Since its birth the concept of multicast has been being a hotspot on the research of Internet applications, people regarded it as a effective prescription to solve the problems that caused by unicast communication such as the serious waste of bandwidth and inefficiency. In 1988 Steve Deeing first proposed the idea of using routers in the network layer to carry out the multicast idea, this is called IP Multicast. In this model, routers responsible for the maintenance of multicast tree, when the data packets are sent out from the data sources, they are reproduced in different tree node router before finally achieve the terminal. In IP Multicast protocol, incorporating transmission of repetitious information effectively reduces the replication of data packets, reducing bandwidth waste and improving the efficiency of service. So IP Multicast has been considered to be the most effective. However, after 10 years′ research, IP Multicast, not as expected will be widely applied to the Internet, come to an end as the technical areas it relates to and its monitoring and management are too complicated, also the service and charge model are lack of market-driven and customer demand. For this reason, the focus of the study of Multicast has transferred to the application layer in recent years.

The data of application layer multicast are reproduced in the terminal system and then transmitted, instead of using router to maintain the status of multicast group, terminal system is used to take charge of the administration of the members of the multicast group, in this way the problem of the expansion of the business is overcome. At the same time, application-layer multicast can be deployed at any time, without upgrading or updating the network equipment. Although the status of bandwidth waste of application layer multicast is more serious than that of IP Multicast, the development and deployment with simple, therefore, bring it a good prospect. Application layer multicast, from the whole, are divided into two modes single-source and multi-source Multicast. This paper proposes an algorithm design for single-source multicast protocol SSMPAL (Single Source Multicast Protocol of Application Layer), which support the application of large-scale user streaming media.

2 The Framework Design and Characteristics of SSMPAL

SSMPAL inherited the concept of application layer multicast, can be more efficient in multicast transmission, without the participation of router or any change in the network infrastructure, the restrictions of application is also greatly reduced. SSMPAL is a protocol that aims at the single source application layer multicast, that is to say there can be only one sender at the same time in a specific multicast group, this, to a certain extent, reduces the complexity of the protocol and greatly simplifies the process of the realization of the protocol, All this together make SSMPAL a protocol of better capability. Group members are organized together in a tree topology preferential way. In this protocol, each node is a member of the client and at the same time a server, each node must achieve three functions: receiving data stream, actualizing the feedback, transmitting data stream. In general stream media application continues for a long time, and needs a stable bandwidth, so this protocol algorithm proposes a new view in the maintenance of the reliability of application layer multicast transmitted tree.

3 The Main Algorithm Design of SSMPAL

In the protocol each node is logically divided into two topology structure: Control topology and Data-transmit topology, each line in the topology graph stands for a unicast connection. The Control topology is used to maintain the multicast transmit tree and enhance the reliability of the protocol by periodically exchanging the control information. The Data-transmit topology is mainly used to show the transmission route of the multicast data packet. A shared data topology tree is firstly established in the protocol. The primary task for each node who wants to join in is to find a suitable father node. Then, after some control information between the members got added, the Control topology is established. According to the process of members of the multicast nodes participating in multicasting, the main description of the algorithm of SSMPAL is as following.

3.1 Nodes join multicast group

To join the group the entrant must in some way know the IP address of the root of the group which they want to join, when a new member is joining, it will get the related information of the root node from a RIP (Reserve Information Point, this point will retain the information of the root node, its parent node, its all children, and some other adjacent points), then after a series of action such as requesting to connect, sending the inquiry, feeding back etc, its father nodes is finally chosen in accordance with the evaluation and the guideline that established by the protocol. When a particular member in the group becomes the father node of the new node, the new one is successfully added to the multicast group. At the moment of successfully get the father node, the new node also have to go through some calculation to find its "spare parent node." The steps of parent node selecting arithmetic are as following:

(1) Suppose the member node A receives enquiry message that generated by the new entrant who is joining the group and inquires about the relevant information about the root of the group.

(2) After receiving the query message, node A searches for the address of the root and the restrictions that established by the root for QOS (such as delay, bandwidth, etc.) according to the information that mastered by itself, and then create feed-back messages, send the query results back to the new entrant.

(3) After getting the address of the root, the entrant creates query-father message and send it to the root.

(4) The root will inquire about the information of its adjacent points it maintained, finding out the nodes whose number of children is less than the number of their degree making them "the possible father node" (PF, Possible Father), if this step is not carried on to the next step, until it find out all of the PF of the tree, and then create feed-back message sending the information such as the set of "the possible father node" (SPF, Set of Possible Father), the root node, and other relevant

information back to the entrant.

(5) According to the feed-back message, the entrant send detection message to every PF, detecting the delay and bandwidth of the direct path between the entrant and the PF.

(6) Every PF sends feed-back messages to the entrant, returning the result of detection. When the entrant in succession receives these feed-back messages from the PF, the entrant chooses its own father node under the principle that choosing the maximum bandwidth after choosing the minimum delay. Then through some algorithms, one node of the members is chosen as the spare father node of the entrant, in case of any emergency, to reconstruct the tree.

(7) After the father node is chosen, the entrant sends Subscribe message to the father node. And the father node enrolls the entrant on the list of his children which will be maintained by him .In this way the entrant is added to the transmit tree.

(8) When the action of join is completed, the father node sends succeed-Join message to the entrant showing that the join is successful. In the algorithm, the scope from which the father-node is chosen is the whole transmit tree. two conditions must be met before choosing a suitable father node for an entrant: one is that if a node is chosen as the father node of a new member, there should not be any cycle in topology graph of the multicast tree; Another, if the entrant is added to the children of a node of the tree, the degree of the member node should be legitimate. If the new member finds out a number of suitable father nodes, it will choose the most suitable one according to the requirements of measurement in specific conditions, and it also goes through some algorithms to find out another node as the spare father node of it.

3.2 The text of the broadcast transmission and playback of multimedia streams

When a node joins the multicast tree, It will be able to receive the data from the root node. The multimedia streams from the Data source will generate multicast text through acquisition, encoding, and other processes, then the text will be sent to root along the direct path, next via multicast tree to leave node from top to bottom. After receiving data texts from its father node, each node will transmit them to all its sub-node, and it ends in the leaf nodes. If the received data at one certain node is not from the parent node, it will be discarded. When a node is receiving and sending data streams, it gives the multicast data back to multimedia streams to display. You can choose to adopt TCP or UDP According to the upper data applications when transmitting, if the upper class is the document distribution system, you can choose the reliable TCP protocol; if it is the audio / video broadcasting system, and you can choose the UDP protocol which consumes little.

3.3 The voluntary drop-out of the multicast member node and the multicast tree maintenance from accidental failure

In the application layer multicast tree, the tree ruptures because of the voluntary drop-out of the nodes and the accidental failure, which is inevitable. Therefore, an effective mechanism is requested to rebuild the entire multicast tree and reduce the data lose to the minimum after the nodes failure. Multicast tree can be reconstructed in two ways, that is the backwards and proactive approach the former one means to rebuild the tree after the node failure. This remedial approach will spend a lot of time and the data will lose simultaneously. So the proactive approach is taken in this algorithm, which means to work out the new father node in advance before the node failure, once the father node of the node drops out, you can immediately find out the new father node according to the recalculated value, thus avoiding data lose and spending too much time searching the new node.

It is necessary to preschooler a spare node, once the father node fails and drops out, you can start the spare node at once; The thought of PUP(Pre-using-Parent)algorithm is introduced to the choice of “standby father node”, which is described as follows:

Assume the n is the node of the “standby father node” and the p is the index that refers to certain node.

(1) Algorithm starts are the parameter which represents the wanted node p.

(2) First refer index n to node p.

(3) If p has no brother node, go along the tree upwards.

(4) P is the wanted node if p has a reserved link, algorithm ends; otherwise, if node p has a brother node, the cycle ends as well. Otherwise, continue to walk up, it will reach the root node with ending cycle if there is no appropriate node along the path. It is purposeful to find out the brother node. The "standby father node" of the node chooses the ancestor node except the father node in priority, and then is the brother node of the ancestor node. Among the upper class nodes beyond n, only the father node of n can not act as the "standby father node", while all other nodes can.

(5) If no suitable p appears until the root node, the process ends and returns 0.

(6) If such a p emerges, write down the information of its brother node into a queue q, where lists the candidate nodes. and node n has the priority to take out nodes in queue q for comparison and selection

(7) Start to take out nodes in turn in queue q.

(8) Refer p to the node from the queue.

(9) If there is reserved link of node p, p is the wanted node, skip the cycle. Otherwise, join the son node in the queue q, restart the cycle from the beginning. if no suitable node is found, return 0,and algorithm ends.

The algorithm is realized with PARSCAL language as follows:

```
procedure PUP(n)
p = n.parent;
while p.siblings is null
do
{p = p.parent;
if p.dp>0 then
return p;
end if
until p=null;
if p.siblings = null then
no link provided;
exit;
else
add_queue_item(q,p.siblings);
end if
}
while q <> null
do
{p = get_queue_item(q);
  if p.dp > 0 then
return p;
  else
add_queue_item(q,p.children);
loop;
return 0; }
end procedure;
```

When the "standby father node" is found with the above algorithm, it can be started immediately once some nodes leave or fail, so that the multicast tree will not collapse, which takes the effect of maintenance and precaution.

The algorithm steps for the establishment of links are described as follows:

(1) Assume that n is the node of the "standby father node". Then work out the node p that acts as "standby father node" through the PUP algorithm.

(2) Node n sends a "JOINT" text to node p;

(3) Node p sends Accept or Refuse text to node n to tell it whether to establish the obligated link path.

(4) After being received, node n refers its "standby father node" to node p, and then sends ACKNOWLEDGEMENT text to node p.

(5) Node p revises its relevant information and establishes the obligated link path to node n when it receives the confirmed message.

(6) Algorithm ends.

It is necessary to state that the following method is taken in the maintenance of standby link:

When a node finds out its "standby father node" thorough the PUP algorithm, the father node is required to send keep alive text in certain intervals so as to keep in contact with each other. Once the node can not receive the keep -alive text in certain time, the "standby father node" can be considered as a failure, thus stimulating the PUP algorithm to function again for the new spare node.

According to the two algorithms above, in the reconstruction strategy of the multicast tree, the member node can also receive the data text transmitted over the obligated link path, then lots of nodes will get two identical data texts at the same time, while the node can discard the text from the obligated link path. As long as the father node of the node fails, the node will apply for a data transmission to the father node, and then start to use the data from the obligated link path. The multicast text will not lose because of the absence of certain nodes in this way.

4 Concluding remarks

SSMPAL protocol has been put into practice in the laboratory on less than 10 machines. It turned out to that this protocol operated well in a small area. Along with the experiments, we also compared the maintenance and fault-tolerant mechanisms between the PUP algorithm and reserved chains algorithm which used in the agreement and the multicast trees used in the multicast protocol of some other applications. The result of the comparison proved that there is a slight increase in the average delay of the transmitted trees' node. But the reliability of the maintenance is superior to other algorithms. The algorithms are simple and easy to achieve. When a new node is added, it will launch the PUP algorithm automatically to find a backup link instead of affected by its generations. The constructions of the trees' control topology and date transmit ion topology are clear and nature.

The tests of SSMPAL on large-scale machines haven't been completed. The purpose of the agreement is to realize large-scale live broadcast video systems based on application layer multicast. It needs further test and valid on whether it can apply to the internet and the correctness and performance of the protocol.

The inaction of this in the come up of the single source application layer multicast protocol SSMPAL. In this protocol, come up a algorithm which called PUP algorithm that pre-select a "standby father node" for each tree node to set reserved chain mot only make the protocol inherits the merits of the application layer multicast, it also, to some extent, overcome the instability features of the application layer multicast and greatly promote the stability and reliability of the multicast.

References

[1] Miao Zhang, Mingwei Xu, JianPing Wu. "Application layer multicast Review," Electronic journals, 2004, 32 (12A): 22-23

[2] Pendarakis D, Shi S, VermaD, et al. ALMI: "An Application Level Multicast Infrastructure," Proceedings of 3rd, USENIX Symposium on Internet Technologies and Systems, 2001

[3] BANERJEE S, BHATTACHARJEE B. "Analysis of the NICE Application Layer Multicast Protocol," Technical report, UMIACSTR 2002260 and CS2TR 4380, the Department of Computer Science University of Maryland, College Park, June 2002

[4] Chang Jin-pyng. Such as Xu-Dong Ma. "Intelligent building system software and network video monitoring service Works for," Micro-computer information, 2004, 20 (11): 145

[5] Suman Banerjee, Bobby Bhattacharjee, Christopher Kommareddy. Scalable Applicati-on Layer Multicast. In: ACMSigcomm, 2002-08

[6] JXTA Jave Standard Edition2.3.4Javadoo[EB/OL]. http: //platform.jxta.0rg/n0nav/java/api/jndex.html,2005

[7] Gong L. "JXTA: A Network Programming Environment". IEEE Internet Computing,2001,15(3): 88-95

[8] K. L. Morse, M. Zyda. "Multicast grouping for data distribution management". In Proc of Computer Simulation Methods and Applications Conference, 2000

[9] Y. H. Chu, S.G. Rao, S. Seshan, H. Zhang. A case for end system multicast. IEEE Journal on Selected Areas in Communications, Vol. 20, No. 8, 2002: pp.1456-1471

[10] L. Wei, D. Estrin. The trade-offs of multicast trees and algorithms. In Proc of International Conference on Computer Communications Networks. 1994

[11] P. Francis, Y. Pryadkin, P. Radoslavov, R. Govindan, B. Lindell. YOID: your own internet distribution. In Proc of Peer to Peer workshop, 2002

[12] D. Waitzman, C. Partridge, S. Deering. Distance Vector Routing. RFC 1075, 1988

An Improved Pastry Network Model*

Sijia Liu[1] Xingwei Liu[1,2] Fang Xia[1]

1 School of Mathematics & Computer Engineering, Xihua University, Chengdu, Sichuan ,610039, P. R. China

2 Key Lab of Information Coding & Transmission Sichuan province, Southwest Jiaotong University, Chengdu, Sichuan ,610031, P. R. China
Email: scarlett805@gmail.com

Abstract
Pastry is a generic peer-to-peer distributed object location and routing scheme, based on a self-organizing overlay network of nodes connected to the Internet. However, the traditional Pastry design is a flat overlay network and limited-scope searching, in which all nodes have identical capabilities and responsibilities and all communication is symmetric. But in practice, because of a variety of available bandwidth and delay, each node has different capabilities. Therefore, in this paper, a hierarchical Pastry network model is presented to address this limitation. Furthermore, the routing procedure and handling algorithm of the arrival and departure of nodes will be carefully discussed. Finally, some elementary experimental results obtained with a prototype implementation on JXTA platform indicate that the improved Pastry network model is superior in routing performance.

Keywords: Distributed system, P2P, Pastry, DHT, JXTA

1 Introduction

Pastry is a generic peer-to-peer object location and application-level routing scheme, based on a self-organizing overlay network of nodes connected to the Internet. Each node in the Pastry peer-to-peer overlay network is assigned a 128-bit node identifier (nodeId) which is used to indicate a node's position in a circular nodeId space. NodeIds could be generated by computing a cryptographic hash of the node's public key or its IP address. Each Pastry node maintains a routing table, a neighborhood set and a leaf set. In Pastry system, files are kept on certain nodes, and the system routes to the node using a mapping principle between file identifier and nodeId. The routing procedure can get the searching result efficiently, and the expected number of routing steps is O(logN), where N is the number of nodes in the Pastry network [1,2].

However, the traditional Pastry design is a completely decentralized flat overlay network and limited-scope searching, in which all nodes have identical capabilities and responsibilities and all communication is symmetric. But in practice, because of a variety of available bandwidth and delay, each node has different capabilities [3,[4]. Therefore, in this paper a hierarchical overlay network model is presented to address this limitation in traditional Pastry.

The remainder of this paper is organized as follows: section 2 presents an improved Pastry model, including a careful description of the routing procedure and handling algorithm of the arrival and departure of nodes. Then, some elementary experimental results with a prototype implementation on JXTA [5] platform are presented in section 3. Related work is discussed in section 4, and this paper ends with conclusions drawn from this work and future work [6,7].

2 An Improved Pastry Model

We have learned that Napster and Gnutella use two

* This research was partly supported by the Scientific Research Fund of Sichuan Provincial Education Department under Grant 07ZA115，07ZC050.

diametrically opposite approaches for locating content. Napster uses a centralized directory server and always locates content when it is present in some participating node. Gnutella uses a fully distributed architecture, but only locates content in nearby nodes in the overlay network [8]-[10]. An improved Pastry borrows ideas from both Napster and Gnutella, resulting in a hierarchical overlay network, which is shown in Fig 1.

A hierarchical Pastry network model resembles Gnutella in the sense that it does not use a dedicated server for tracking and locating content. However, unlike Gnutella, in improved Pastry not all nodes are equal. The more powerful nodes, which have high bandwidth connections and high Internet connectivity, are designated as group leaders (in this paper, named as search node) and have greater responsibilities. If a node is not a search node, then it is an ordinary node or index node (including internal and external index nodes) and is assigned to a search node. Typically a search node will have up to a few hundred ordinary nodes or index nodes. In this way, each search node becomes a mini Napster-like hub. Furthermore, the search nodes interconnect themselves, creating an overlay network among the search nodes. Thus the search nodes create a network that resembles a Gnutella network.

In Figure 1, unlike traditional Pastry nodes, each ordinary node only maintains a record table about the search node of this group, and each internal index node only maintains a record table about all ordinary nodes' resources and nodeIds of relative external index nodes. While each external index node maintains a record table about nodeIds of ordinary nodes and search node in this group. In this improved Pastry network model, only the search node, like traditional Pastry node, still maintains a routing table, a neighborhood set and a leaf set.

2.1 Search node state

The routing table is organized into $\log_{2^b} N$ rows with $2^b - 1$ entries each. However, unlike traditional Pastry, the shaded cell in each row of the routing table shows the corresponding digit of the external index node's nodeId, and the other cells contain the nodeId of search node. Each entry in the routing table contains the one IP address of the potentially nodes whose nodeId have the appropriate prefix.

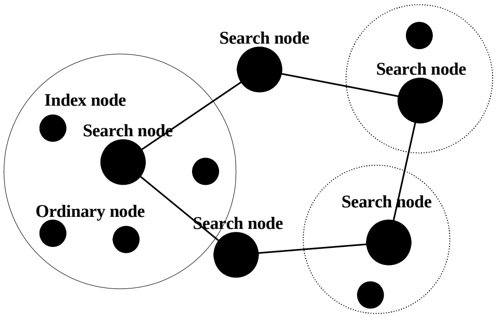


Figure 1 A hierarchical Pastry network model

The neighborhood set contains the nodeIds and IP addresses of the search nodes of neighborhood groups, and the neighborhood set is not normally used in routing messages. The leaf set is composed of the external index nodes and search nodes, and the leaf set is used during the message routing.

2.2 Locating and routing

One of the key problems in large-scale peer-to-peer applications is to provide efficient algorithms for object location and routing within the network. The procedure is executed whenever a message with key sent by an ordinary node arrives at a search node:

1) Firstly, the search node checks to see if the key falls within the range of nodeIds covered by internal index node. If so, the procedure directly goes to step 3).

2) Then, if the key is not covered by the internal index node, then the search node first checks to see if the key falls within the range of nodeIds covered by its leaf set. If so, the message is forwarded directly to the destination node. If the key is not covered by the leaf set, then the routing table is used and the message is forwarded to a node that shares a common prefix with the key by at least one more digit. In certain cases, it is possible that the appropriate entry in the routing table is empty or the associated node is not reachable, in which case the message is forwarded to a node that shares a prefix with the key at least as long as the local node, and is numerically closer to the key than the present node's id. Such a node must be in the leaf set unless the message has already arrived at the node with

numerically closest nodeId.

3) Finally, the further key matching procedure is executed among several external index nodes, and the last locating results are sent back to the request node.

2.3 Node arrival

Pastry is completely decentralized, scalable, and self-organizing, and it automatically adapts to the arrival, departure and failure of nodes. When a new node arrives, it needs to initialize and then inform other nodes of its presence. In this section, we consider the most general case. We assume the new node knows initially about a nearby search node, each ordinary node maintains a share-list based on its shared resource.

When the node finds a search node, and asks it to route a special “join” message with a share-list. In response to receiving the “join” request, the search node classifies the received resource information, and updates it state tables. Furthermore, the “join” message with share-list is forwarded directly to a certain external index node. Then the external index node checks the information and updates its record table, and so on. Similarly, the internal index node also updates its record table. Up to now, a node has joins into Pastry system.

2.4 Node departure

In Pastry system, nodes may fail or depart without warning. In this section, we discuss how to handle such node departures. A node is considered failed when its immediate neighbors in the nodeId space can no longer communicate with the node. Here, the failure of nodes is divided into ordinary node failure and super node failure (including search node and index Node).

The failure of an ordinary node that appears in the routing table of search node is detected when search node attempts to contact the failed node and there is no response. Then, the search node contacts the internal index node in this group. While receiving the message, the index node updates its own record table accordingly to deleting a failed node. To see if the index node is still alive, the search node attempts to contact each one periodically. If an index node is not responding, the search node asks other spare index nodes, checks the distance of each of the newly discovered spare nodes, and updates its own neighborhood set accordingly.

3 Experimental Results

In this section, we present experimental results obtained with a prototype implementation of the improved Pastry. All experiments were performed on Pentium 4.0 computers, running Windows XP. The Pastry node software was implemented on JXTA platform. Any computer that is connected to the Internet and runs the Pastry node software can act as a Pastry node.

3.1 Average number of routing hops

The first experiment shows the number of routing hops as a function of the size of the Pastry network. We vary the number of Pastry nodes from 10 to 60 in a local network. In each trial, two Pastry nodes are selected at random and a message is routed between the pair using Pastry.

Figure 2 shows the average number of routing hops taken, as a function of the network size. “$\log(N)$” shows the value $\log_{2^b} N$ and is included for comparison. ($\log_{2^b} N$ is the expected maximum number of hops required to route in a network containing N nodes). The results show that the number of route hops scale with the size of the network as predicted and the improved Pastry is superior in average number of routing hops.

3.2 Average distance length

The second experiment evaluated the locality properties of Pastry routes. It compares the relative distance a message travels using Pastry, according to the proximity metric, with that of a routing scheme that maintains complete routing tables. The distance travelled is the sum of the distances between consecutive nodes encountered along the route in the emulated network. For the routing scheme, the distance travelled is simply the distance between the source and the destination node. The results are normalized to the distance travelled in the routing scheme. The goal of this experiment is to quantify the cost, in terms of distance travelled in the proximity space, of maintaining only small routing tables in Pastry.

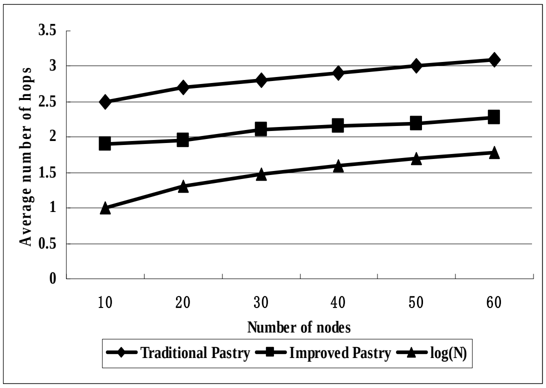


Figure 2 Average number of routing hops versus number of improved Pastry nodes

Figure 3 shows the results of average distance lengths. Because the improved Pastry design is a hierarchical overlay network, thus the average locating distance length of is often larger than traditional Pastry.

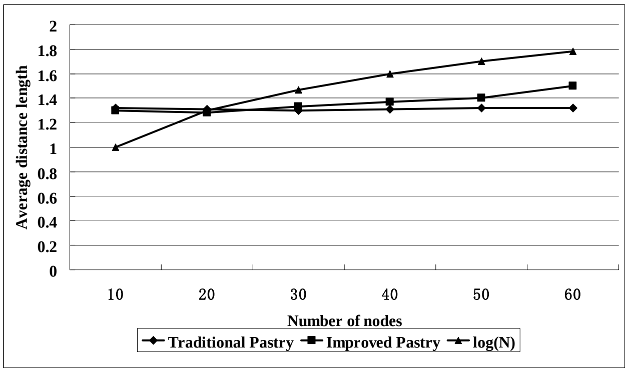


Figure 3 Average distance length of traditional Pastry versus improved Pastry

4 Related Work

P2P network is factually an overlay network for distributed object store, search and sharing. There are currently several peer-to-peer systems in use and many more are under development [8]. Among the most prominent are Napster [9], Gnutella [10]-[13], Freenet [14], Tapestry[15], Chord [16], CAN [17] and Pastry [1]. According to differences of object location, routing and P2P logical topology, the current P2P architecture can be classified to three types: Centralized-All, Decentralized but unstructured and Decentralized and structured. Most of the Decentralized and structured topologies are usually constructed using distributed hashing table (DHT) techniques, for example [1] and [15]-[17].

But the argument is that DHT-based systems do not capture the semantic object relationships between its name and its content or metadata [18]. When a node’s single hop crosses through different AS, it may lead to high routing delay. Though Pastry and Tapestry have did some work on this problem, it doesn’t solve it completely. Chord and CAN currently have no specific mechanism to heal partitioned rings, such rings could appear locally consistent to the stabilization procedure. In Chord, it only maintains a light overlay network protocol and hasn’t taken physical topology into account so that it may cause a long physical relay. Since Pastry and Tapestry record their physical neighbours in routing tables, the capability of system has improved greatly. However, compared to Chord, Pastry needs to keep more overlay network information. Moreover, it has no idea on the influence on realistic network caused by complex, dynamic and unsymmetrical position topology [19] and there is no research on this topic yet.

5 Conclusions

This paper presents an improved Pastry, a scalable, distributed object location and routing substrate for wide-area peer-to-peer applications. We also put forward the concept of index node and search node so as to address the limitation in traditional Pastry. Furthermore, a prototype is implemented on JXTA platform, and the elementary experimental results indicate that the improved Pastry network model is superior in routing performance. On the other hand, improved routing mechanism in Pastry has some shortcomings, such as how to incent nodes to share more resources, and so on. All of the questions need further study in the future.

References

[1] A. Rowstron, P. Druschel, “Pastry: scalable, decentralized

object location and routing for large-scale peer-to-peer systems[R]," IFIP/ACM International Conference on Distributed Systems Platforms, 2001

[2] A. Iamnitchi, M. Ripeanu, and InaForster, "Locating data in (Small-World) Peer-to-Peer scientific collaborations," 1st International Workshop on Peer-to-Peer Systems, IPTPS'02, Canmbridge, MA, March 2002

[3] C. Schmitz, "Self-organization of a small world by topic," 1st International Workshop on Peer-to-Peer Knowledge Management, Boston, MA, August 2004

[4] J. Kleinberg, "The small-world phenomenon and decentralized search," The Mathematics of Networks, 37(3), 2004

[5] Project JXTA. http://www.jxta.org/

[6] Kurose, K.W. Ross, Computer Networking: A Top-Down Approach Featuring the Internet (Third Edition), High Education Press, Beijing, 2005

[7] A. S. Tanenbaum, Computer Nerworks(Third Edition), Prentice Hall International, Inc.,Beijing,2002

[8] C. Wang, B.L, "Peer-to-Peer Overlay Networks: A Survey," 2003.http://citeseer.ist.psu.edu/706822.html

[9] Napster. http://www.napster.com/

[10] "The Gnutella protocol specification," 2000. http://dss.clip2.com/GnutellaProtocol04.pdf

[11] N. Kiran, S. Rollilns, and M. Khambatti, "From the editors: peer-to-peer community, looking beyond the legacy of Napster and Gnutella," Distributed Systems Online, IEEE, 7(3), March, 2006

[12] Yutang. Guo, Lv-Wanli, and Bin Luo, "Improved resource discovery algorithm on Gnutella based on P2P networks", Control Conference, CCC, Chinese 2007

[13] M.Ripeanu, "Peer-to-peer Architecture Case Study: Gnutella," In Proceedings of International Conference on P2P Computing, 2001

[14] I. Clarke, O. Sandberg, B. Wiley, and T. W. Hong, "Freenet: A distributed anonymous information storage and retrieval system," In Workshop on Design Issues in Anonymity and Unobservability, 2000, pp.311-320

[15] B. Y. Zhao, J. D. Kubiatowicz, and A. D. Joseph, "Tapestry: An infrastructure for faultresilient wide-area location and routing," Technical Report UCB//CSD-01-1141, U. C, Berkeley, April 2001

[16] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: A scalable peer-to-peer lookup service for Internet applications," In Proc. ACM SIGCOMM'01, SanDiego, CA, Aug. 2001

[17] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A scalable content-addressable network," In Proc. ACM SIGCOMM'01, San Diego, CA, Aug. 2001

[18] S. Ratnasamy, ScottShenker, "Routing algorithms for DHTs: some Open Questions," In IPTPS'02, January 2002

[19] Eng Keong Lua, Jon Crowcroft, Marcelo Pias, Ravi Sharma and Steven Lim, "A survey and comparison of peer-to-peer overlay network schemes," IEEE Communications Survey and Tutorial, March 2004

Research of Grid Portal Based on Web Framework

Lin Hu Qingping Guo

School of Computer Science and Technology, Wuhan University of Technology, Wuhan, Hubei, 430063, China
Email: suifeng115@163.com

Abstract

In this paper the problems of the developing effective methods for building the Grid Portal are discussed. Enlightened by the idea of building the Web Portal, this paper puts forward a model of the Grid Portal based on Struts Framework and has introduced Struts Bridge to solve those problems between the Struts Framework and the JSR-168 compliant Portlet. Compared with the normal web application based on the Struts Framework, the paper illustrates the ways of how to integrating Portlet with the Struts Framework and how to implement the model of Grid Portal based on Struts Framework in detail. The use of the Struts based on Struts Bridge in the process of building Grid Portal will enhance the extendibility, probability and reusability of the Portlet and shorten the period of the development.

Keywords: Grid Portal, Portlet, MVC, Struts Bridge, JSR-168

1 Introduction

The term “the Grid” was coined in the middle of 1990s to denote a proposed distributed computing infrastructure for advanced science and engineering [1]. Considerable progress has since been made on the construction of such an infrastructure, but the term “Grid” has also been conflated to embrace everything from advanced networking to artificial intelligence. As a new developing technology, the Grid is popular with many governments, research organization and the information industry community. Through the researchers’ endless explorations and efforts, the Grid has begun to march from the obscurely academic community to the business community and has really facilitated human being’s life. Nowadays, the research of Grid is attracting more and more people. However, no matter observing the Grid from the technologies the Grid used, the inside operating mechanism of the Grid or the implemented purpose of the Grid, the Grid is so extraordinarily complicated that users can not make good use of the gigantic shared resources in the Grid system. Therefore, the Grid Portal appearing is the result of the Grid development.

2 The Development of Grid Portal

By shielding the complex internal details of the resources in Grid system for the users and enabling the users to use the grid system and gain the grid services through the friendly Web interface, the consistent operating mode and the effectively and conveniently accessing mechanism, the Grid Portal has solved the difficult problems about how to share resources conveniently and effectively in Grid system [2]. Therefore, the researches about Grid Portal and its related technologies are practically and meaningfully.

At the beginning of the Grid technology development, the Grid Portal [3, 4, 5] is always built on some special Grid middleware, such as Globus which provides services by the Globus toolkits(GT2) [6] programmed by C. However, this kind of Grid portal has many limitations, such as lacking customization, limited Grid services and static Grid services [7]. In order to overcome the Grid Portal’s limitations, the modern Grid Portal is built with Portlet [8]. Using the interfaces Java Portlet API provided, Grid application researchers can implement the concrete application logic according to the Grid services interfaces that OGSA has defined and add the Portlet into the Grid Portal after packaging every Grid application into a Portlet. For the users, they can obtain the Grid application they are interested in by adding or deleting Portlet directly.

At present, the Grid Portal, to a certain extent, is valuable and it is easy for user to access, but many problems exist in the following aspect: the performance ability, the extendibility and the reusability.

3 MVC-Based Gird Portal Model

The Portlet technology brings a bright future for the Grid development. To make good use of Portlet to build Grid Portal, and ensuring the performance ability, the extendibility and the reusability of the Portlet components, we can learn something from the idea of Web Portal. It’s a good way to introduce MVC pattern to build the Grid Portal.

3.1 MVC pattern

The MVC Pattern [9, 10] has its roots in SMALLTALK. It separates core business model functionality from the presentation and control logic. Such separation allows multiple views to share the same data model, which makes supporting multiple clients easier to implement, test, and maintain. MVC model breaks up an application into three parts: model, view and controller.

1) The model represents data and the business rules that govern access to and update the data. Often the model serves as a software approximation to a real world process, so simple real world modeling techniques apply when defining the model.

2) The view renders the contents of a model. It can access data through the model and specify how the data should be presented. It is the view's responsibility to maintain consistency in its presentation when the model changes. It’s the interface that where the user interacts with the system.

3) The controller is the means by which the view interacts with the model. The controller receives the inputs from the user and instructs the model to perform actions based on the corresponding input.

The following Figure 1 represents the MVC Pattern and the relationship among them (the dashed represents Events and the solid represents Method Invocations).

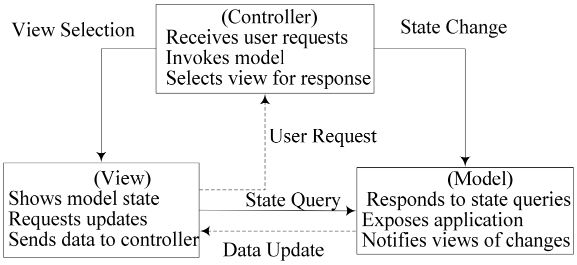


Figure 1 MVC architecture

3.2 MVC-based gird portal model

When building Grid Portal using MVC pattern, the model of development transferred from the model based on Java Portlet API to the model composed of three parts which are the view layer, the model layer and the controller layer. Using the JSP technology, the view layer provides users with abundance interfaces and expands the performance ability of the program. The model layer mainly defines the data structure and provides the Grid services, namely takes the responsibility of communicating with the Grid middleware. In the layer of controller, Servlet and the module which has implemented the Portlet interfaces dominate the interactions between the users and the Grid services through the OGSA client proxy or the standard communication protocol (such as SOAP and Java RPC). In this model, the process of developing Portlet is similar to Servlet. Playing controller’s role, the Portlet receives the requests from the view layer, and then through the event processing methods in the Portlet, it invokes model, updates the view and responses user’s requests. The following Figure 2 represents Gird Portal model based on MVC pattern.

The Grid Portal built with MVC pattern can solve the existing problems in building Grid Portal. The advantages of this model mainly show as followings:

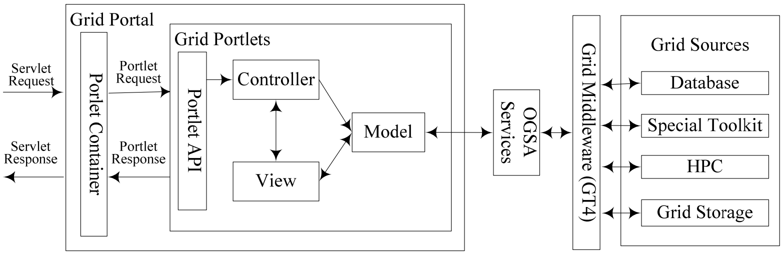


Figure 2 Gird Portal model based on MVC pattern

1) Separating the program into three layers, the model will ease the coupling of the whole program and enhance the extendibility of the program;

2) The separation of the view, the model and the controller will enhance the reusability of the modules;

3) The responsibilities of each layer are very clear, so it is facilitating for researchers to develop;

It's so easy for testers to understand the structure of the program that they can test the program conveniently and effectively.

4 Application of Struts Framework in Grid Portal Building

As one of the MVC Frameworks based on Sun J2EE [10, 11], Struts is mainly realized by Servlet and JSP. Because of its' advantages such as fully satisfying the application and development demands, easy to use and deploying effectively, Struts is popular with many company and organization in the past several years. The Struts compacts the four parts (Servlet, JSP, the labels defined by the developers and the message resources) into a uniform framework. By the Struts, the developers can directly extend the framework instead of implementing the MVC pattern by themselves, so it will save the time. If we can utilize the Struts in the Grid Portal's construction, it will shorten the period of the development and will enhance the component's extendibility, probability and reusability.

4.1 Integrating the struts framework with the portlet

The Struts Framework is designed according to the MVC pattern, but it does not support the JSR-168 standard, so it is incompatible with Portlet based on JSR-168 standard. For example, the controller of the Portlet returns fragments that can assemble into JSP pages instead of JSP pages directly. To solve the problem, a good way is to introduce the Struts Bridge [12] composed of some java interfaces. The Struts Bridge allows Struts and Struts Tiles applications to be run as JSR-168 compliant Portlet. Existing or new Struts Applications can be easily deployed as Portlet Application or Web Application. The Bridge wraps and enhances the native Struts Framework to overcome its limitations and incompatibilities with the Java Portlet Standard requirements. The Struts Bridge is developed to be independent from specific Portals and uses only a very small interface to the Portal to be able to get access to the Servlet environment at runtime. As all JSR 168 Portlet Containers are required to build upon the Servlet specification, providing this interface for a specific Portal is usually very easy to do, if not done already. To overcome the incompliant between Struts and JSR-168, the Struts Bridge provides the following improvements on Struts for developing and running Struts Applications as Portlets:

1) Providing a "virtual" page context and a "normal" page URL to the Struts Application;

2) Allowing Servlet invocation from an AcitonRequest using an separate lightweight interface to the Portal;

3) Providing extension points if a more complex interface to the Portal is required;

4) Separating Struts Action Processing and View Rendering when accessed from an ActionRequest automatically;

5) Transparent communication of request attributes between ActionRequest and RenderRequests;

6) Transparent translation of "normal" Web application resource and action link urls to valid Portlet URLs.

4.2 Building grid portal on struts framework

Building Grid Portal on Struts Framework, the most difficult thing we will meet is how to integrate the Struts with Portlet. After expanded by the Struts Bridge, a Struts application can work as a JSR168 compliant Portlet. Compared with the normal application program based on Struts, the model of Struts based on Struts Bridge is showed by the following Figure 3.

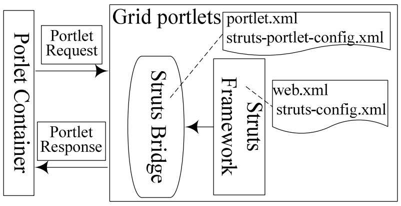


Figure 3 Struts based on Struts Bridge

Now, we can build the Grid Portal according to the mode of Struts based on Struts Bridge. The Grid Portal model based on Struts and Struts Bridge is showed by the following Figure 4.

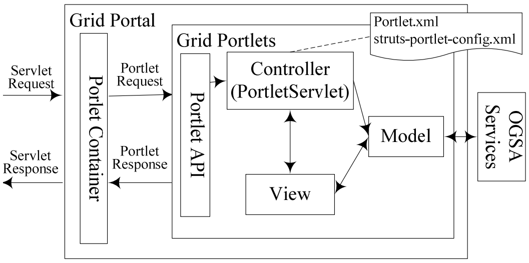


Figure 4 Grid Portal based on Struts and Struts Bridge

The following section will discuss how to build the Grid Portal based on the Struts Framework by researching the way that Struts1.2 integrates with Portlet based on JSR 168 standard.

(1) Modifying the related configuration files

When we do some normal projects based on Struts Framework, there are tow important configuration files: web.xml and struts-conFigure xml. To converting a Struts application to a Portlet, the first step is to modify web.xml to use PortletServlet instead of ActionServlet. We can make use of the package of org.apache.portals.bridges.struts.PortletServlet included in Struts Bridge to modify the configuration files. In struts-conFigure xml a new Portlet aware controller should be introduced. We can add a <controller> element just above the <message-resources> element to override the default RequestProcessor using PortletRequestProcessor which has also been included in Struts Bridge.

(2) Creating the related configuration files

A new configuration file called struts-portlet-conFigure xml should be created. The purpose of this file is to help the Struts Portlet identify which URLs are portal actions and which are portal views, it also identifies which objects should be copied from a portal action to the subsequent portal rendering request. The file defines an attribute element named portlet-url-type and its three child element, namely action, render and resource. In order to configuring the Portlet URL, the file also appoints the attribute of the path. To deploy all Portlets, another configuration file named portlet.xml is also be added to the corresponding project. This file includes one or more Portlets definition and the information about Portlet itself, such as the name of the Portlet, the initial parameter and the Portlet class.

(3) Create an instance of ServletContainerProvider for the portal platform

To get the struts bridge working, modifying and creating the above configuration files are not enough, we must implement the interface named org.apache.portals.bridges.common.ServletContextProvider. The Struts Bridge has contained the package, the only thing we should is to implement it. Of course, before we do these, we should configure portlet.xml like this:

```
<init-param>
<name>ServletContextProvider</name>
<value>ca.mun.portal.bridges.PortalServletContext
Provicer</value>
</init-param>
```

After implementing the above steps, the model of Grid Portal based on Struts has been created. The researchers can make use of this model to build the needed Grid Portal according to the specific project.

5 Conclusions

This paper puts forward a model of Grid Portal based on the Struts Framework, and has introduced the way of how to implement the model. Through utilizing the Struts based on the Struts Bridge to build the Grid Portal, the advantages of the Struts will enhance the extendibility, probability and reusability of the Portlet components. The researchers can make good use of Portlet integrated with the Struts to build the Grid Portal they want easily and effectively.

References

[1] Ian Forster, Carl Kesselman, Steven Tuecke, “The Anatomy of the Grid”, *Journal of Supercomputer Applications*, 15(3), 2001, pp.1-6

[2] Yong Yin, Zude Zhou, Yihong Long, “Research on Grid Portal in Manufacturing Grid Environment”, DCABES 2007 PROCEEDINGS. WuHan, 2007, pp. 677-680

[3] Xiaobo Yang, Xiao Dong, Robert Allan. “Development of standards-based grid portals Part 1-3”, http://www-128. ibm.com/developerworks/grid/library/gr- stdsportal1/, 2007.4

[4] Yossathorn Phumisuth, Tiranee Achalakul, “Grid Portal Design and Usability Evaluation”, *Communications and Information Technologies, ISCIT’06*, 2006, pp.189-193

[5] D.Gannon, J.Alameda, O.Chipara, Building Grid Portal Applications from a Web Service Component, Proceeding of the *IEEE*, 93(3), 2005, pp.551-562

[6] Globus Projects, http://www.globus.org/toolkit/

[7] Zheng Feng, Shoubao Yang, Shanjiu Long, “Research on Integrating Service in Grid Portal”, *Computer Science*, 2004, pp.821-824

[8] Xin Chen, Shoubao Yang, Xiaochun Zhao, “Portlet Research in Grid Environment”, *Journal of Software*, 14(6), 2003, pp.1148-1151

[9] Weiqin Sun, Master Struts, BeiJing: Publishing House of Electronics Industry, 2006, pp.1-190

[10] Lei Ji, Li Li, Wei Zhou, Master J2EE, BeiJing: POSTS&TELECOM PRESS, 2006, pp.1-91

[11] John T.Bell, James T.Lambros, J2EE Open Source Toolkit, Wiley: Wiley Publishing, 2005, pp.12-35

[12] Struts Bridge, http://www.ja-sig.org/wiki/display/PLT/ Struts+ Bridge

Lin Hu is a graduate student in school of computer science & technology, Wuhan University of Technology. He got his bachelor's degree in Gannan normal’s college in 2006. His main research interests are in distributed parallel processing and grid computing.

Qingping Guo is a Full Professor and a head of Parallel Processing Lab, dean of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology. He graduated from Wuhan University in 1968; from Huazhong University of Science and Technology in 1981 with specialty of wireless technology. He is a holder of K. C. Wong Award of UK Royal Society (1994); was a visiting scholar of City University and University of West Minster (1986~1988), Visiting Professor of the UK Royal Society (1994), Visiting Professor of Queen Mary and Westfield College, London University (1997~2000), Visiting Professor of National University of Singapore (2000), Visiting Professor of University Greenwich (2003). He is one of the DCABES international conference founder, was the chairman of DCABES 2001, 2004, and 2007, co-chair of DCABES 2002, and will be the co-chair of DCABES 2008. He has published two books, over 80 Journal papers, edited four DCABES Proceedings. His research interests are in distributed parallel processing, grid computing, network security and e- commence.

Parallel Data Mining Technology under Grid Environment

Xiufeng Jiang[1]   Meiqing Wang[2]

College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian 350002
Email:1 jxf1963@21cn.com; 2 mqwang@fzu.edu.cn

Abstract

The traditional data mining is unable to solve the problem of distributional magnanimous data mining and the distributional system is very difficult to solve the question of heterogeneous operating systems and proposals. Under the grid environment, parallel data mining can solve heterogeneous, distributed and magnanimous data mining problem. In this paper the existing strategy and the association rule algorithm of parallel data mining is analyzed and a new association rule parallel mining plan is proposed based on the grid environment. The new strategy is based on the agent technology, multi-thread and the centralism vote technology. The experimental results in a simulated grid platform show the feasibility and the effectiveness of the new strategy.

Keywords: Data mining, Grid, parallel, RMI, Agent

1 Introduction

Traditional data mining algorithm[1] is often based on a single database, and it was performed out on a single computer. The situation of the actual application is the magnanimous databases were distributed in many different places. To analyze and mine these magnanimous data which were distributed widely, it requests that the data mining system should have better characteristics of distribution and extendibility, which is the shortcoming of traditional data mining pattern and technology. On the other hand, some new mining strategy and algorithm meet the huge challenge of the system's computation ability and the central data mining pattern which make it necessary to use the distributional computing technology. In this paper, a distributional data mining system is proposed to parallel process subtasks for each distributed database in each computing nodes and synthesizes each partial processing result to form the final output in the central node.

But the existed distributed computing system is very difficult to solve the heterogeneity on operating systems and protocols. In this paper, the solution of these problems is based on Globus Toolkit. Globus Toolkit [2][3] is a Grid middleware, which is equal to grid operation system[4], can hide the network heterogeneity and is convenient to parallel processing tasks among heterogeneous machines. In this way, the existed network resources can be fully used to execute the distribution parallel computation.

2 Parallel Data Mining Strategy Choice

Parallel computing strategy based on the grid can be classified as tight coupling parallel program and the loose coupling[4]. The tight coupling parallel program requests the close communication between each parallel node. Loose coupling parallel service is suitable for the parallel application which a task may be divided into independent sub-tasks, and each sub_task can be defined as a grid service to be distributed in each node. The grid client can call the grid service by starting multi_threads to accomplish the parallel computing application.

Based on Globus, MPICH-G2 is the parallel processing MPI grid edition. It has provided completely consistent parallel programming interface with MPI in heterogeneous grid environment. But in Globus version 4.0, the close coupling MPICH-G2 is not supported.

*This work was partially supported by Provincial Natural Science Funding, Fujian, China, under Grant No. A0510005

And the loose coupling parallel service strategy can not be used completely, because parallel data mining service needs communication repeatedly due to the dependency among the sub-tasks.

In this paper parallelization based on agent [5] is proposed under the grid environment. An agent is a program using for performing query for frequent item set. Using the agent, the data mining service is deployed to the Globus container, and the database in each node is issued as data service through OGSA-DAI[6] which can be inquired through the grid portal. The node which has a database is used as a server of RMI service, and the node which provides data mining service is used as the central node which is also the consumer of RMI service. The consumer calls RMI service by using the agent as a parameter in order to let this agent running in the RMI server node.

3 Parallel Association Rule Data Mining Algorithm

The existing distributional association rule algorithm mostly is based on Apriori algorithm [7][8]. Based on the Apriori algorithm, Agrawal et al. have designed three kinds of distributional association rule algorithms: the counting distribution, the data distribution and the candidate distribution. In the counting distributed algorithm, each node includes the complete frequent item set, reduced system's communication and the synchronized expenses. The drawback is that the redundancy saving of the frequent item in various node decreases the utilization ratio of resources. In the data distribution algorithm, the entire frequent item set disperses evenly in each node; the utilization ratio of resources is high, but communication and the synchronized expenses increase between the node. In the candidate distributed algorithm, the algorithm will distribute the entire frequent item set to each node according to the semantic information, thus load in various nodes is achieved balance and synchronized expenses between the node is decreased. Another improving method based on the Apriori algorithm is FDM (fast distributed association rules mining) algorithm proposed by Cheung et al. The scale of candidate frequent item set produced by FDM may be smaller than the counting distribution algorithm, and communication expenses between the nodes also will be reduced.

When seeking for the whole frequent item set, the above algorithms are all based on using the distributional voting plan, so their communication is based on point to point pattern. Although each node's communication load is quite balanced, but each node's communication load is big which causes the entire Internet communication expenses to be big.

To further reduce the communication load between the nodes, the central vote plan is used in this paper. The central node directly communicates with each node to obtain the partial frequent item set, then unite the result to generate whole frequent item set. We also implement the parallel association rule algorithm based on the improvement of Apriori algorithm. In the new algorithm each node has the subset of whole database evenly, and the database is deployed as a service in GT4 service registration by OGSA-DAI. Figure 1 explained the process of parallel data mining under the OGSA architecture.

(1) Users submit the task of data mining through grid portal[8]. The function of the grid portal is to submit jobs, look up existed services, and collect the information about the situation of hardware, software, the network resource, and running jobs. The user first look up the text association rule service and data service through the grid portal, and then submit the job of data mining. When running job submission service, the grid portal requires the user to choose the mining service and data service, the minimum support and confidence.

(2) The job submission service searches for the provider of text association rule service in the service registration center, then create the instance of the service in the node which provide the service.

(3) As controller which vote centralism, the service provider, create multi threads to parallel perform the mining task. The function *run()* in each thread searches the server instance in the RMI registry, call its *receive()* function, where the instance of Agent acted as a parameter will be transported to each node. The function of Agent is to seek

the frequent one item set. In each node, Agent scan local partial database, generate partial frequent 1 item set . $L_{i,1}$

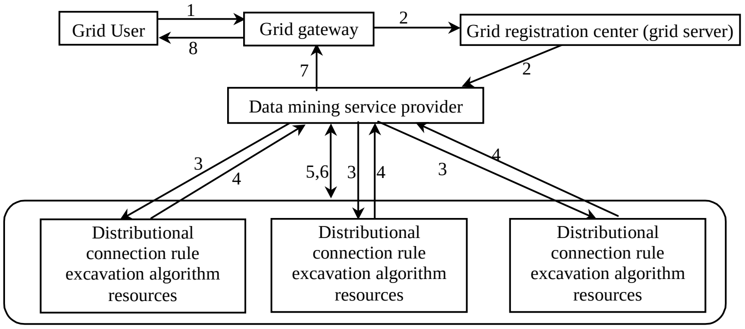

Grid User
1
8
Grid gateway
2
Grid registration center (grid server)
2
7
Data mining service provider
3
4
5,6
3
4
3
4
Distributional connection rule excavation algorithm resources
Distributional connection rule excavation algorithm resources
Distributional connection rule excavation algorithm resources

Figure 1 The process of parallel data mining under the OGSA architecture

(4) The central node by using the *join()* method of thread waits for each partial node computation finished, then call the partial node's *getData()* method to obtain frequent 1 item set, and finally produce overall frequent 1 item set L_1 according to the minimum support.

(5) The central node starts multi-threads once more, and the function *run()* of each thread searches for the server instance in the RMI registry, call its *receive()* function, where the instance of Agent acted as a parameter will be transported to each node. The function of Agent is to seek the frequent *k* item set. In each node, Agent take $k-1$ frequent item set from central node, generate local $k-1$ frequent item set

$L_{i,k-1}$, generate *k* candidate item set $C_{i,k}$ by using connection step $L_{i,k-1}\infty L_{i,k-1}$, and scan local partial database to generate *k* frequent item set.

(6) The central node by using the *join()* method of thread waits for each partial node computation finished, then calls the partial node's *getData()* method to obtained frequent k item set, and finally produces overall frequent k item set L_k according to the minimum support.

The algorithm repeats step (5) and (6) until some value *r* causes L_r spatial. The algorithm stopped and the central node produces frequent item set *L*.

(7) The central node generates an association rule according to minimum confidence, and returns the result of the association rule to the job submission service.

(8) The job submission service returns the result to the client.

4 Parallel Connection Rule Service Realization and Service Issue

(1) The interface of text association rule mining service is defined by using the WSDL and its service[10] name is called *DataMining*. The service provides two functions, one is *ArrayList* aprior (AaaryList servers, ArrayList databases, double minSup), the other is *generateRules* (ArrayList d, float minCon). The aprior function is used to generate frequent item set which satisfies minimum support and accepts three parameters: instances of the server, data services, minimum support. The *generateRules* function is used to generate association rule which satisfies minimum confidence.

(2) The class of this service implementation is *DataMiningServiceImpl,* which must realize the interface of Resource and ResourceProperties. This class has four resources attributes: minimum support, minimum confidence, list of data services and list of server instance. Each resources attribute should have *get/set* methods. Finally, the methods of aprior and generateRules should be implemented.

(3) In order to realize the frequent item set parallel processing, the program had to be transferred to the remote host where the database runs. We use the agent-based RMI to achieve this goal. The method of *aprior()* act as the control node. The concrete realization is as follows:

①Defines the Agent interface, the interface functions is *execute ()*.

②Agent1 is defined to realize Agent interface. Agent1's constructor accepts the data services name and the minimum support. Its method of *execute()* generates 1 frequent item for the specified data service and minimal support.

③Agent2 is defined to realize Agent interface. Agent2's constructor accepts the data services name, $k-1$ frequent item set and the minimum support. Its method of *execute*() generates *k* frequent item for the specified data service, $k-1$ frequent item set and minimal support.

④The interface of Server class is defined as *ServerInterface*. The function of the interface is receive (Agent a) and *getData()*. The function of receive (Agent a) is to execute the program Agent, and the function of *getData()* return data to central node.

⑤The class of Server is defined to implement the interface of *ServerInterface*. The instance of this class is generated and registered to the RMI registry in *main()* function.

⑥The class of *NodeThr* is defined to extend thread. Its constructor accept Agent interface and the name of Server instance as parameters. According to the name of Server instance, the function of *run()* searches in the RMI registry, obtains the instance handle of Server on the remote host, and call receive(Agent) by this handle.

(4) The Realization of apriori as follows:

①Traversing set of Server instance name and data service name, instance of Agent1 will be generated according to the name of data service and minimum support, instance of NodeThr will be generated according to the name of Server instance and Agent1 instance in Cycle, and start each thread instance to run.

The execution of each thread instance *run()* function has been separated from the host thread which lets the main loop be able to start next node instance rapidly, achieved multiple nodes to find the frequent item set at the same time.

②To know all the *execute ()* method on nodes is finished, can cycle through executing *node.join ()* method on each node.

③Collecting all the frequent set on various nodes to vote. In order to avoid more complex communication between threads and conflict while Task of computing nodes threads in active transmission results. We gather data from computing nodes in central node by call *getData()* function of Server class , instead of transmitted the results to the central node in the sub-task.

④Instance of Agent2 will be generated according to the name of data service, $k-1$ frequent item set and minimum support, and instance of *NodeThr* will be generated according to the name of Server instance and Agent2 instance in Cycle , and start each thread instance to run , to produces *k* frequent itemset parallelly.

⑤To know all the *execute ()* method on nodes is finished or not, the method *node.join ()* should be executed repeatedly on each node.

⑥Collect the *k* frequent item set of various nodes to vote. Generated full k frequent item set L_k .

⑦Repeat ④~⑥ until the return of *k* frequent item set is empty .

(5) Job submission service calls text association rule service. we create a *EndpointReferenceType* object to represent *EndpointReference* of this service, Then, we have invoked the service *portType*, and call methods of the service by this *EndpointReferenceType* object.

(6) Use WSDD and JNDI to define release parameters, and use Ant to compile and generate file of services gar, finally we deploy Service to GT4 containers.

5 Test Environments and the Result Analysis

(1) Grid platform simulation

A grid environment is simulated for data mining services. The experimental hardware configuration is shown in Table 1 and the software configuration in Table 2.

Table 1 Experimental Hardware Configuration

Host name	IP address	OS	Machine Configuration
LiJie	210.34.55.181	RedHat Linux 9.0	P4 2.93G/1GB/120G
Node1	210.34.55.222	RedHat Linux 9.0	P4 1.7G/512MB/80G
Node2	210.34.55.175	RedHat Linux 9.0	P4 1.7G/512MB/80G
Node3	210.34.55.161	Windows XP2	P4 2.93G/1GB/120G

Table 2 Experimental Software Configuration

Software Type	Software name and version
Grid container	Globus Toolkit 4.0.2
Database	MySQL 5.0, PostgreSQL 8.0.4, RRD 1.2.15
Programming tool	JDK1.5.0_07
Information surveillance Software	Ganglia 3.0.3
Web container	Tomcat 5.5.17
Construction, deployment tool	Ant 1.6.5
Data access and integration tool	ogsadai-wsrf-2.2

(2) The experimental results and analysis

In this environment, we execute the Data Association Rules Mining and have corresponding tests, Table 3 is the test results of parallel association rule mining exection time. Table 4 is test results of the parallel data mining speedup model. The results show that on the distribution grid computing environment, the running time of large-scale parallel computing is greatly smaller than the single one. So, using the distribution parallel computation strategy which we proposed to realize parallel computation is feasible.

Table 3 Parallel Association Rule Mining Exection Time

Node Number Time (ms) Transact	N=1	N=2	N=4	N=6	N=8
400	968	1424	864	701	712
800	2452	2851	1657	1277	1179
1600	5060	4961	2811	2162	1917
3200	16323	13603	7420	5232	4251

Table 4 Parallel Association Rule Mining Speedup

Node Number Speedup ratio Transact	N=2	N=4	N=6	N=8
400	0.68	1.12	1.38	1.36
800	0.86	1.48	1.92	2.08
1600	1.02	1.80	2.34	2.64
3200	1.20	2.20	3.12	3.84

6 Conclusion

This paper presents a parallel data mining solution under grid environment GT4. Data mining program will be represented as grid services, and multi-thread means are used to perform the parallel data mining, agent technology is used to implement the operation of program transportation from the control nodes to the local data node, and the centralized vote on the data mining results. This strategy reduces the system's communication time, avoids shortcoming which GT4 not being able to use MPICH-G2 to realize the parallelism, also avoids that loose coupling parallel services is difficult to use in this situation that a sub-task is an independent, request repeated communication. Compared to the distributional vote plan, the centralism voting reduces the system communication time. The experimental results show that the strategy is feasible for parallel data mining.

References

[1] Jiawei Han, Micheline Hamber. Data Mining Concepts and Techniques. Beijing: China Machine Press, 2001

[2] globus Project, http://www.globus.org

[3] A Globus Toolkit Primer, http://www.globus.org/toolkit/docs/

[4] LIN Weiwei, ZHANG Zhili, QI Deyu. Strategies for distributed parallel computing on Grid computing environments, Computer Engineering, 2006.5

[5] M.L.Liu. Distributed Computing: Principles and Applications. Tsinghua University Press, Beijing, March 2004, page: 310-315

[6] OGSA-DAI. http://www.ogsadai.org.uk/

[7] Su Yuanyuan, Parallel Algorithm for Mining Association Rules

[8] Peter Brezany, Jürgen Hofer, A Min Tjoa, Alexander Wöhrer. GridMiner: An Infrastructure for Data Mining on Computatutional Grids, 2004

[9] Jason Novotny, Michael Russell and Oliver wehrens. GridSphere: A Portal Framework

[10] Borja Sotomayor. The Globus Toolkit 4 Programmer's Tutorial, http://docs.huihoo.com/globus/gt4-tutorial/

Numerical Based on Genetic Algorithm and Application of Inverse Problem

Yamian Peng Xiujuan Xu Nan Ji Aimin Yang Junhong Ma

College of Science, Hebei Polytechnic University ,Tangshan, Hebei, 063000 China

Email: pengyamian@heut.edu.cn

Abstract

A new technique is based on Genetic Algorithm (GA) to solve inverse problem of partial differential equation parameter identified is given in the text. The original iterated values are obtained through across gene and variance gene of GA, and put the original iterated values in the beginning of the best disturbed iteration method for the inverse problem of partial differential equation parameter identified, then we can get the steady numerical solution of the parameter that need seek for. The results shows that the approximate solution have good astringency and high precision, and can apply broadly.

Keywords: Genetic Algorithm (GA); inverse problem; partial differential equation; the best disturbed iteration method

1 Introduction

In the process of solving inverse problem of partial differential equation by the best disturbed iteration method,at the first need to ascertain the guessing original values for the iteration, but the un-suitable guessing original values could made a big infection to the precision of iterate solution, so how to choose appropriate original values is the key condition for the problem solved. In recent years, A has broad appliance in every subject, the feasibility of GA solving Inverse problem will be discussed in the text and discuss a new technique for solving inverse problem: the best disturbed iteration method based on GA.

In the filed of environmental hydraulics, control equations of many systems can described by the ODE or PDE with fitting initialization conditions and boundary conditions. They formed posed problem in math, and can get the rules of distributing in space and evolution in time of dependent variable in the process of solve it. Realization for forecast of dependent variable was positive problem. Supposed region Ω encircle with boundary Γ, and then the classical positive problem can express:

$$\begin{cases} L(\Phi)=0 \quad in \;\; \Omega \\ B(\Phi)=0 \quad on \;\; \Phi \\ I(\Phi)=0 \quad in \;\; \Omega \;\; and \;\; on \;\; \Phi \;\; when \;\; t=0 \end{cases}$$

Φ is dependent variable of speed of water or tension etc, t is time，L，B，I are control equation and differential operator and initialization condition operator and boundary condition operator. The target of posed problem is for $\Phi=\Phi(x,t)$ fill forward formula and x is vector of space.

With the development of economy, the environmental problems increasing more and more such as: control liquid waste let problem, control heat waste let problem, optimization problem of liquid waste let system design, inverse problem of filter coefficient of groundwater, etc. It is difficult to solve these problems used forecast method, and then it must put inverse theory into the research. Show in mathematics, it is settled the positive problem certain factor that must known change into variable of unknown solution for solve and the part of unknown solution $\Phi(x,t)$ (such as distributing in the sub domains of Γ or Ω) as condition must to satisfy and already known. In this way, portion of positive problem solution become condition already known and the factor of positive problem must known become the solution to solve. The realization of control of the system through the assurance to carry on an estimate to become to regulate system of some factor cause a vegetable to attain a purpose of resolve the actual problem .

2 Get the Original Numbers

Consider the hyperbolic equation inverse problem of parameter identified as follows:

The initial boundary problem of hyperbolic partial differential equation in the rectangle region

$$R=\{(x,t):0\le x\le a,0\le t\le b\}$$

$$p(x)\frac{\partial^2 h}{\partial t^2}=\frac{\partial}{\partial x}(q(x)\frac{\partial h}{\partial x})+c(x,t)\frac{\partial h}{\partial x}+d(x,t)h+f(x,t)$$

(I)

$$h(\alpha,t)=g_1(t),h(\beta,t)=g_2(t);\ 0\leqslant x\leqslant a,0\leqslant t\leqslant b$$

$$h(x,0)=h_0(x);\ h_t(x,0)=h_1(x);$$

Functions $\mathrm{p(x),q(x),c(x,t),d(x,t),f(x,t),}$ $\mathrm{g_1(t),}$ $\mathrm{g_2(t),}$ $\mathrm{h_0(x),h_1(x)}$ are known. The inverse problem is confirmed $q(x)$ by system of equations (I) and other add-information. It often translate into the nonlinear optimize problem depend on observation values of $\mathrm{h=h(x,t)}$.A few values instances bases different observe modes and different optimize indexes adoptively in practice as follows:

（1）Numbers of one place but different times

$$\mathrm{h(x_0,t_i)}\text{，}\mathrm{(i=1,2\cdots n)}$$

are observed.

（2）Numbers of one time but different places

$$\mathrm{h(x_i,t_0)}\text{，}\mathrm{(i=1,2\cdots n)}$$

are observed.

（3）Numbers of different times and different places

$$\mathrm{h(x_i,t_j)}\text{，}(i=1,2\cdots n, j=1,2\cdots m)$$

are observed.

If $\mathrm{p(x),q(x),c(x,t),d(x,t),f(x,t),}$ $\mathrm{g_1(t),g_2(t),}$ $\mathrm{h_0(x),h_1(x)}$ are fixedness, $\mathrm{h(x,t)}$ is depend on $\mathrm{q=q(x)}$ in system of equations (I).So solution of (I) can express as $\mathrm{h=h(q(x),x,t)}$ and translate problems of identify parameter $\mathrm{q=q(x)}$ into optimize problems. According to the situation for values of different times and different places

$$\mathrm{h(x_i,t_j)}\text{，}(i=1,2\cdots n, j=1,2\cdots m)$$

Basis the differ measure of c(x,t), the identify problem of k(x) can transform the optimization problem as following:

（4）$\min\limits_{k\in M}\sum\limits_{i=1}^{n}\left[h(q(x),x_i,t_0)-h(x_i,t_0)\right]^2$

Get data $\mathrm{h}(x_i,t_0),i=1,2\cdots n$.

（5）$\min\limits_{K\in M}\sum\limits_{i=1}^{n}\left[h\left(q\left(x\right),x_0,t_i\right)-h\left(x_0,t_i\right)\right]^2, i=1,2,\ ......n$

Get data s $\mathrm{h}(x_0,t_i)\ i=1,2\cdots n$,

（6）$\min\limits_{K\in M}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}\left[h\left(q\left(x\right),x_i,t_j\right)-h\left(x_i,t_j\right)\right]^2,$

$i=1,2,......n,\ j=1,2,......,m$

Get data $\mathrm{h}(x_i,t_j),\left(i=1,2\cdots n, j=1,2\cdots m\right)$.

If take （4）as the adapted function in GA, it can translate the parameter identify problem into the nonlinear functional optimize problem and can get the best approximate solution of the unknown parameter $\mathrm{q=q(x)}$.

The system of primary functions

$$\Phi(x)=(\varphi_1(x),\varphi_2(x),......,\varphi_n(x))^T$$

is used and take the combination

$$\mathrm{c^*(x)}=\sum_{\mathrm{i=1}}^{\infty}\mathrm{k_i^*}\varphi_{\mathrm{i}}\mathrm{(x)}$$

as the simulated solution of unknown parameter $\mathrm{q=q(x)}$,then through iteration method to seek for coefficients $k^T=(k_1,k_2,...,k_n)\in R^n$ in the combination and hope to get the most approximate solution that most nearly the precise solution of $\mathrm{q=q(x)}$.So the question is how to give appropriate original coefficients of $k^T=(k_1,k_2,...,k_n)\in R^n$ to get the best approximate solution of $q = q(x)$. The strategy is through the GA to get appropriate original coefficients of $k^T = (k_1, k_2, ..., k_n) \in R^n$.

3 Genetic Algorithms

The parameter identify of the convection-diffusion equation of the environment hydraulics can be the example explains the application of Genetic Algorithms.

Consider the initial boundary problem of convection-diffusion equation:

$$(II)\begin{cases} \frac{\partial h}{\partial t} - \frac{\partial}{\partial x}\left(q(x)\frac{\partial h}{\partial x}\right) + a(x,t)\frac{\partial h}{\partial x} \\ \qquad = f(x,t) \quad 0 < x < 1, t > 0 \\ h(x,0) = h_0(x) \quad 0 \le x \le 1 \\ h(0,t) = g_0(t), h(1,t) = g_1(t) \end{cases}$$

The equation with important application background and many actual problems in the environment hydraulics was classified this equation. When k、a、f、c0、g0、g1 were known and satisfied certain regularization and consistent, the above problem was posed. Now the question is how to identify these unknown parameters according the equations (II) and other known information if $q(x)$ is unknown. This make up a kind of parameter identify inverse problem that important in study water pollution model. The identification of parameter plays an important rule of model accuracy improvement. For recognizing these parameters, it is often convert it an optimization problem through the measure data of $q = q(x,t)$.

If fixed a，f，c0，g0，g1，when k=q(x) in equations (II) get differ values, the respond solution be differ too. For express this relationship, sign the solution of equations (II) q=q(k(x),x,t). Basis the differ measure of q(x,t), the identify problem of k(x) can transform the optimization problem as following:

$$\min_{K \in M} \sum_{i=1}^{n} \left[h\left(q(x), x_i, t_j\right) - h\left(x_i, t_j\right) \right]^2,$$

$$i = 1, 2, ...n, j = 1, 2, ..., m$$

When get data of difference time and difference points $h(x_i, t_j)(i = 1, 2 \cdots n, j = 1, 2 \cdots m)$.

The material steps as following:

1 The adoption real amount codes strategy, each individual contains parameter for estimate and its number according to real problem. It takes three ones in this text and 20～50 individuals in a group;

2 Take the excellent select policy that is the previous generation individual substitute the present generation individual when the previous generation individual is better;

3 Genetic operator include cross and variation only;

4 The control parameter of GA in the numerical example of the text is cross probability $P_c = 0.8$ and variation probability Pm=0.2;

5 The shut standard of GA is when evolve generate number bigger than maximum evolve generate number G=1000 or no ameliorate in fifty generations.

Material algorithm description is as follows:

```
Procedure: General Genetic Algorithms.
Begin
Randomization initial group，P(0)；t=0
Calculate adapt value of individuals in P(0)；
Keep the best individual；
While (dissatisfied stop condition) do
Circulate cross operator；
Circulate variation operator；
Circulate adapt value of individuals in P(t+1)，t=t+1；
If the previous generation individual is better than the present generation individual, then the previous generation individual substitute the other one;
end
end
```

4 Numerical Simulation

Consider the inverse problem of hyperbolic equation parameter identified as follows:

$$p(x)\frac{\partial^2 h}{\partial t^2} = \frac{\partial}{\partial x}(q(x)\frac{\partial h}{\partial x}) + c(x,t)\frac{\partial h}{\partial x} + d(x,t)h + f(x,t),$$

$0 < x < 1, 0 < t < 0.5$

$p(x) = 1 + x, \quad c(x,t) = x + t$,

$d(x,t) = x - t$, $\mathrm{f}(x,t) = -(x^2 + 3x)e^{x+t}$,

The boundary conditions are:

$$h\big|_{x=0} = g_1(t) = e^t, \qquad h\big|_{x=1} = g_2(t) = e^{1+t},$$

$$h\big|_{t=0} = f_1(x) = e^x, \qquad \frac{\partial h}{\partial t}\bigg|_{t=0} = f_2(x) = e^x,$$

The add-conditions are:

$$\mathrm{h}(\mathrm{x_i}, 0.5)(\mathrm{i} = 0,1,2,\ldots\mathrm{N}), \ \mathrm{x_i} = \mathrm{ih}, \mathrm{h} = \frac{1}{\mathrm{N}}, \quad \mathrm{N} = 10 .$$

The system of primary functions $\{\varphi_i(x)\}$ chooses as linear piecewise functions:

$$\varphi_0(x) = \begin{cases} 1 - nx, & x \in \left[0, {}^{1}\!/\!{}_{n}\right] \\ 0, & others \end{cases},$$

$$\varphi_i(x) = \begin{cases} n(x - \dfrac{i-1}{n}), & x \in \left[\dfrac{i-1}{n}, \dfrac{i}{n}\right] \\ n(\dfrac{i+1}{n} - x), & x \in \left[\dfrac{i}{n}, \dfrac{i+1}{n}\right] \\ 0, & others \end{cases},$$

$$\varphi_n(x) = \begin{cases} n(x - \dfrac{n-1}{n}), & x \in \left[\dfrac{n-1}{n}, 1\right] \\ 0, & others \end{cases}$$

Take $\mathrm{n} = 10$, the parameter $\mathrm{q} = \mathrm{q}(\mathrm{x})$ precise solution adopt as:

$$\begin{aligned} q(x) = {} & 2\varphi_0(x) + 2\varphi_1(x) + \varphi_2(x) + \varphi_3(x) \\ & + 0.5\varphi_4(x) + 0\varphi_5(x) + 0.5\varphi_6(x) + \\ & \varphi_7(x) + \varphi_8(x) + 2\varphi_9(x) + 2\varphi_{10}(x) \end{aligned}$$

Take the original parameter as $q_0(x) = \sum_{i=0}^{10} k_i^0 \varphi_i(x)$ and $\mathrm{k_0^0, k_1^0, \ldots\ldots, k_{10}^0}$ are global variations in search interval to search the values approach the precise solution. The number of variations take as nvars=11, the size of group take as psize=20,the maximum evolution number take as maxgen=7000,the cross probability take as pc=0.8,the variance probability take as pm=0.2,and used the GA seek solution as:

$$\begin{aligned} q(x) = {} & 2.177\varphi_0(x) + 2.209\varphi_1(x) + 1.565\varphi_2(x) \\ & + 1.155\varphi_3(x) + 1.161\varphi_4(x) - 0.361\varphi_5(x) \\ & + 1.073\varphi_6(x) + 1.105\varphi_7(x) + 1.024\varphi_8(x) \\ & + 2.470\varphi_9(x) + 1.712\varphi_{10}(x) \end{aligned}$$

The above solution takes as original values at the beginning of the best disturbed iteration method. When the iteration time $\mathrm{n} = 1000$, it can obtain the parameter solution and the approximate solution error is

$$\mathrm{e} = \frac{\|\mathrm{q} - \mathrm{q}_{1000}\|_2}{\mathrm{n}} = 8.17497\times 10\text{-}3.$$

The precise solution and the approximate solution of $\mathrm{q} = \mathrm{q}(\mathrm{x})$ are shown as Figure 1:

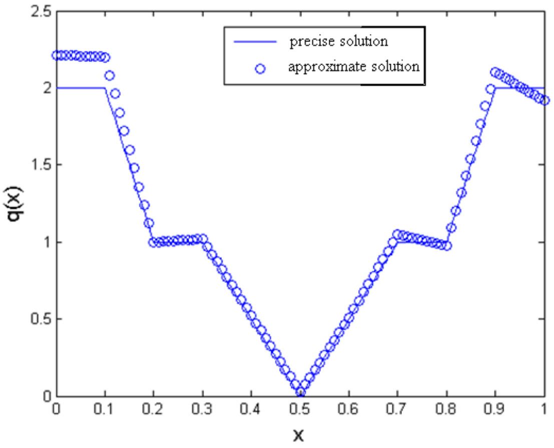


Figure 1 Parameter Indentify Problem

5 Summarize

The problem of parameter identify is a kind of inverse problems of partial differential equation and apply broadly. The best disturbed iteration method is a classical method for inverse problem solving and the method has virtues as high stability and good astringency, but how to obtain the original values for iteration is a hard work.

The text has attempt at this work and bring forward the best disturbed iteration method based on GA to solve parameter identify problem. Because of GA take the group search policy and choose the probability variance rule to control the search directions, so it can search to the values approach precise solution. But the GA is instability and has big calculation, so it can combine with the best disturbed iteration method to find the best optimal solution. The numerical example outcome showed that the solution have good astringency and high precision.

Acknowledgement

Fund item: Hebei Polytechnic University science study imbrues item（z200718）

References

[1] M. S. Pilant, W. Rundell. An inverse problem for a nonlinear parabolic equation. *Commum Partial Differ Equations*. 1986, 11:445~457

[2] Victor Isakov. Stefan Kindermann.. Identification of the diffusion coefficient in one-dimensional parabolic equation. *Inverse Problem*. 2000, 16:665-680

[3] Ting-yan Xiao. Shen-gen Yu. Yan-fei Wang. Numerical Method for Inverse Problem. *Science publishing company*.2003

[4] Xiao-ping Wang. Li-ming Cao. Genetic Algorithms. *Xi'an Jiaotong University publishing company*. 2000

[5] Chao-wei Su. Numerical Method for Inverse Problem of Partial Differential Equation and Application . *Northwestern Polytechnical University publishing company* . 1995.10. Copyright forms

[6] Jin ZQ, Zhou ZF. Inverse problem of engineer hydraulics [M], *Hehai university publisher*, 1997

[7] Li W. Headway of environmental hydraulics [M], *WuHan university publisher*, 1999

[8] Typhoons A, Arsenin V. Solution of ill-posed problems. Winston, Washington,1977

[9] Min T, Zhou XD. A Numerical Solution to Initial Condition Inverse Problem of One-dimensional Unstable Furbulent Diffusion. *Journal of Xi'an University of Technology* (2001) Vol.17No.2

[10] Wang ZZ. Bo T. JinHua JiSuan[M].ChangSha, *National University Of Defence Technology Publisher*，2000

On Numerical Simulation of the Vortex in BEC

Yimin Tian[1*] Mengzhao Qin[2] Xiaofeng Zhu[1] Yongming Zhang[1]

1 Mathematics and Physics Division, Beijing Institute of Graphic Communication, Beijing 102600, China

2 Institute of Computational Mathematics and Scientific/Engineering omputing Academy of
Mathematics and Systems Sciences Academic Sinica
Email: tym66105@yahoo.com.cn

Abstract

Two kind of symplectic difference schemes was used to study the evolution of the vortex in Bose-Einstein Condensate(BEC), the case of one vortex was simulated by numerical method and the evolution of the vortex can be seen in the simulation. We can see that implicit scheme is better than explicit scheme, though they are both symplectic difference schemes.

Keywords: symplectic methods, Bose-Einstein, Condensate; vortex

1 Introduction

The existence of quantized vortices in the Bose- Einstein Condensate(BEC) has been well documented[1,2,3,4].The study of elementary excitations is a task of primary importance of quantum many-body theories. In the case of Bose fluids, in particular, it plays a crucial role in the understanding of the properties of superfluid liquid helium and was the subject of pioneering work by Landau, Bogoliubov, and Feynman[1]. In the last few years, after the experimental realization of BEC in trapped Bose gases, there has been a great surge in the studies of quantized vortices in the BEC, both experimentally and theoretically [4,5,6,7,8,9,10,11]. Theoretical studies of vortices in the BEC experiments have often been made in the frame work of the nonlinear Gross-Pitaevskii equation [9,10,11], well known for superfluids, but which provides a very good description of BEC.

Structure preserving algorithm is a kind of promising method developed recently[12,13,14], YM Tian and MZ Qin used an explicit symplectic scheme for two-dimensional Gross-Pitaevskii equation to investigate the evolution of vortices in a rotating BEC[14], but the discrete conservation law is only discrete quasi norm: $(U^{k+1},U^{k}) = (U^{k},U^{k-1})$. So we try to construct a symplectic methods for two-dimensional Gross-Pitaevskii equation, which conserve the discrete norm: $(U^{k+1},U^{k+1}) = (U^{k},U^{k})$ and investigate the evolution of vortices in a rotating BEC by the symplectic methods presented in this paper.

2 Hamiltonian Formulation

Now we consider the time-dependent G-P equation

$$i\frac{\partial u}{\partial t} = (\nabla - iA)^2 u + \frac{1}{\in^2} |u|^2 u - \frac{a_\in(r)}{\in^2} u = 0 \qquad (1)$$

in D with initial condition $u(r,0) = u_0(r)$ in D and boundary condition $u = 0$. The equation is 2n + 1 dimension Shrodinger equation. The constraint $\int_D |u|^2 =1$ is automatically preserved at all time. Where a(r) $=\alpha_0-(x^2 + \lambda^2 y^2)$ for some constant α_0 which is chose so that the integral of a(r) on the ellipsoid D={a>0}= $\{x^2 + \lambda^2 y^2 < \alpha 0\}$ is equal to 1, which leads to $\alpha^2 = \frac{2\lambda}{\pi}$. If $\lambda = 1$ then $a_\in(r) = a(r) - \in^2 \Omega^2 (x^2 + y^2)$, D is a disc of radius R_0 with $R_0^4 = \frac{2}{\pi}$ and A is a vector potential defined by

[*] Corresponding author, this paper is supported by Beijing education committee founds km200710015013, BIGC founds E-b-07-03 and BIGC teaching reform key project: On thought about second teaching form of mathematics

$$A=\begin{pmatrix}A_1\\A_2\end{pmatrix}=\begin{pmatrix}y\\-x\end{pmatrix}\Omega$$

$cur1\vec{A}\neq 0$, or there will be div(A)=0, and A will be some function's gradient, therefore $\left(\nabla\text{-iA}\right)$ will be some function's gradient too, so equation (1) will be an ordinary Shrodinger equation, where $\in^2=\dfrac{\hbar^2}{2Ngm}\Omega=\dfrac{\Omega_0}{\in\Omega}$, see [9]. The nonlinear Gross-Pitaevskii equation(1) can be written as an infinite-dimensional Hamiltonian system and has a Hamiltonian structure as follows.

Let u = p + iq, here p(x, y, t), q(x, y, t) are real functions of x, y, t, then (1) has the hamiltonian formulation

$$\frac{dz}{dt}=J\frac{\delta H(z)}{\delta z}, \qquad (2)$$

where z = (p, q), J = $\begin{pmatrix}0&1\\-1&0\end{pmatrix}$ and the Hamiltonian is

$$\begin{aligned}\text{H(z)=}&\int\left(\frac{1}{2}\left(p_x^2+p_y^2+q_x^2+q_y^2\right)\right)-p\left(qA_1\right)_x-\\&p(qA_2)_y-pA_1q_x-pA_2q_y-q(pA_1)_x-q(pA_2)_y-\\&qA_1p_x-qA_2p_y+\frac{1}{2}(A_1^2+A_2^2)(p^2+q^2)-\\&\frac{1}{2\varepsilon^2}(R^2-(1+\varepsilon^2\Omega^2)(x^2+y^2)-\\&(p^2+q^2))(p^2+q^2)-\frac{1}{4\varepsilon^2}(p^2+q^2))dxdy\end{aligned}$$

We present a symplectic integrators of order o$((\Delta t)^2+(\Delta x)^2+(\Delta y)^2)$

$$Z_{k+1}-Z_k=\Delta tJ\nabla_zH\left(\frac{Z_{k+1}-Z_k}{2}\right) \qquad (3)$$

let

$$\begin{aligned}U^k&=(u_{1,1}^k,\ldots,u_{1,n}^k,u_{2,1}^k,\ldots,u_{2,n}^k,\ldots,\\&u_{n,1}^k,\ldots,u_{n,n}^k),\\V^k&=(v_{1,1}^k,\ldots,v_{1,n}^k,v_{2,1}^k,\ldots,v_{2,n}^k,\ldots,\\&v_{n,1}^k,\ldots,v_{n,n}^k)\end{aligned}$$

we define inner product (·) and norm ||· || respectively by

$$(\text{U}^{\text{ k}},\text{V}^{\text{ k}})=h\sum_{i,j=1}^{n}u_{ij}^k\bar{v}_{ij}^{k}$$

$$\left\|U^k\right\|^2=(\text{U k ; U k }).$$

We can prove that schemes (3) satisfies discretenorm conservation law:

$$(\text{U}^{\text{ k+1}};\text{U}^{\text{ k+1}})=(\text{U}^{\text{ k}};\text{U k }).$$

3 Numerical Experiments

In this section, we present the numerical simulation results of the G-P equation(1) by using symplectic integrator (4) to study the evolution of vortices. The following pictures show some instances of our numerical experiments of vortices solutions and their evolution behavior of G-P equation in time.

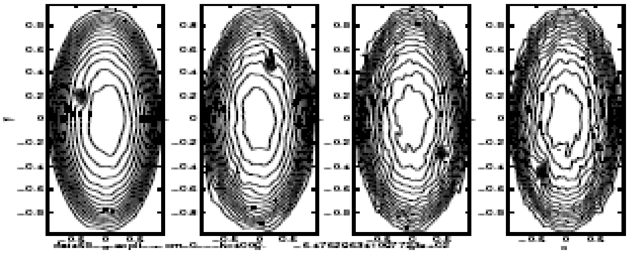


Figure 1 This is the case of one vortex's evolution in real time when Ω= 0 by explicit scheme, from the figure we can see that the vortex migrates to the edge of the trap in the process and the vortex is unstable

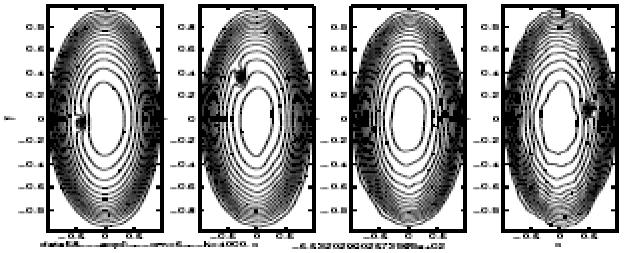


Figure 2 This is the case of one vortex's evolution in real time when Ω =6 by explicit scheme. The vortex moves clockwise in the positive sense

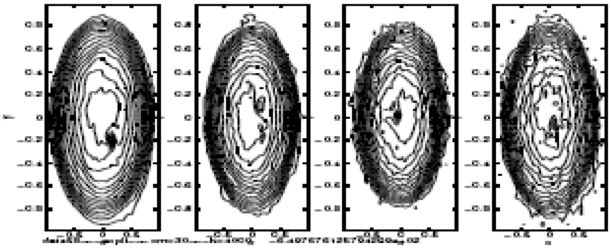


Figure 3 This is the case of one vortex's evolution in real time when Ω = 30by explicit scheme. The vortex moves back towards the center and eventually disappear

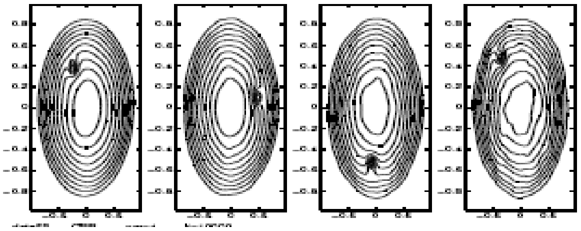


Figure 4 This is the case of four vortices's evolution in real time when Ω=4 by implicit scheme, the vortices moves clockwise in the positive sense and keep quite well the in explicit scheme

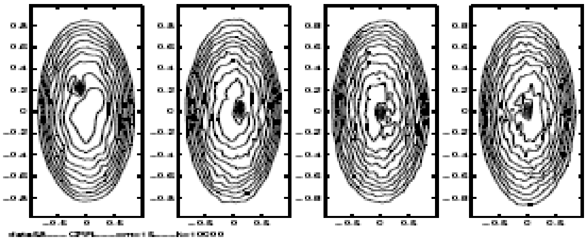


Figure 5 This is the case of four vortices's evolution in real time when Ω = 15by implicit scheme. The vortices moves clockwise in the positive sense and moves back towards the center and still exist

References

[1] F.Dalfovo, S.Giorgini, Theory of Bose Einstein condensation in trapped gases, Rev.Mod.Phys.71, 1999, pp.463-512

[2] K.W. Madison, F. Chevy, W. Wohlleben, and J. Dalibard, Vortex Formation in a Stirred Bose-Eistein Consate, Phys. Rev. Lett. Vol 84,2000, pp.806-809

[3] Chen Shuai, Zhou Xiaoji, Yang Fan, Xia Lin, Sun Yaya, Wang Yiqiu, Chen Xuzong, Analysis of runaway evaporation cooling and Bose-Einstein Condensation by time of flight absorption imaging, Chin. Phys. Lett., Vol.21, 2004, 2105

[4] N.G. Parkerand Adams,Respons of anatomic Bose-Einstein condensate to rotating elliptical trap J. Physic. B, AL. Mol. Opt. Phys.39,2006,43-55

[5] D. L.Feder, C. W. Clark and B.I.Schneider, Vortex Stability of Interacting Bose-.Einstein Condensates Confined in Anisotropic Harmonic TrapsPhys.Rev.Lett., 82, 1999, pp.4956-4959

[6] A.A.Svidzinsky and A.L.Fetter, Stability of a Vortex in a Trapped Bose-Einstein Condensate, Phys.Rev.Lett., 84, 2000, pp.5919-5923

[7] F. Minardia, C. Forta, P. Maddalonib, M. Modugno and M. Inguscio, Time-domain Ramsey interferometry with Bose-Einstein condensates, Sciences - Series IV - Physics Volume 2, Issue 4 , June 2001, pp. 605-612

[8] Yvan Castin and Christopher Herzog, Bose-Einstein condensates in symmetry breaking states, Sciences- Series IV - Physics Volume 2, Issue 3 , April 2001, pp.419-443

[9] Amandine, QiangDu, Vortices in a rotating Bose - Einstein condensate:critical velocities and energy diagrams in the Thomas0Fermi regime, srXiv:condomat/0103299 v1 14 Mar. 2001, pp.1-10

[10] Makoto Tsubota, Kenichi Kasamatsu, Masahito Ueda, Dynamics of vortex lattice formation in a rotating Bose-Einstein consdensate, Physica B 21-22,2003, pp.329-333

[11] D.E. Pelinovsky, P.G. Kevrekidis, D.J. Frantzeskakis, Persistence and stability of discrete vortices in nonlinear Schrodinger Lattices, Physica D 212 (2005) 20-53

[12] Jing-Bo Chen, Qin M Z and Yi-Fa tang, Symplectic and multi-symplectic method for the Nonlinear Schrodinger Equation, Computers and mathematics with Applications 43,2002, pp.1095-1106.

[13] Jing-Bo Chen, A multisymplectic integrator for the periodic nonlinear Schrodinger equation, Appl.Math. Comp. 170(2005), 1394-1417

[14] Yi-Min Tian, Meng-Zhao Qin, Explicit Symplectic Schemes for Investigating the Evolution of Vortices in a Rotating Bose - Einstein Condensate, Computer Physics Communications 155 (2003) 123-143

Existence and Asymptotic of Solution for Third Order Boundary Value Problem

Guocan Wang

Department of Mathematics, Dalian Jiaotong University, Dalian, Liaoning 116028, P.R.China
Email: wanggc@dl.cn

Abstract

Third order singularly perturbed boundary value problem by means of differential inequality theories is studied. Based on the results of second order nonlinear boundary value problems of Volterra type integro-differential equation, the upper and lower solutions method of third order nonlinear boundary value problems was established. Specific upper and lower solutions were constructed, and existence and asymptotic estimates of solutions under suitable condition were obtained. The result showed that is seems new to apply these techniques to solving these kinds of third order singularly perturbed boundary value problem. An example is given to demonstrate the applications.

Keywords: nonlinear differential equation, singular perturbation, third order boundary value problem

1 Introduction

The singularly perturbed boundary value problems for the third order nonlinear ordinary differential equations, which are important not only for theoretical purpose but also for applications in fluid dynamics, have been studied in [1-4], We consider the boundary value problem involving the third order nonlinear ordinary differential equation with a small parameter $\varepsilon > 0$,

$$\varepsilon x''' = f(t, x, x', x'', \varepsilon) \tag{1}$$

and the following boundary conditions.

$$x(0) = A, x'(0) = x'(1), x''(0) = x''(1) \tag{7}$$

In the present paper, we study the existence and asymptotic estimates of solutions, in the general sense, for the singularly perturbed boundary value problems (1), (2) by making use of Volterra type integral operator and differential inequality techniques. The use of these techniques to study for these kinds of singularly perturbed boundary value problems seems to be new.

2 Preliminaries

The purpose of this section is to state some preliminaries which be needed in the sequel.

Let us consider the following Volterra type nonlinear boundary value problem

$$u'' = f(t, Tu, u, u') \tag{3}$$

$$u(0) = u(1), u'(0) = u'(1) \tag{4}$$

where

$$[Tu](t) = \varphi(t) + \int_0^t K(t,s)u(s)ds, \quad \varphi(t) \in C[0,1]$$

and $K(t,s) \geq 0$ for $(t,s) \in [0,1] \times [0,1]$.

We call $\beta(t)$ an upper solution of the equation (3) on [0,1], if $\beta(t) \in C^2[0,1]$ and for $0 \leq t \leq 1$

$$\beta''(t) \leq f\left(t, [T\beta](t), \beta(t), \beta'(t)\right)$$

and we call $\alpha(t)$ a lower solution of the equation (1) on [0,1] if $\alpha(t) \in C^2[0,1]$ and for $0 \leq t \leq 1$

$$\alpha''(t) \geq f\left(t, [T\alpha](t), \alpha(t), \alpha'(t)\right)$$

Lemma 1. Assume that

(1) $f(t, v, u, w) \in C([0,1] \times R^3)$ and it is nonincreasing in v.

(2) $\alpha(t)$, $\beta(t)$ be lower and upper solution relate to, satisfying the following hypotheses

$$\alpha(0) = \alpha(1), \beta(0) = \beta(1)$$

$$\alpha'(0) \geq \alpha'(1), \beta'(0) \leq \beta'(1)$$

then the boundary value problem (3),(4) has a solution $u(t)$ such that

$\alpha(t) \le u(t) \le \beta(t)$, $0 \le t \le 1$

Proof: First, we assume that $\alpha(0) = \beta(0)$. Thus, by $\alpha(t) \le \beta(t)$ for $0 \le t \le 1$, we imply that $\alpha\,'(0) \le \beta'(0)$, $\alpha'(1) \ge \beta'(1)$. On the other hand, $\alpha'(0) \ge \alpha'(1)$, $\beta'(0) \le \beta'(1)$ from condition (2). This is $\alpha'(0) = \alpha'(1), \beta'(0) = \beta'(1)$. Hence, the solution $u(t)$ of (3) on [0.1], which satisfies that $u(0) = \alpha(0)$, $u(1) = \beta(1)$, $\alpha(t) \le u(t) \le \beta(t)$, is a solution for the boundary value problem (3) and (4).

Next we consider that $\alpha(0) < \beta(0)$. Let $\Omega(A)$ denote the set of soution $u(t)$ to Eq.(3) such that

$u(0) = u(1) = A, \alpha(0) \le \beta(0)$

and $\alpha(t) \le u(t) \le \beta(t)$ on [0,1].

By Theorem 1 of [3], it is clear that Ω is nonempty for all $\alpha(0) \le A \le \beta(0)$. For $u(t) \in C^3\,[0,1]$, define the function

$$h(u'(0), u'(1)) = u'(0) - u'(1).$$

We notice that $u(t) \in \Omega(\beta(0))$ implies $u'(0) \le \beta'(0) \le \beta'(1) \le u'(1)$, which yields $h(u'(0), u'(1)) \le 0$. Similarly, $u(t) \in \Omega(\alpha(0))$ imply $u'(0) \ge \alpha'(0) \ge \alpha'(1) \ge u'(1)$ so that $h(u'(0), u'(1)) \ge 0$. Conesquenently, employing similar arguments as in the proof of Theorem1.5.2 of [5], for any $\varepsilon > 0$, $\alpha(0) \le A \le \beta(0)$, we can prove that Eq.(3) has a solution $u(t,\varepsilon)$ which satisfies the following results:

$u(0,\varepsilon) = u(1,\varepsilon) = A, \left|h(u'(0,\varepsilon), u'(1,\varepsilon))\right| < \varepsilon$

$\alpha(t) \le u(t,\varepsilon) \le \beta(t), 0 \le t \le 1$

Hence , let $\varepsilon \to 0$, the proof is complete.

Lemma 2. Assume that

(1) $f(t,x,x',x'')$ is nonincreasing in x and continuous on $[0,1] \times R^3$,

(2) there exist two functions $\alpha(t), \beta(t) \in C^3[0,1]$, such that

$\alpha'''(t) \ge f\left(t, \alpha(t), \alpha'(t), \alpha''(t)\right)$, $\beta'''(t) \le f(t, \beta(t), \beta'(t), \beta''(t))$ for $0 \le t \le 1$, and

$\alpha(0) \le A \le \beta(0), \alpha'(0) = \alpha'(1), \beta'(0) = \beta'(1)$, $\alpha''(0) \ge \alpha''(1), \beta''(0) \le \beta''(1)$

then the boundary value problem

$$x''' = f(t,x,x',x'') \tag{5}$$

$$x(0) = A, x'(0) = x\,'(1), x''(0) = x''(1) \tag{6}$$

has a solution $x(t)$ such that $\alpha(t) \le x(t) \le \beta(t)$, $0 \le t \le 1$.

Proof: Let $x' = u$, then $x(t) = A + \int_0^t u(s)ds$. Thus, boundary value problem (5) with (6) can be written as the following second order integrodifferential equation of Volterra type given by

$$u'' = f\left(t, A + \int_0^t u(s)ds, u, u'\right) \tag{7}$$

$$u(0) = u(1), u'(0) = u'(1) \tag{8}$$

However, for the successful employment of the result of Lemma 1, we need to construct lower and upper solution for (7) by using $\alpha(t)$, $\beta(t)$ and hypotheses (1)-(3). Therefore, we set

$\hat{\alpha}(t) = \alpha(t) + \delta_1, \hat{\beta}(t) = \beta(t) - \delta_2$

where $\delta_1 = A - \alpha(0)$, $\delta_2 = \beta(0) - A$. Then, it is clear that $\hat{\alpha}(0) = A = \hat{\beta}(0)$. Moreover, if we write $\hat{\alpha}'(t) = \alpha_*(t)$, $\hat{\beta}'(t) = \beta_*(t)$, it is easy to show that $\alpha_*(t) \le \beta_*(t)$, for $0 \le t \le 1$.

Note that

$\hat{\alpha}(t) = A + \int_0^t \alpha_*(s)ds$, $\hat{\beta}(t) = A + \int_0^t \beta_*(s)ds$, we obtain

$$\alpha_*''(t) \ge f\left(t, A + \int_0^t \alpha_*(s)ds, \alpha_*(s), \alpha_*'(s)\right)$$

$$\beta_*''(t) \le f\left(t, A + \int_0^t \beta_*(s)ds, \beta_*(s), \beta_*'(s)\right)$$

$\alpha_*(0) = \alpha_*(1), \beta_*(0) = \beta_*(1)$

$\alpha_*'(0) \ge \alpha_*'(1), \beta_*'(0) \le \beta_*'(1)$

Hence, we see that $\alpha_*(t)$, $\beta_*(t)$ are lower and upper solutions for (7) with (8). Consequently we obtain a solution $u(t)$ of (7) and (8), such that $\alpha_*(t) \le u(t) \le \beta_*(t)$. From the relation $x'(t) = u(t)$, we can recover

$$x(t) = A + \int_0^t u(s)ds$$

Obviously, $\alpha(t) \le x(t) \le \beta(t)$, $0 \le t \le 1$. So, we know that the proof is complete.

3 Existence and Asypampotic Estimate Results

The existence and asymptotic estimates of solutions for the boundary value problem (1), (2) are discussed in this section.

When $\varepsilon = 0$, the boundary value problem (1), (2) is reduced to

$$0 = f(t, x, x', x'', 0), x(0) = A \qquad (9)$$

Let $\varepsilon_0 > 0$ be a constant and

$$D = \left\{ (t, x, x', x'', \varepsilon) \left| \begin{array}{l} 0 \le t \le 1, \\ -\infty < x, x', x'' < \infty, \\ 0 \le \varepsilon \le \varepsilon_0 \end{array} \right. \right\}$$

We make the following hypotheses:

(1) The reduced problem (9) has a solution $x_0(t) \in C^3[0,1]$.

(2) $f(t, x, x', x'', \varepsilon)$ is continuously differentiable and bounded on D, and for $(t, x, x', x'', \varepsilon) \in D$,

$-l \le f_x(t, x, x', x'', \varepsilon) \le 0$, $f_{x'}(t, x, x', x'', \varepsilon) \ge m$, $f_{x''}$ $(t, x, x', x'', \varepsilon) \ge 0$ with some constants $m > 0$, $l > 0$.

Theorem Assume that (1)-(2) hold. Then for sufficiently small $\varepsilon > 0$, the boundary value problem (1), (2) has a solution $x(t, \varepsilon)$, such that $\left| x(t, \varepsilon) - x_0(t) \right| \le D_1 e^{\lambda_1 t} \ + D_2 e^{\lambda_2 (t-1)} + D_3$, where λ_1, λ_2 are two roots of equation $\varepsilon\lambda^3 - m\lambda + l = 0$, and such that $-2\sqrt{\dfrac{m}{\varepsilon}} < \lambda_1 < \quad -\sqrt{\dfrac{m}{\varepsilon}}, \dfrac{1}{2}\sqrt{\dfrac{m}{\varepsilon}} < \lambda_2 < \sqrt{\dfrac{m}{\varepsilon}}$, $D_i = O(\sqrt{\varepsilon}), i = 1,2,3$ ·

Proof: Assume that for $(t, x, x', x'', \varepsilon) \in D$, there exist positive number M_1 , M_2 ,so that $\left| f_\varepsilon(t, x, x', x'', \varepsilon) \right| \le M_1$, $\left| x_0'''(t) \right| < M_2$.

Obviously, there exists real number $s_1, s_2 \in (-\infty, +\infty)$, $s_1 < s_2$ such that

$$s_1 < x_0^{'}(0) < s_2, s_1 < x_0^{'}(1) < s_2$$

and we let $c_1{}^0 = x_0\,(0) - s_1, c_2{}^0 = x_0\,(1) - s_1,$

$$c_3{}^0 = s_2 - x_0\,(0), c_4{}^0 = s_2 - x_0\,(1)$$

$$k_i = 1 + \max\left\{ c_i{}^0, c_{i+2}{}^0 \right\}, i = 1,2 \ \ k = \max\left\{ k_1, k_2 \right\}.$$

$$F_0(c_1, c_2, \varepsilon) = x_0'(0) - c_1 - c_2 e^{-\lambda_2} - \frac{2(M_1 + M_2 + \dfrac{k}{\sqrt{m}} + 1)\lambda_3\sqrt{\varepsilon}}{l} - s_1$$

$$F_1(c_1, c_2, \varepsilon) = x_0'(1) - c_1 e^{\lambda_1} - c_2 - \frac{2(M_1 + M_2 + \dfrac{k}{\sqrt{m}} + 1)\lambda_3\sqrt{\varepsilon}e^{\lambda_3}}{l} - s_1$$

$$G_0(c_3, c_4, \varepsilon) = x_0'(0) + c_3 + c_4 e^{-\lambda_2} + \frac{2(M_1 + M_2 + \dfrac{k}{\sqrt{m}} + 1)\lambda_3\sqrt{\varepsilon}}{l} - s_2$$

$$G_1(c_3, c_4, \varepsilon) = x_0'(1) + c_3 e^{\lambda_1} + c_4 + \frac{2(M_1 + M_2 + \dfrac{k}{\sqrt{m}} + 1)\lambda_3\sqrt{\varepsilon}e^{\lambda_3}}{l} - s_2$$

where λ_1, λ_2, λ_3 are three roots of $\varepsilon\lambda^3 - m\lambda \quad + l = 0$, and such that

$$-2\sqrt{\frac{m}{\varepsilon}} < \lambda_1 < -\sqrt{\frac{m}{\varepsilon}}, \ \frac{1}{2}\sqrt{\frac{m}{\varepsilon}} < \lambda_2 < \sqrt{\frac{m}{\varepsilon}},$$

$$\frac{l}{m} < \lambda_3 < \frac{l+m}{m}$$

then $F_i(c_1{}^0, c_2{}^0, 0) \equiv G_i(c_3{}^0, c_4{}^0, 0) \equiv 0, i = 0,1$. In addition, we have

$$\frac{\partial(F_0, F_1, G_0, G_1)}{\partial(c_1, c_2, c_3, c_4)} = \begin{vmatrix} -1 & -e^{-\lambda_2} & 0 & 0 \\ -e^{\lambda_1} & -1 & 0 & 0 \\ 0 & 0 & 1 & e^{-\lambda_2} \\ 0 & 0 & e^{\lambda_1} & 1 \end{vmatrix} \qquad (10)$$

$$= (1 - e^{\lambda_1 - \lambda_2})^2 \ne 0$$

Hence, there exist a positive number $\varepsilon_1 > 0$ and unique continuous function set $c_i = c_i(\varepsilon)$ such that $c_i(0) = c_i{}^0$, ($i = 1,2,3,4$), and

$$\begin{aligned} &F_i(c_1(\varepsilon), c_2(\varepsilon), \varepsilon) \equiv G_i(c_3(\varepsilon), \\ &c_4(\varepsilon), \varepsilon) \equiv 0, i = 0,1, 0 \le \varepsilon \le \varepsilon_1 \end{aligned} \qquad (11)$$

Thus we can select sufficiently small ε_1>0, we obtain

$$\frac{1}{2}c_i{}^0 \le c_i(\varepsilon) \le c_i{}^0 + 1, i = 1,2,3,4,$$

$$0 < c_i(\varepsilon), c_{i+2}(\varepsilon) \le k_i, i = 1,2$$

For any $\varepsilon \in (0, \varepsilon_1]$, we take the following upper and lower solution:

$$\alpha(t) = x_0(t) - \frac{c_1(\varepsilon)}{\lambda_1}[e^{\lambda_1 t} - 1] - \frac{c_2(\varepsilon)}{\lambda_2}e^{\lambda_2(t-1)} - \frac{(M_1 + M_2 + \dfrac{k}{\sqrt{m}} + 1)\sqrt{\varepsilon}}{l}(2e^{\lambda_3} - 1)$$

$$\beta(t) = x_0(t) + \frac{c_3(\varepsilon)}{\lambda_1}[e^{\lambda_1 t} - 1] + \frac{c_4(\varepsilon)}{\lambda_2} e^{\lambda_2(t-1)} +$$

$$\frac{(M_1 + M_2 + \frac{k}{\sqrt{m}} + 1)\sqrt{\varepsilon}}{l}(2e^{\lambda_3} - 1)$$

Now, for $(t,\varepsilon) \in [0,1] \times [0,\varepsilon_0]$, it is obvious that $\alpha(t,\varepsilon) < \beta(t,\varepsilon)$, $\alpha'(t,\varepsilon) < \beta'(t,\varepsilon)$, and

$\alpha''(t,\varepsilon) < \beta''(t,\varepsilon)$

$$f\left(t,\beta(t,\varepsilon),\beta'(t,\varepsilon),\beta''(t,\varepsilon),\varepsilon\right) - \varepsilon\beta'''(t,\varepsilon)$$

$$= f\left(t,\beta(t,\varepsilon),\beta'(t,\varepsilon),\beta''(t,\varepsilon),\varepsilon\right) -$$

$$f\left(t,\beta(t,\varepsilon),\beta'(t,\varepsilon),x_0''(t),\varepsilon\right) +$$

$$f\left(t,\beta(t,\varepsilon),\beta'(t,\varepsilon),x_0''(t),\varepsilon\right) -$$

$$f\left(t,\beta(t,\varepsilon),x_0'(t),x_0''(t),\varepsilon\right) +$$

$$f\left(t,\beta(t,\varepsilon),x_0'(t),x_0''(t),\varepsilon\right) -$$

$$f\left(t,x_0(t),x_0'(t),x_0''(t),\varepsilon\right) +$$

$$f\left(t,x_0(t),x_0'(t),x_0''(t),\varepsilon\right) -$$

$$f\left(t,x_0(t),x_0'(t),x_0''(t),0\right) - \varepsilon\beta'''(t,\varepsilon) \ge$$

$$m(\beta'(t,\varepsilon) - x_0'(t)) - l(\beta(t,\varepsilon) - x_0(t)) -$$

$$\varepsilon(M_1 + M_2) - \varepsilon(\beta'''(t,\varepsilon) - x_0'''(t))$$

$$= \frac{c_3}{\lambda_1} e^{\lambda_1 t}(m\lambda_1 - l - \varepsilon\lambda_1^3) + \frac{c_4}{\lambda_2} e^{\lambda_2(t-1)}(m\lambda_2 - l - \varepsilon\lambda_2^3) +$$

$$\frac{2(M_1 + M_2 + \frac{k}{\sqrt{m}} + 1)\sqrt{\varepsilon}}{l} e^{\lambda_3 t}(m\lambda_3 - l - \varepsilon\lambda_3^3) +$$

$$(M_1 + M_2 + \frac{k}{\sqrt{m}} + 1)\sqrt{\varepsilon} - \varepsilon(M_1 + M_2) + \frac{c_3}{\lambda_1} > 0$$

Simiarly, we obtain that

$$f\left(t,\alpha(t,\varepsilon),\alpha'(t,\varepsilon),\alpha'(t,\varepsilon),\varepsilon\right) - \varepsilon\alpha'''(t,\varepsilon) < 0$$

From construct of c_i and $c_{i+2}(i = 1,2)$, it is clear that $\alpha(0) < A < \beta(0)$, and

$$\alpha\,\mathrm{n}(0) = F_0(c_1(\varepsilon),c_2(\varepsilon),\varepsilon) + s_1 = s_1$$

$$\alpha'(1) = F_1(c_1(\varepsilon),c_2(\varepsilon),\varepsilon) + s_1 = s_1$$

$$\beta'(0) = G_0(c_3(\varepsilon),c_4(\varepsilon),\varepsilon) + s_2 = s_2$$

$$\beta'(1) = G_1(c_3(\varepsilon),c_4(\varepsilon),\varepsilon) + s_2 = s_2$$

thus $\alpha'(0) = \alpha'(1), \beta'(0) = \beta'(1)$, by (11) and the representations of $\beta(t,\varepsilon)$ and $\alpha(t,\varepsilon)$, reducing ε_1 ,

if necessary, we have

$\beta''(1) \ge 0$, $\beta''(0) \le 0$, $\alpha''(0) \ge 0$, $\alpha''(1) \le 0$,

This is $\alpha''(0) \ge \alpha''(1), \beta''(0) \le \beta''(1)$ .Thus , the conditions of Lemma 2 are all satisfied, and hence the boundary value problem (1)-(2) has a solution $x(t,\varepsilon)$ satisfying the inequality $\alpha(t,\varepsilon) \le x(t,\varepsilon) \le \beta(t,\varepsilon)$, $0 \le t \le 1$.

It follows from the representations of $\alpha(t,\varepsilon)$ and $\beta(t,\varepsilon)$ that the estimate:

$\left(x(t,\varepsilon) - x_0(t)\right(\le D_1 e^{\lambda_1 t} + D_2 e^{\lambda_2(t-1)} + D_3$, $(x,\varepsilon) \in$ $[0,1] \times [0,\varepsilon_1]$.

4 Examples

A example illustrating the applicability of the main results in the present paper is given in this section. The following example with certain generality shows that the assumptions of these results are appropriate and convenient for applications and that the results can be applied in a wide range.

Let us consider the boundary value problem

$$\varepsilon x''' = x' \exp[\arctan(x')^2] + (x'')^3 - \frac{x\sin^2(\varepsilon t)}{\sqrt{1+x^2}}$$

$$x(0) = 0, \quad x'(0) = x(1), \quad x''(0) = x''(1)$$

It is easily seen that the reduced problem of the boundary value problem (12),(13)

$x' \exp[\arctan(x')^2] + (x'')^3 = 0$, $x(0) = 0$ has a solution $x_0(t) = 0$

Let $f(t,x,x',x'',\varepsilon) = x' \exp[\arctan(x')^2] + (x'')^3$

$$-\frac{x\sin^2(\varepsilon t)}{\sqrt{1+x^2}}$$

then $f(t,x,x',x'',\varepsilon)$, and their first order partial derivatives with respect to x,x',x'',ε are continuous and $f_{x''} \ge 0, f_{x'} \ge 1, \quad -1 \le f_x \le 0.$ Therefore, by theorem , for sufficiently small $\varepsilon > 0$, problem (12), (13) has a solution $x(t,\varepsilon)$ satisfying asymptotic estimates of Theorem .

Acknowledgements

The author thanks professor Li Yong for helpful refereeing this paper.

Reference

[1] Zhao Weili, Singular perturbations for third order nonlinear boundary value problrm, Nonlinear Analysis, 17(10),1994, pp.1225-1242

[2] Wang Guocan, Asymptotic estimation of Robin boundary value problem for third nonlinear equation, Soochow Journal of Mathematics, 23(1),1997, pp.73-80

[3] Zhou Qinde, Singular perturbations for Volterra type integrodifferential equation, Appl. Math. JCU, 3(3),1988, pp.392-400

[4] Zhang Xiang, Singular perturbations for a third order boundary value problem, J.of Anhui Normal University, 18(1),1995, pp.1-5

[5] Bernfeld S.R. and Lashmikanthan V., An introduction to nonlinear boundary value problems, New York, Academic press, 1974

[6] Wang Guocan,Third order nonlinear singularly perturbed boundary value problem, Applied Mathematics and Mechanics,23(6),2002, pp.670-677

[7] Wang Guocan, Some new results for nonlinear boundary value problems of mixed type integro-differential equation, Ann.of Diff.Eqs19(2),2003, pp.202-208

[8] Ma Jiaqi, Singularly perturbed Robin boundary value problems for nonlinear systems, Mathematic Applicata,11(2),1998, pp.113-115

[9] Cheng Xiu, Singularly perturbed of nonlinear Robin boundary value problems ,13(3),1997, pp.43-45

[10] Wang Xiuqun, Three-order nonlinear Singular perturbation Robin boundary value problem, Journal of Fujian Teachers University(natural science),15(2),1999, pp.1-4

Nonmonotone AdaptiveTrust-region Method for Nonlinear Equations

Dongjin Yuan Haiyan Zhao Fu Wang

Yangzhou Univercity JiangSu Province China
Email: djyuan@yzu.edu.cn; zhaohaiyan-41@163.com; wangfu666@sohu.com

Abstract

The trust region method used to solve the nonlinear programming has been found for last 20 years and is now viewed as an efficient optimization method .It is a key channel to solve Maratos effect phenomenon. In this paper ,an adaptive trust region method with nonmonotone technique for nonlinear equations is proposed and analyzed.

The globle method is efficient. Convergence results of the algorithm are established. Numerical results show that the new Algorithm is successful. Therefore , the algorithm in the life sciences, water sciences, earth sciences, engineering and technology, natural sciences and social sciences, such as the economic and financial fields have extensive and important application, therefore, this algorithm has a very good theoretical and practical significance.

Keywords: Trust region method, Nonlinear equations, Globle convergence, Nonmonotone methods, adaptive

1 Nonmonotone Adaptive Trust Region Algorithm

We consider the following nonlinear equations

$$F(x)=0 \tag{1.1}$$

Where F: $R^n \to R^n$ is twice continuously differentiable.

The following optimization problem of nonlinear equations is

$$\min_{x\in R^n}\psi(x)=\frac{1}{2}\|F(x)\|^2 = \frac{1}{2}F(x)^T F(x) \tag{1.2}$$

If X^* is not empty and $x^*\in X^*$, x^* is the solution of Eq.(1.2).

Trust region method is a robust iterative method. It requires the calculation of a trail step by solving the following subproblem which is

$$\min\phi_k(d)=\frac{1}{2}\left\|F(x)+J(x)^T d\right\|^2 \quad s.t\ \|d\|\le\Delta_k \tag{1.3}$$

To find an approximate point x_{k+1} from the following equation which is

$$\min\phi_k(d)=\frac{1}{2}F(x)^T F(x)+(J(x)^T F(x))d + \frac{1}{2}d^T(J(x_k)^T J(x_k))d \tag{1.4}$$

$$s.t\ \|d\|\le\Delta_k$$

The gradient at the current iterative point x is

$$g(x)=J(x)^T F(x)=\sum_{i=1}^m f_i(x)\nabla f_i(x)$$

The symmetric matrix $G(x)$ is either the Hessian of $\psi(x)$ or an approximation to it, and $G(x)$ may be not positive definite,where $G(x)$ is

$$G(x)=\sum_{i=1}^m(\nabla f_i(x)\nabla f_i(x)^T + f_i(x)\nabla^2 f_i(x)) = J(x)^T J(x)$$

If d_k is the solution of Eq(1.3),we set

$x_{k+1}=x_k+\alpha_k d_k$, otherwise, we set $x_{k+1}=x_k$,at the same time we adjust Δ_k to Δ_{k+1} .

But we can adopt a nonmontone technique.

Recently, the nonmonotone technique for optimization and this technique has been combined with the trust region algorithm to deal with optimization problem ([19,20,21,22,23,24,25]).

* This work was supported by National Natural Science Foundation of China(60774073)

According to r_k^p, the new function value $\psi(x_k + d_k^p)$ is smaller than the previous iterations $\psi(x_{l(k)})$

When d_k^p is accepted , it is not sure that

$$\psi(x_k + d_k^p) < \psi(x_k),$$

But we are sure $\psi(x_k + d_k^p) < \psi(x_{l(k)})$,

$\{\psi(x_{l(k)})\}$ is a monotone nonincreasing subsequence.

In this paper, we will discuss a nonmonotone version of adaptive trust region algorithm proposed in [Hong wei Li]. The numerical results show that our algorithm is effective.

This paper is organized as follows.Next section we describe our nonmonotone adaptive trust region algorithm.In section 2 ,we discuss the global convergence results of the proposed algorithm.The numerical results are reported in section 3 ,which show that our algorithm is effective.

At the current iterative point x_k,the subproblem with adaptive radius is[9,10]

$$\min \phi_k(d) = \frac{1}{2}\left\|F(x) + J(x)^T d\right\|^2$$
$$s.t. \quad \|d\| \le \frac{c^p \|F(x_k)\|^2}{\|J(x_k)F(x_k)\|} \tag{1.5}$$

where $0 < c < 1$, p is a nonnegative integer.

We use the nonmonotone trust region technique [1,2,3,4,5,6].

Let

$$\psi(x_{l(k)}) = \max_{0 \le j \le m(k)} \psi_{k-j}$$
$$\psi_{k-j} = \psi(x_{k-1})$$

where $m(0) = 0$, $0 \le m(k) \le \min\{m(k-1)+1, M\}$, $k \ge 1$

d_k^p is the solution of (1.5). And the actual reduction of $\psi(x)$ is defined by

$$Ared_k(d_k^p) = \psi(x_{l(k)}) - \psi(x_k + d_k^p)$$

And the predictive reduction of $\psi(x)$ is

$$\Pr ed_k(d_k^p) = \psi(x_k) - \phi_k(d_k^p)$$

Further,set: $r_k^p = \dfrac{ared_k}{\Pr ed_k} = \dfrac{\psi(x_{l(k)}) - \psi(x_k + d_k^p)}{\psi(x_k) - \phi_k(d_k^p)}$

Algorithm1.1:

Step 1:Given $0 < c < 1$, $\varepsilon > 0$, $0 < \eta < 1$, $c_2 > 0$, $x_0 \in R^n$,and compute $J(x_0), F(x_0)$, $p := 0$, $k := 0$

Step2:If $\left\|J(x_k)^T F(x_k)\right\| \le \varepsilon$, stop . Otherwise, solve the subproblem (1.5) approximately with d_k^p

Step3:Set $\psi(x_{l(k)}) = \max\limits_{0 \le j \le m(k)} \psi_{k-j}$,

compute $r_k^p = \dfrac{Ared_k}{\Pr ed_k} = \dfrac{\psi(x_{l(k)}) - \psi(x_k + d_k^p)}{\psi(x_k) - \phi_k(d_k^p)}$

Step 4:If $r_k^p > \eta$,then we set $x_{k+1} = x_k + d_k^p$ go to step 5, otherwise set $p := p+1$,go to step2

Step5:Set $m(k) = \min\{m(k-1)+1, M\}$

Step6: $k := k+1$ $p = 0$ go to step1.

Assumption1.1[7]:

(1) $F(x)$ is twice continuously differentiable.,and X^* is not empty.

(2) $\{x_k\}$ is non-increasing.

From Assumption(1.1), ∍ $M > 0$,

$$\forall\ k > 1, \left\|J(x_k)J(x_k)^T\right\| \le M .$$

Lemma1.1[8]: Satisfied Assumption 1.1 and in the level set $\Omega_0 = \{x \in R^n \mid \psi(x) \le \psi(x_0)\}$,the $\psi(x)$ is continuously differentiable,then $\{\psi(x_{l(k)})\}$ is monotone nonincreasing.

Lemma1.2: If Assumption 1.1 holds, then

$$\left|Ared_k(d_k^p) - \Pr ed_k(d_l^p)\right| \le O(\left\|d_k^p\right\|^2)$$

Lemma1.3:

$$\Pr ed_k(d_k^p) \ge \frac{c^p \|J(x_k)F(x_k)\|^2}{2M}$$

Proof: From the definition of d_k^p and setting $\alpha \in [0,1]$,we have

$$\Pr ed_k(d_k^p) = \psi(x_k) - \varphi_k(d_k^p)$$
$$\ge \psi(x_k) - \varphi_k\left(-\frac{\alpha c^p \|F(x_k)\|^2}{\|J(x_k)F(x_k)\|^2}\right)J(x_k)F(x_k)$$

$$= -\frac{1}{2}\frac{\alpha^2 c^{2p}\|F(x_k)\|^4}{\|J(x_k)F(x_k)\|^4}\left\|J(x_k)^T J(x_k)F(x_k)\right\|^2$$

$$+\alpha c^p \|F(x_k)\|^2$$

$$\geq \alpha c^p \|F(x_k)\|^2 - \frac{1}{2}\frac{\alpha^2 c^{2p}\|F(x_k)\|^4}{\|J(x_k)F(x_k)\|^4}$$

$$M\|J(x_k)F(x_k)\|^2$$

$$\geq \alpha c^p \|F(x_k)\|^2 - \frac{1}{2}\frac{\alpha^2 c^{2p}\|F(x_k)\|^4 M}{\|J(x_k)F(x_k)\|^2}$$

because of $\alpha = \dfrac{\|J(x_k)F(x_k)\|^2}{M\|F(x_k)\|^2} \leq 1$

so $\Pr ed_k(d_k^p) \geq \dfrac{c^p\|J(x_k)F(x_k)\|^2}{2M}$.

Lemma1.4:Suppose that Assumption 1.1 holds.

Then Algorithm 1.1 cannot stop infinitely between Step 2 and Step 4.

Proof: Suppose that Algorithm 1.1 stop infinitely between Step 2 and Step 4 at iteration times with k.And we define the cycling index at iteration k by $l(k)$.Then we have

$\forall i = 1,2,\cdots$ having $x_{k+i} = x_k$, $p = i$ $\|J(x_k)F(x_k)\| > \varepsilon$

$\therefore$ when $c^i \to 0$, $(i \to \infty)$

$$r_k^i = \frac{\psi(x_{l(k)}) - \psi(x_k + d_k^i)}{\psi(x_k) - \phi_k(d_k^i)} \leq \eta \qquad i = 2,3\cdots$$

then (1.6).

On the other hand having

$$\left|\frac{\psi(x_{l(k)}) - \psi(x_k + d_k^i)}{\Pr ed_k(d_k^i)} - 1\right|$$

$$= \left|\frac{\psi(x_{l(k)}) - \psi(x_k + d_k^i) - \Pr ed_k(d_k^i)}{\Pr ed_k(d_k^i)}\right|$$

$$\leq \frac{O(\|c_i\|^2)}{O(c_i)} \to 0$$

and for sufficiently large i ,we have

$$\frac{\psi_{l(k)} - \psi(x_k + d_k^i)}{\Pr ed_k(d_k^i)} > \eta$$

This is a contradiction.

Lemma1.5: Satisfied Assumption 1.1 , $\{x_k\}$ is generated from Algorithm 1.1.Then we have $\{x_k\} \subset \Omega_0$ $\forall k$,where

$\Omega_0 = \{x \in R^n \mid \psi(x) \leq \psi(x_0)\}$ boundary, and $\psi(x)$ is constiously and differentable in level set Ω_0.

Proof: when $k = 0$,from Ω_0, we can know $x_0 \in \Omega_0$ obviously.

Suppose to $k \leq m$ and $\forall$ m is positive integral. Then we have $x_k \in \Omega_0$.

Then we know $\psi(x_k) \leq \psi(x_0)$ $\quad k = 0,1,2,\cdots,\ m.$ Then when $k = m+1$, from Algorithm 1.1 we know

$$\psi_{l(m)} \leq \psi(x_0)$$

By induction,we can show that $\psi_{m+1} \leq \psi_0$, then $x_{m+1} \in \Omega_0$.

From inducing ,we can know an assumption, $\{x_k\} \subset \Omega_0$ $\forall k$.

Th1.1 If Assumption 1.1satisfied, then Algorithm as limited step terminate in $\psi(x_k)$ of stability point , $\{x_k\}$ generated from Algorithm1.1 satisfied

$$\lim_{k\to\infty}\|J(x)F(x)\| = 0 \tag{1.7}$$

Proof: $\because$ $\forall x \in R^n$, $\psi(x) \geq 0$, $\{\psi(x_k)\}$ is boundary. Suppose that Algorithm1.1

produce a little endless sequence $\{x_k\}$ to satisfy (1.7). The type step established, then existing and making

$k \subseteq \{0,1,2,3,\cdots\}$ and $\varepsilon > 0$ make

$$\|J(x_k)F(x_k)\| \geq \varepsilon \quad k \in K \tag{1.10}.$$

For $\left\|J(x_k)J(x_k)^T\right\| \leq M$ and $\|J(x_k)F(x_k)\| \geq \varepsilon$ ($k \in K$) to know

$$\|F(x_k)\| > 0 \qquad (k \in K).$$

From Lemma 4.2, $H_i(B)$ from Assumption1.1 and Lemma 4.3 we know

$$\Pr ed_k \geq \frac{c^p\|J(x_k)F(x_k)\|^2}{2M} \geq \frac{c^p\varepsilon}{2M}.$$

Setting $a = \dfrac{\varepsilon}{2M}$,then it generate $\Pr ed_k \geq a.c^p$

and $\psi_{l(k)} > \psi_{k+1} + \eta \Pr ed_k$

$$\geq \psi_{k+1} + \eta a c^p$$

$$\geq \psi_{k+1} + \eta a c^{p_{l(K}}$$

$$\psi_{k+1} < \psi_{l(k)} - \eta a c^{p_{l(k)}}.$$

Above mentioned ,we know $\psi_{l(k)+1} < \psi_{l(l(k))} - \eta a c^{p_{l(k)}}$ and $p_{l(k)}$ is the biggest of P generated from Step 4.

$\because$ $0 < c < 1$

When $k \to \infty$, $p_{l(k)} \to \infty$, supposing $p_{l(k)} \geq 1, l(k) \to \infty$ $(k \to \infty)$and

and according to Algorithm1.1,we know $\tilde{d}_{l(k)}$ generated from sub- problem not accepted

$$\min \phi_{l(k)}(d) = \frac{1}{2}\left\| F(x_{l(k)}) + J(x_{l(k)})^T d \right\|^2$$

$$\text{s.t.} \quad \|d\| \leq \frac{c^{p_{l(k)-1}} \left\| F(x_{l(k)}) \right\|}{\left\| J(x_{(k)})^T F(x_{(k)}) \right\|}$$

We can prove it by reduction to absurdity . $\tilde{d}_{l(k)}$ is the solution of above-mentioned problem.Then we have

$$\frac{\psi_{l(l(k))} - \psi(x_{l(k)} + \tilde{d}_{l(k)})}{\psi(x_{l(k)}) - \phi_{l(k)}(\tilde{d}_{l(k)})} < \eta \qquad (1.11)$$

From Lemma1.2 ,we know

$$\Pr ed_k \geq \frac{c^{p_{l(k)-1}} \varepsilon^2}{2M} \psi_{k+1} < \psi_{l(k)} -$$

$$\eta \Pr ed_k < \psi_{l(k)} - \frac{\eta c^{p_{l(k)-1}} \varepsilon}{2M} \psi_{l(k)+1} <$$

$$\psi_{l(l(k))} - \frac{\eta c^p \varepsilon}{2M} Ared_k{}^p - \Pr ed_k d_k^p \leq \qquad (1.12)$$

$$\frac{1}{2}\left| \left\| F(x_{l(k)}) \right\|_2 - \left\| F(x_k) \right\|_2 \right| + O\left\| d_k^p \right\|^2$$

$$\left| \frac{\psi(x_{l(k)}) - \psi(x_{l(k)} + \tilde{d}_{l(k)})}{\psi(x_{l(k)}) - \phi_{l(k)}(\tilde{d}_{l(k)})} - 1 \right| \leq$$

$$\frac{\frac{1}{2}\left| \left\| F(x_{l(k)}) \right\| - \left\| F(x_k) \right\|_2 \right|}{\Pr ed_k(\tilde{d}_{l(k)})}$$

From lemma1.5,1.6 ,we know that $F(x_k)$ converges and $F(x_{l(k)})$ which is sub-sequence of $F(x_k)$ converges too.

Then (1.12)

$$= \frac{\frac{1}{2}\left| \left\| F(x_{l(k)}) \right\|_2 - \left\| F(x_k) \right\|_2 \right| + O(c^{2(p_{l(k)-1})})}{O(c^{p_{l(k)-1}})} \to 0$$

$$(k \in K, k \to \infty) \quad \frac{\psi(x_{l(k)}) - \psi(x_{l(k)} + \tilde{d}_{l(k)})}{\psi(x_{l(k)}) - \phi_{l(k)}(\tilde{d}_{l(k)})} \to 1$$

With Eq(1.9),this is a contradiction.

2 Numberical Example

Example2.1

Consider
$$\begin{cases} f_1(x_1, x_2, x_3) = 2x_1 - x_2^2 + 5x_3 = 0 \\ f_2(x_1, x_2, x_3) = x_1^2 + x_2^2 - 10x_3 - 60 = 0 \\ f_3(x_1, x_2, x_3) = x_1 + 2x_2 - x_3^2 - 4 = 0 \end{cases}$$

Table 1 The result of modified Newton method

The first point	times	result
(6,6,6)	4	(7.57,6.64,4.58)
(10,10,10)	8	(8.33,6.69,4.93)
(1,100,-100)	12	(8.53,5.46,3.96)

Table 2 The result of Nonmonotone adaptive trust-region method

The first point	times	result
(6,6,6)	2	(8.00,6.00,4.00)
(10,10,10)	2	(8.00,6.00,4.00)
(1,100,-100)	4	(6.94,1.23,-2.13)

References

[1] N.Y.Deng,Y.Xiao,F.J.Zhou, Nonmonotonic trust region algorithm ,Journal of Opimization Theory and Application 26(1993) 259-285

[2] L.Grippo,F.Lamparillo,S.Lucidi,A nonmonotone line search technique for Newton’s method,SIAM Journal of Numerical Analysis 23(1986)707-716

[3] Brezinski C.A classification of quasi-Newton methods [J].Numerical Algorithms,2003,33:123-135

[4] Yamashita N and Fukushima M.On the rate of conver-gence of the Levenberg-Marquardt method[J] Computing, 2001, 15:237-249

[5] Zhang JL and Wang Y.A new trust region method for nonlinear equations[J].Mathematical Mathods of operat-ions Research,2003,58:283-298

[6] Wenyu sun. Nonmonotone trust region method for sovling optim-Ization.Applied Mathmatics and comput-ation 156 (2004) 159-174

[7] M.E.Gilpin,Spiral chaos in a predator-prey model, Am.Na-turalist, 113(1997),306-308

[8] P.Korman,Dynamics of the Lotka-Volterra systems with diffusion,Appl.Anal.44(1992),191-207

[9] H.Amann,Dynamic theory of quasilinear parabolic equ-ations,reaction diffusion systems,Differential Integral Equations,3(1990),13-75

[10] Y.P.Wu,Travelling waves for a class of cross-diffusion systems with small parameters,J.Differential Equations, 123(1995),1-34

Parallel Numerical Simulation for the Multi-group Particle Transport Equations*

Jie Liu Lihua Chi Jing Chen

Section 605，College of Computer, National University of Defense Technology , Changsha, Hunan 410073, China
Email: liujie.nudt@163.com

Abstract

In the conditions of high temperature and high pressure, the unknown particle fluxes of transport equations were defined in energy, time, velocity phase-space and high dimensions geometry space. This paper presents an effective way to implement the scalable parallel numerical simulation on the clusters by combining the energy groups and the space domain decomposition. Based on the list schedule we first design a multi-group parallel method to solve the load unbalance problem. Then we present a parallel algorithm based on geometry domain decomposition. Experiments indicate that the algorithms get a better parallel efficiency. A parallel code combining those two algorithms was designed. Using the code, we solved a two dimension particle transport equations on a cluster, and the results show that the algorithms have well scalability. Parallel efficiency relative to 256 is 82% on 2048 processors.

Keywords: multi-group particle transport equations, unstructured grid, parallel computing, load balancing

1 Introduction

Due to the complexity of physical problems, the time-dependent particle transport equations are complex differential and integral equations involved a lot of physical variable. The transport equations can be solved by numerical simulation through a large number particles transport, and the distributions of particles can be determined on the geometry space, energy group, velocity phase-space and time space, then the whole movement behavior can be described. The large amount computation is the most important problem that impedes the practical applications of the time-dependent particle transport. It is urgent to solve the scalable parallel computing problems.

The discrete ordinate (Sn) method on the unstructured grids is the effective method to solve the equations. The standard iterative technique for solving discretized transport equation is source iteration, in which one alternates between solving for the local scattering source and inverting the global streaming-plus-collision operator. During organizing the parallel computation from the geometry space decomposition, the certain orders f the grids cause the data dependences, efficient organization of parallel algorithms for the sweep process is difficult.

In the past years, the scalable parallel computation of Sn method on the structured meshes has been made the big progress. On the orthogonal hexahedral meshes, the KBA sweep algorithm [1][2][3] has been developed by the Los Alamos national laboratory, and the algorithm can be scaled to thousands of processors to solve the millions of meshes. Based on the KBA sweep algorithm, the kernel Benchmark code called SWEEP3D[4] and the PARTISN code [5] are developed, and the parallel efficiency of the PARTISN code is 80% on 3000 processors.

Compared with the structured meshes, the scalable parallel computation of Sn method on the unstructured meshes is much more difficult. On the three dimensions Cartesian coordinate, a parallel pipelining Sn sweep algorithm [6] is presented to solve the radiation transport calculations, and for the largest problem the

* This research work is supported by the National Natural Science Foundation of China under grant No.60673150.

algorithm has over 80% efficient on 256 processors. Recently, Pautz developed a new algorithm[7] for performing parallel Sn sweeps on the unstructured meshes, and the algorithm uses a low-complexity list ordering heuristic to determine a sweep ordering on any portioned mesh. For the typical problems and the with normal mesh partitionings, parallel efficiencies are 50% on 126 processors. Mo Zeyao presents a parallel flex sweep algorithm [8][9][10] for the neutron transport on unstructured grid, for the two different scale applications , the parallel solver has respectively achieved speedup larger than 72 using 92 processors and 78 using 256 processors on two different parallel computers.

The above research jobs show that the parallel algorithms on the unstructured meshes have some similarities and implement the modest levels of parallelism. Especially, due to the strong dependence of the meshes, the current algorithms are difficult to scale the more processors, which rely on the communicating delay, sorting algorithms and inserting algorithms[11].

The purpose of this paper is to develop implement the scalable parallel numerical simulation of multi-group particle time-dependent transport equations on the unstructured meshes by combining the energy group and the space domain decompositions. Based on the list schedule we first design a multi-group parallel method to solve the load unbalance problem. Then we present a parallel algorithm based on geometry domain decomposition. Experiments indicate that the algorithms get a better parallel efficiency. A parallel code combining those two algorithms was designed. Using the code, we solved a two dimension particle transport equations on a cluster, performance results show that the algorithms have well scalability. Parallel efficiency relative to 256 is 52% on 2048 processors.

The rest of the paper is organized as follows. In sec.2, we discuss the particle transport equations and the numerical method. In sec.3, we develop a multi-group parallel method to solve the load unbalance problem. In sec.4, we present an algorithm for parallel sweeps on unstructured meshes. We give the computational results in sec.5. Finally, in sec.6 we make some conclusions and recommendations for the future work.

2 Multi-Group Time-Dependent Tarticle Transport Equations

Under 2-D cylindrical Lagrange coordinates, multi-group time-dependent particle transport equation can be defined by

$$\frac{1}{\mathrm{V_g}}\frac{\partial\varphi_{\mathrm{g}}}{\partial \mathrm{t}}+\frac{\mu}{\mathrm{r}}\frac{\partial(\mathrm{r}\varphi_g)}{\partial r}+\xi\frac{\partial\varphi_g}{\partial x}-\frac{1}{r}\frac{\partial(\zeta\varphi_g)}{\partial\omega}+\sigma_g^{tr}\varphi_g=\mathrm{Q}_{fg}+\mathrm{Q}_{sg} \tag{1}$$

where $\varphi_{\mathrm{g}}=\varphi_g(x,r,\xi,w,t)$ represent the particle flux of gth group along velocity angular direction (ξ,ω) at time t and at location (x,r). Q_{fg} and Q_{sg} represents the fission source and the scatter source. G represents the total number of particle groups. Ω_{xr} represents the geometry domain.

Equation (1) is defined in energy, time, velocity phase-space and high dimensions geometry space. And equation (1) has been written with the form of independent groups. For the time, equation (1) is discretized by the implicit difference format. For the velocity phase-space, equation is discretized by discrete ordinate (Sn) method. And for the geometry space, equation is discretized by discontinuous finite element of arbitrary triangles and quadrangles.

3 A Multi-Groups Parallel Method for the Load Unbalance

The computation of fusion source Q_{fg} is composed of the angle fluxes for all energy groups, and the computation of each Q_{fg} is equal. On the other hand, scatter source Q_{sg} is acumulated from No.1 to No.g group, which means the computational amount of each Q_{sg} is increased by linear degrees to No.g group. Due to the above two sides, when we design the parallel algorithm according to the energy group decomposition, we ensure that each processor has not only the almost equal number of energy groups, but also the sum of the ranks of energy groups.

Suppose the total number of energy groups is G, the rank of each energy group is from 1 to G, the number of processors is P and the rank of each processor is from 0 to $P-1$. According to the multiple and non-multiple relation of G and P, we design a multi-group parallel load balance method based on the list schedule, which divides into three conditions.

(1) when $G=(2k+2)P$, $k=0,1,2,\ldots$, each processor has the $2k+2$ number of energy groups. From the No.1 of energy group, first we assign one energy group to each processor according to the increscent rank number of processors. Then we assign one energy group to each processor according to the decreasing rank number of processors. And so on. Table 1 gives the sketch map. In this case, each processor has not only the equal number of energy groups, but also the sum of the ranks of energy groups, and the loads of processors are balance.

Table 1 The list schedule method when $G=(2k+2)P$, $k=0,1,2,\ldots$

No. of processors	No. of energy groups				
0	1	2P	2P+1	…	(2K+2)P
1	2	2P-1	2P+2	…	(2K+2)P-1
2	3	2P-2	2P+3	…	(2K+2)P-2
…	…	…	…	…	…
P-1	P	P+1	3P	…	(2K+1)P+1

(2) when $G=(2k+1)P$, $k=0,1,2,\ldots$, each processor has the $2k+1$ number of energy groups. From the No.1 to the No.P of energy group, we assign one energy group to even number processors and one of energy group to odd number processors. And From the No.$(P+1)$ to the No.$2P$ of energy group, we assign one energy group to odd number processors and one of energy group to even number processors. Then from the No.$(2P+1)$ to the No.$3P$ of energy group, we assign one energy group to each processor according to contradictory order. Finally, for the residual energy groups, we assign them to processors used method (1). Tables 2 and 3 give the sketch map. The expression of energy groups is different to the parity of P. In this case, each processor has the equal number of energy groups, and one rank difference between the sums of the ranks of energy groups in each processor, which the loads of processors are almost balance.

Table 2 The list schedule method when $G=(2k+1)P$, $k=0,1,2,\ldots$ and P is odd

No. of procesors	No. of energy groups				
0	1	(3P-1)/2+1	3P	…	(2K+1)P
1	(P-1)/2+2	P+1	3P-1	…	(2K+1)P-1
2	2	(3P-1)/2+2	3P-2	…	(2K+1)P-2
3	(P-1)/2+3	P+2	3P-3		(2K+1)P-3
…	…	…	…	…	…
P-1	(P-1)/2+1	2P	2P+1	…	2KP+1

Table 3 The list schedule method when $G=(2k+1)P$, $k=0,1,2,\ldots$ and P is even

No. of processors	No. of energy groups				
0	1	3P/2+1	3P	…	(2K+1)P
1	P/2+1	P+1	3P-1	…	(2K+1)P-1
2	2	3P/2+2	3P-2	…	(2K+1)P-2
3	P/2+2	P+2	3P-3		(2K+1)P-3
…	…	…	…	…	…
P-1	P	2P	2P+1	…	2KP+1

(3) when $G=nP+r$, $k=0,1,2,\ldots$, and $r \geq 1$, the number of energy groups in each processor are not equal because G is not integer multiple of P. In order to implement load balance, we also divide the energy groups into two parts. For the first part energy groups from the No.1 to the No.r, we assign one energy group in contradictory order to the processor with the rank from 0 to $r-1$. For the residual nP pieces of energy groups, we assign them to processors used method (1) and (2) due to the integer multiple relations between G and P. Table 4 gives the sketch map. ,In this case, there is only 1 group difference between the number of energy groups in each processor, and min(p-r-1, r) ranks difference between the sums of the ranks of energy groups in each processor.

Table 4 The list schedule method when $G=nP+r$, $k=0,1,2,\ldots$, and $r \geq 1$

No. of processors	No. of energy groups				
0	r	r+1	r+2P	…	r +nP
1	r-1	r+2	r+2P -1	…	r +nP –1
2	r-2	r+3	r+2P -2	…	r +nP -2
…	…	…	…	…	…
r-1	1	2r+1	2P+1	…	nP+1
…	…	…	…	…	…
P-1		r+P	r+P+1	…	r+ (n-1)P+1

4 SN Parallel Algorithm Based on Geometry Domain Decomposition

Based on the space meshes, we usually calculate the problem by the method of "sweeping". During a sweep, it is locally solved for each spatial cell in the mesh in a specified order for a single direction in the discrete ordinates set. This order is constrained by the interaction of the discrete ordinates set with the spatial mesh. Because of the constraints placed on the sweep ordering, it is difficult to solved in parallel. Tasks (cell-angle pairs) assigned to a processor cannot begin until cells upstream from them have been solved; if their neighbor meshes at the geometry boundary are assigned to other processors, this processor must waiting for communication.

We present an algorithm to solve 2-D neutron transport equation in parallel on the space domain decompositions. We organize the parallel computation according to the arranged order, enlarging the computation granularity and reducing the communication time, increasing the dependence on the delay of communication. All these can bring a good efficient algorithm , and can gain better parallel computing performance combining with the load balance algorithm. We apply the partition software Chaco to decompose the system domain into subdomains distributed to different processors. By applying the software Chaco we can distribute the meshes on unstructured grid automatically with small communicating surface. After finishing the domain decomposition, the keys of the parallel algorithm are: (1) how to design the priorities of the tasks to gain a efficient sorting algorithm; (2) how to design the mode of the communication in the parallel computing to gain a good ratio of computing and communication.

4.1 Prioritization Algorithm

The determining factor in the performance of the parallel algorithm based on geometry domain decomposition is the assignment of priorities to meshes. The dependencies between the unstructured meshes are more complicated the structured meshes, just because there is not obvious what a "columnar" decomposition like KBA method on an unstructured mesh is or what the corresponding ordering should be. The purpose of this chapter is to develop a good sweep ordering method before the parallel computation of the meshes.

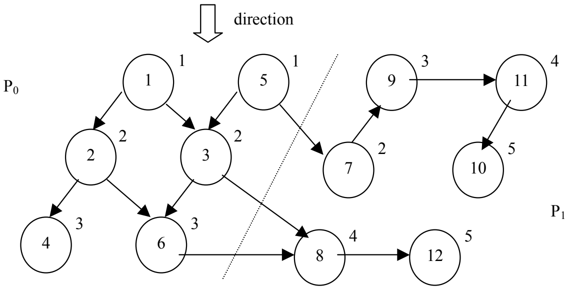


Figure 1 Priorities of a single direction

(the number of meshes given in the circles, the priorities given on the top right corner of the circles)

The prioritization algorithm is presented as follows:

First for each direction, the sequence calculating orders of all meshes is determined by the data dependencies between meshes. We can calculate the meshes that are belonged to the domain boundary and have no dependencies from any meshes, and we define the priority of those meshes to 1. Then we can calculate the meshes that are the direct down meshes of the meshes with the priority 1, and we define the priority of those meshes to 2. And so on. According to the calculating orders, we assign to each mesh a priority equal to the highest priority of its father meshes plus 1. The value is smaller, and the priority is higher. Figure 1 gives the priorities of a single direction, in the Figure 1 the No.8 mesh's father meshes are No.3 and No.6 meshes. The priorities of No.3 and No.6 meshes are 2 and 3 respectively, so the priority of No.8 mesh is 4.

Then we assign the priorities for all directions and all meshes. When there are no dependent between directions, the priorities are same to the single direction. If there are dependent between directions, we must assign the priorities over again. We take a single mesh and a single direction as a task. According to the dependencies between directions and meshes, after determining the father and son tasks, we assign to each task a priority equal to the highest priority of its father task plus 1. Also the value is smaller, and the priority is higher. Figure 2 gives the priorities of the tasks.

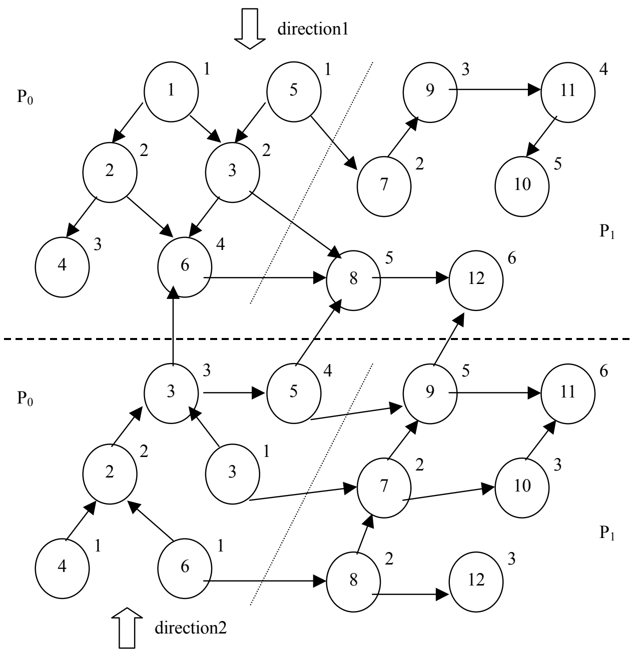


Figure 2 Priorities of all directions and meshes

We can calculate the meshes that are belonged to the domain boundary and have no dependencies from any meshes. The priority of a mesh can be assigned by its father meshes' priorities. We first calculate the meshes that have the higher priorities. The meshes that have the same priorities can be calculated simultaneously, and the meshes that have lower priorities must wait until their father meshes have been calculated. The meshes on the decomposition border need data communication. In this paper, we assign the priorities to all meshes and all directions, so all the directions can be calculated simultaneously. Compared with the old algorithm, the new parallel algorithm that exploits the simultaneity used the priorities avoids to manage the queues of meshes and use the time-consuming inserting algorithm, and has better parallel performance.

4.2 SN parallel algorithm

After arranging all meshes in the order of the priority algorithm, we will develop the communication and computation methods to exploit the parallelism. For the meshes with the same priority belong to the same processor, we should calculate the meshes at the boundary at first for two reasons. First, if some processors are waiting for the communication of the meshes at the boundary, it can get the data to calculate as soon as possible to reduce the waiting time, At the other hand, when a processor finished the calculation of the meshes at the boundary, it can begin to send out the data and calculate the other meshes with the same priority in the same processor, which can overlap the communication and calculation. All these can increase the parallel efficiency.

It is unavoidable to communicate between processors when we solve the particle transport equations in parallel. But the large amount of communications will decrease the parallel efficiency. So when we design the parallel algorithm, we hope the communication cost can be reduced. So designing a good communication mode is one of the important ways to increase the parallel efficiency. It is necessary to transmit data between the neighbor processors. If we determine the way of geometry domain decomposition, the data amount of communication is fixed. Then we can pack the messages with the same priority into a long message in order to reduce the time of communication. Basing on the description above, we describe the algorithm on geometry domain decomposition in Figure 3.

According to the relation of meshes, determine the serial calculating sequence of all meshes for all directions.

Evaluate the parallel calculating priority of all meshes for all directions.

According to the priority of each element, calculate in order:

```
Do i=1,ipromax (the largest priority)
    Calculate the meshes that need to send out data and need not
    receive data from other meshes;
    Receive the data from meshes at the partition boundary;
    Calculate the meshes that need to receive and send out data;
    Send out the data at the partition boundary;
    Calculate the rest meshes with the priority i ;
Enddo
```

Figure 3 Sn parallel algorithm on geometry domain decomposition

5 Computational results

We have implemented our sweep scheduling algorithm in a new parallel Sn code using Message Passing Interface (MPI), and conducted our timing studies on a Xeon cluster. The Xeon cluster contains 2048 processors and uses the InfiniBand interconnected network with latency equal to 5~6 microseconds.

Table 5 lists the parallel executing time for 150 time steps, speedup and efficiency for the multi-groups parallel method for the load unbalance problem. The test problem is defined on a ball constituted by 225 meshes, S4 (16 angular directions), and 24 groups. The data show that the list scheduling method has the linear speedups and high parallel efficiencies. The number of energy groups limits the number of processors that may be used. So we must combine the multi-groups parallel method with the Sn domain decomposition parallel method to implement the scale parallel computation.

Table 5 Executing results for the multi-groups parallel method

No. of processors	Executing Time (s)	Speedup	Parallel efficiency
1	1584.1	1.00	100%
2	768.6	2.06	103%
4	390.7	4.05	101%
8	195.2	8.11	101%
12	138.0	11.48	96%

In order to test the performance of the Sn domain decomposition parallel method, we test the problem define on a ball constituted by 900 meshes, S4 (16 angular directions), and 24 groups on the same Xeon cluster. Table 6 list the parallel executing time for 150 time steps, speedup and efficiency. The test time shows that the speedup of 128 processors is only 40.06 and the parallel efficiency is only 31.3%.

Table 6 Executing results for the Sn parallel algorithm

No. of processors	Executing Time (s)	Speedup	Parallel efficiency
1	4836.2	1.00	100%
2	2253.5	2.14	107%
4	1097.6	4.40	110%
8	669.6	7.22	90.3%
16	351.5	13.76	86.0%
32	233.5	20.71	64.7%
64	177.8	27.20	42.5%
128	120.7	40.06	31.3%

Figs.4 and 5 respectively gives the testing time and parallel efficiency for the parallel algorithm combining the energy group and the geometry domain decomposition in difference size of the meshes on the same Xeon cluster. The problem is defined on a ball constituted by 225, 900, 3600, 14400, and 57600 meshes, S4 (16 angular directions), and 24 groups. The testing result in Figs.4 and 5 show that the parallel algorithm combining the energy group and the geometry domain decomposition has well scalability and the parallel efficiency relative to 256 is 82% on 2048 processors.

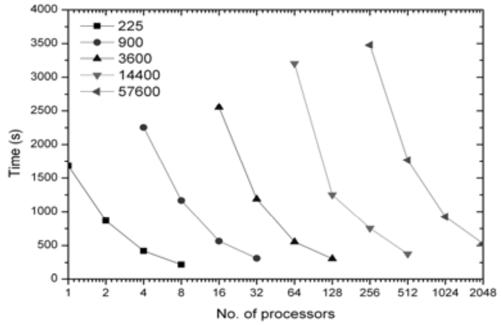


Figure 4 Time of parallel algorithm combining group and domain decomposition

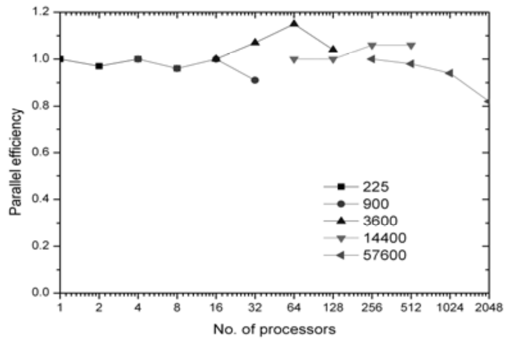


Figure 5 Parallel efficiency of parallel algorithm combining group and domain decomposition

6 Conclusion

This paper presents an effective way to implement the scalable parallel numerical simulation on the clusters by combining the energy groups and the space domain decomposition. Based on the list schedule we first design a multi-group parallel method to solve the load unbalance problem. Then we present a parallel algorithm based on geometry domain decomposition. Experiments indicate that the algorithms get a better parallel efficiency. A parallel code combining those two algorithms was designed. Using the code, we solved a two dimension particle transport equations on a cluster, performance results show the algorithms have well scalability. Parallel efficiency relative to 256 is 82% on 2048 processors.

There are a number of directions for future research in this area. The method utilizing asynchronous message passing to exchange the ghost data can increase scaling in many clusters. Second, the low complexity prioritization method need be developed. Finally, there is a continuing need for the development of more effective spatial decompositions.

References

[1] K.R.Koch, T.S.Baker, and T. E. Alcouffe, A Parallel Algorithm for 3D Sn Trnsport Sweeps, LA-CP-92-406, Los Alamos National Laboratory, 1992

[2] R. S. Baker and K. R. Koch. An Sn Algorithm for the Massively Parallel CM-200 Computer [J]. Nuclear Science Engineering, 1998, 128: 310-320

[3] R. S. Baker and K. R. Koch, An Sn Algorithm for the Massively Parallel CM-200 Computer, SuperComputing 2000, Dallas, Texas, Novemaber 4-10, 2000

[4] http://www.llnl.gov/asci_benchmarks/asci/limited/sweep3d/asci_sweep3d.html

[5] PARTISN, http://www.ccs.lanl.gov/CCS/CCS-4/code.shtml

[6] Plimpton S. , Hendrickson B. , Burns S. , McLendon W. . Parallel algorithms for radiation transprt on unstructured grids. In : Proceedings of SuperComputing’2000 , Dallas , Texas , 2000

[7] Shawn D. Pautz. An Algorithm for Parallel Sn Sweeps on Unstructured Meshes. Nuclear Science and Engineering.2002,140(2):111-136

[8] Mo Zeyao, Fu Lianxiang, Yang Shulin. Parallel pipelined Sn sweeping algorithm for neutron transport on unstructured grid. Chinese Journal of Computers, 2004, 27(5): 587-595

[9] Mo Zeyao, Parallel Flux Sweep Algorithm for Neutron Transport on Unstructured Grid, The Journal of Supercomputing, 2004, 30(1): 5-17

[10] Mo Zeyao, Zhang Aiqing, Cao Xiaolin. Towards a Parallel Framework of Grid-based Numerical Algorithms on DAGs[C]. 20th IEEE International Parallel and Distributed Processing Symposium. Rhodes Island, Greece, 2006

[11] V.S. Anil Kumar, et al. Provable Algorithms for Parallel Sweep Scheduling on Unstructured Meshes[C]. 19th IEEE International Parallel and Distributed Processing Symposium. Denver, CO, USA, 2005

The Research of Mobile Supported Synergistic Learning

Yi Jin Zhuying Lin

School of Mathematics and Computer Science, Guizhou Normal University, Guizhou 550001,P. R. China

Email: genekim@126.com,gylinying@126.com

Abstract

M-Learning is called the future of distance education and Synergistic Learning is viewed as a new framework of learning technology system. Based on M-Learning and synergistic learning, Mobile Supported Synergistic Learning (MSSL) came into being as a new research field and MSSL can make both of them supported each other. As a new mode of distance education, MSSL is introduced here，which includes two origins theories. Its definition and its fundamental frame are discussed further and its features are explained all in this paper.

Keywords：M-Learning; Distance education; Synergistic Learning; Mobile Supported Synergistic Learning

1 Introduction

With the developments of the computer network technology, the mobile telecommunication technology, multimedia and modern education thoughts, M-Learning appears as a brand-new assistant learning method by using mobile communication terminal instruments. Essentially M-Learning is the mobilization of learning terminal instruments and it breaks through the choke point of the traditional long-distance learning and e-learning, expanding the multi-media distance education range on the network significantly. Although M-Learning has many advantages, such as convenient, flexible, [1]etc, it is hard for M-Learning to overcome the natural shortcoming of its mobile environment, such as aprosexia and loneliness. So it is necessary to get some help from a better framework of learning technology system.

Synergistic Learning is a new framework of learning technology system, one of its functions is to support the educational mode by using the advanced technologies. Synergistic Learning is a new and useful development based on the traditional learning theories. Synergistic Learning is a new framework, which can adapt the social structure of current times, and content the needs of social developments and education innovations[1]. The five sub-fields of synergistic Learning can eliminate aprosexia, loneliness during the learning process, and promote the efficiency of learning. So it is the right one that M-Learning needs.

As MSSL is a complete new mode of distance education bases on M-Learning and Synergistic learning, MSSL can offset the negative influences during the pure M-Learning process, and can content with the technical needs of Synergistic learning. At the same time MSSL can strengthen the synergistic relationships among the learners and the interaction between the each field of Synergistic learning. The combination of M-Learning and Synergistic learning makes both of them supplement each other accordingly, enhance the efficiency of learning. So it is necessary to study some elementary problems in MSSL research field.

In this paper，Mobile Supported Synergistic Learning(MSSL)，is introduced as a new mode of distance education. Firstly, the basic information of M-Learning and Synergistic learning, both of them are the base of MSSL, is introduced. Next, its definition and its fundamental frame are discussed in detail, and finally, the features of MSSL are explained.

[1] Corresponding author.
Email addresses: genekim@126.com
Cell phone: 0086-13639088977(Yi JIN)

2 Fundamental concept of M-Learning

2.1 The definition of M-Learning

In his book *The future of learning：From E-Learning to M-Learning*, Doctor. Desmond Keenan indicated that the development of M-Learning will make students more free in D-Learning (distance learning). No matter what they are in airport or anywhere, they can learn if they want. Beyond doubt, the next generation of D-Learning will be M-Learning.

At the present, there is no uniform definition about M-Learning. Analyzed are the definition of M-Learning in literature[2] by Alexander Dye and the viewpoint in literature[3] by Clark Quinn. In this paper, the authors present the definition of M-Learning as follows. M-Learning is a kind of learning that learners can study anytime and anywhere with the mobile terminal instruments. The mobile terminal instruments must present the information which learners request efficiently, and must provide the interactive communications among instructors and learners. (The mobile terminal instruments include various smart cell telephones, PDA, and laptops with wireless function).

2.2 The structure and features of M-Learning

2.2.1 The structure of M-Learning

There are four fundamental parts in M-Learning including Internet, Communication networks, Mobile terminal instruments and Education resource Web servers. The structure of M-Learning is in Figure 1:

1) Internet: Abundant education resource can be gained though Internet, and Internet is the main channel through which Communication networks and Mobile terminal instruments can be connect with Education resource web servers.

2) Communication networks: it is the mid-networks between the Mobile terminal instruments and Internet to realize wireless connecting. Communication networks realize the real time instruments connecting and information exchanging.

3) Mobile terminal instruments: it is the characteristic part of M-Learning including wireless and portable digital communication instruments, such as smart phones, PDAs and laptops, etc.

4) Education resource Web servers: the main part to store and transmit education resource, and be connected with Internet.

2.2.2 The features of M-Learning

There are four fundamental elements in M-Learning, learners, instructors, teach contents and the teaching methods. All of them have the same feature: mobility. Compares with the traditional learning methods, M-Learning has following features:

1) Mobility. As long as within the areas covering the mobile telecommunication network services, the learners can study anytime and anyplace. By the same token, the instructors can give their teaching information anytime and anyplace, and also can revise, renew the teaching resource database anytime and anyplace.

2) Real time. If the learners have the needs to get some knowledge, by using some technical methods, the learners can get the knowledge they need at once. So M-Learning is a real time learning method.

3) Interactive. By using mobile terminal instruments and the services of mobile communication, both the learners and the instructors can communicate each other real time, So M-Learning is very interactive.

4) Virtualization. Through M-Learning, the instructors can create a virtual classroom, virtual instructors, and the learners can create a virtual class. The relationship between the instructors and the learners are dynamic and virtual.

5) Digitization. The teaching resource of digital multimedia, the platform of network and the mobile terminal instruments combined determine the digitization of M-Learning.

6) Individuation. M-Learning can provide the individual services according to learners' needs and the features of the subjects.

7) Universalization. With the appearance and the wide spread of advanced mobile terminal instruments, any user who has those instruments can become one member of M-Learning, even they are far away from the

classrooms and instructors.

2.3 The Shortcoming of M-Learning

Although M-Learning has many advantages, such as convenient, flexible and individual, considering that the learning process is individual and emotional, it is hard for M-Learning to overcome the natural shortcoming of its mobile environment, such as aprosexia and loneliness. There are lots of distractive problems during M-Learning, such as terrible learning environment, low quality digital information, complex interface of learning, etc. all of those induce the learners' aprosexia, fidget and frustration, and reduce the efficiency of M-Learning.

To compensate the shortcoming of M-Learning, it is necessary to get some help from the framework of learning technology system maximizing the advantages of mobility, and minimizing the disadvantages of mobility.

Synergistic Learning is just the right framework of learning technology system. The five sub-fields of synergistic Learning can eliminate learners' aprosexia and loneliness efficiently during the learning process, and promote the efficiency of learning. So it is the one that M-Learning really needs.

3 Synergistic Learning

3.1 Base of Synergistic Learning

Synergistic Learning is a new framework of learning technology system, and one of its functions is to support the educational mode by using the advanced technologies. Synergistic Learning is a new and useful development based on the traditional learning theories. As a new framework, it can adopt the social structure of current times, and meet the needs of social developments and education innovations[1].

Synergistic Learning is different from Collaborative Learning and Cooperative Learning. The main difference is that Synergistic Learning is based on the Synergistic theory and knowledge management theory. Synergistic Learning emphasizes particularly on the interactions and cooperation of all elements, and it is self-organizing. Collaborative Learning focuses on sharing some resources and productions among every individual. Cooperative Learning is inclined to distribute the work to each individual, and then accomplish every distributed part by each individual, during the whole period there is few communication among the individuals.

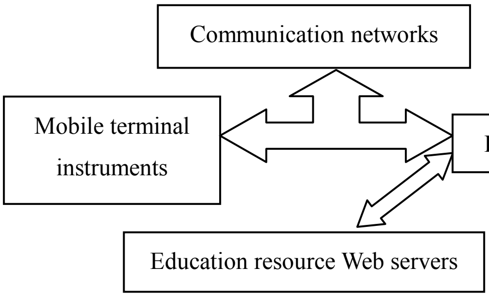


Figure1 Structure of M-Learning

3.2 Framework of Synergistic Learning Field

There are five essentials in the framework of Synergistic Learning including Knowledge, Information, Emotion, Action and Value.

In the framework of Synergistic Learning, for describing the learning activity, Field, one concept of physics, is introduced. The Synergistic Learning Field (SLF) can be comprehended as the space in which contains all essentials of Synergistic Learning.

According those five essentials of Synergistic Learning, there are five sub-fields in Synergistic Learning Field, including Knowledge Field (KF), Information Field (IF), Emotion Field (EF), Action Field (AF) and Value Field (VF). The structure of Synergistic Learning Field is displayed in Figure 2:

With the interactions of every sub-field in learning field, Synergistic Learning has the synergistic manufacture and the knowledge synergistic structure, and then realizes the re-construction of information and the development of the knowledge, finally makes the learners' promotion.

In the process of the entire Synergistic Learning, for the excellent interaction, the learners' attention can be attracted. The efficiency of learning can be guaranteed and promoted.

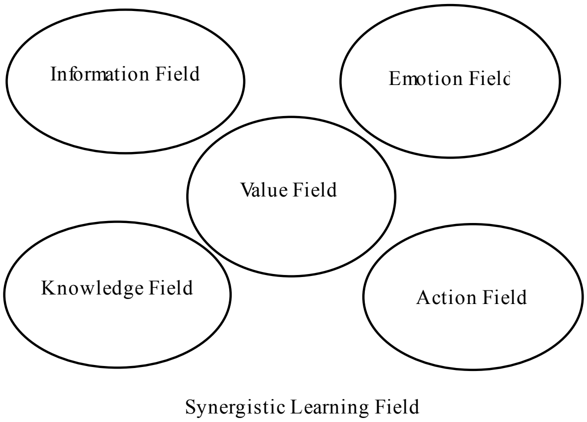


Figure 2 Structure of Synergistic Learning Field

3.3 Technical Needs of Synergistic Learning

The technical needs of Synergistic Learning are described below:

- Deep interaction between learners and contents
- Information aggregation
- Collective thinking operation
- Coordinating between multiple fields
- Collective creating knowledge artifacts

Without the support of modern technologies, as a new framework of learning technical system, Synergistic Learning can not be implemented. The advantages of the technologies of M-Learning are appropriate for that.

The shortcoming of M-Learning and the request of Synergistic Learning give the combination possibility of both, and make both of them supplement each other accordingly, enhance the efficiency of learning. So it is necessary to create a new mode of distance education of MSSL and study some elementary problems in this research field.

4 Mobile Supported Synergistic Learning (MSSL)

4.1 The Definition and Framework of Mobile Supported Synergistic Learning

Mobile Supported Synergistic Learning (MSSL) is a new mode of distance education, the precondition of MSSL is the advanced mobile communication technology. In M-Learning Environment of MSSL the individual information and collectivity information can be expressed and delivered at once, and then the information is aggregated, created and delivered among the individuals and collectivities in the Synergistic Learning field. As a result the final knowledge can be learned by more learners and the efficiency of learning can be promoted.

The author here illustrates the framework of Mobile Supported Synergistic Learning in Figure 3.

MSSL can offset the negative influences during the pure M-Learning process, and can content with the technical needs of Synergistic learning. At the same time the synergistic relationships among the learners and the interaction among the each field of Synergistic learning can be strengthened in MSSL. The combination of M-Learning and Synergistic learning makes both supplement each other, and enhances the efficiency of learning.

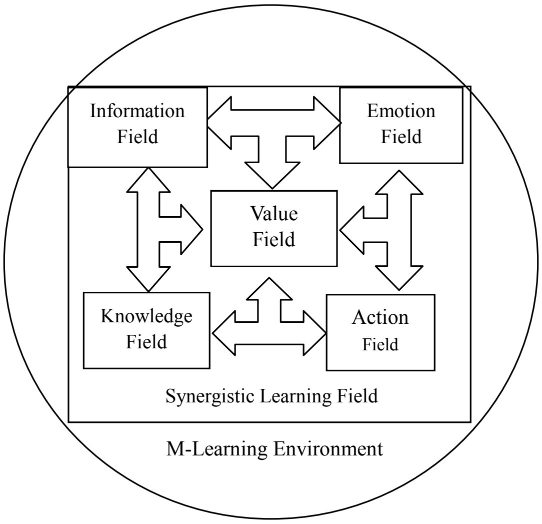


Figure 3 Framework of Mobile Support Synergistic Learning

4.2 Features of Mobile Supported Synergistic Learning

In this section, the authors will introduce the features of Mobile Supported Synergistic Learning.

1) deep interaction between the learners and contents, the individual and individual, individual and collectivity, the collectivity and collectivity, maximal the efficiency of learning.

2) provide the Synergistic fields and space for information and knowledge choosing, aggregating, transforming and delivering. And offset the shortcoming of current learning technical system.

3) append the operation of collective thinking. It is helpful for the information and knowledge collection, aggregation, transformation and delivery. And then form the collective information and knowledge, feed back the individuals and other collectivities.

4) complement the mechanism of information and knowledge collection, aggregation in current learning technical system, MSSL is one new mode of distance education based on relative learning technical system.

5) optimize the structure of learners, instructor, contents and multimedia in education. And supply the tools and methods for the information and knowledge aggregation, transformation and management.

6) coordinate between multiple fields of the Synergistic Learning, and pay more attention to the characteristics of individual.

7) overcome the natural shortcoming of M-Learning, promote the efficiency of learning, and enlarge the scale of modern education efficiently.

5 Conclusions and Future Work

Mobile Supported Synergistic Learning is a complete new mode of distance education, bases on M-Learning and Synergistic learning. MSSL can offset the negative influences during the pure M-Learning process, and can content with the technical needs of Synergistic learning. At the same time MSSL can strengthen the synergistic relationships among the learners and deep the interaction between the each field of Synergistic learning. MSSL can provide the Synergistic fields and space for knowledge choosing, aggregating, transforming and delivering. And append the operation of collective thinking, pay more attention to the characteristics of individual, promote the efficiency of learning, extend the network multi-media distance education range extremely. and enlarge the scale of modern education efficiently.

Indubitably Mobile Supported Synergistic Learning will be an important new learning mode of distance education. And MSSL is significant for modern education with its virtue. The authors believe that Mobile Supported Synergistic Learning will be one of hotspots of modern education in the future. The future work of Mobile Supported Synergistic Learning, as a complete new mode of distance education, should be done on the issues of perfecting the definition and features of MSSL. Another interesting work of Mobile Supported Synergistic Learning is to elaborate the implement in the proposed MSSL definition and framework.

References

[1] Zhi-ting Zhu, You-mie Wang. Synergistic Learning: the frame of learning technology system facing the knowledge times. China Educational Technology. 2006. pp.56－62

[2] Alexander Dye，Mobile Education -- A Glance at The Future [EB/OL]. http://www.nettskolen.com /forskning/ mobile_education. PDF,2001

[3] Clark Quinn, mlearning: Mobile, Wireless, In-Your-Pocket Learning[EB/OL],http://www.linezine.com/2.1/features/cqmmwiyp.htm

[4] Desmond Keenan, The future of learning：From E-Learning to M-Learning [EB/OL]. Http://www.open.edu.cn/ycjy/zx/200306/xzjd/,2003

[5] Donna Abernethy：Get ready for mlearning[EB/OL], http://www.learningcircuits.org

[6] Kaufman, C. (2003). Creating Synergistic Learning Environments for School Administrators. In G. Richards (Ed.), Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education

2003 (pp. 2007-2010). Chesapeake, VA: AACE

[7] Jeremy S Kossen, When eLearning becomes mLearning [EB/OL]http://www.palmpowerenterprises.com/issues/issue 200106/ elearning001.html

[8] Lehner, F, H. Nosekabel, et al. Wireless E-learning and Communication Environment 一 WELCOME at the University of Regensburg, Workshop on M-Services, 2002, Lyon, France

[9] Flexible Learning: Mobile Learning Objects, 2002, http://www.knowledgeanywhere.com

[10] Roibas A.C., Sanchez LA: Design scenarios for m-learning, Proceedings of the European Workshop on Mobile and Contextual Learning (p.53-56), Birmingham, UK, June 2002

[11] M-Learning Forum Meeting [DB/OL] http://www.pjb.co.uk/m-learning/}.html

[12] Truls Fagerberg, Torstein Rekkedal and John Russell. Designing and Trying Out a Learning Environment for Mobile Learners and Teachers [DB/OL], http://www.nettskolen.com/pub/artikkel.xsql?amid=115

The Design for the Architecture of Intelligent Mix Meta Search Engine

Yong Zhang Longbin Xiao Min Wu

School of Computer and Communication, Lanzhou University of Technology, Lanzhou, Gansu, 730050

Email：xlb11111@163.com

Abstract

Search engine is the most effective tool for web information retrieval. This article analyses the current situations of search engine firstly. Then, it proposes a model of an intelligent mix meta search engine. In this system, personal search serves are provided by analyzing the user logs. Furthermore, an intelligent dispatch strategy of the independent search engines is proposed to resolve the problem that the different independent search engine has a different performance to the same user query. At the same time, an improved PageRank algorithm, which is called A-PageRank, is proposed. The algorithm distributes the PageRank value of the source page based on the topic similarity.

Keywords：A-PageRank; mix meta search engine; user interest model; primary feature term; dispatch strategy

1 Introduction

With the development of web information in exponential rate, the information exploding era has come into being. John Naisbitt, a very famous futurologist, has said that people were living in the information ocean but they were hoping for the knowledge. How to locate the required information rapidly and accurately in the internet which contains so much information is a focused question which people are very interesting in all along. From the twentieth web development statistics of China which was published in eighteen July, 2007 by CNNIC, a fact has been exposed that the proportion of using SE (search engine) service is about 74.8 %, exceeding the proportion of using e-mail service which is 55.4 %, however, the developing space of SE service is still spacious because the proportion of using SE service in America has already come to91%[1]. The SE service has become one of the major approaches to obtain information for people. There is a group of considerable datum which comes from the investigation report published by Roper Starch in 2001. From this report, a fact has been known that there are 36% of web users spend 2 hours to look for the required information by using SE service per week, 71% of users encounter problems in using this service and the average time which is spent on finding the search task is failed is 123 minutes[2]. From the statistical data, a conclusion can be obtained that although occurrence of search engine technology satisfies the people need to some extent; there are still many aspects to be improved. Nowadays, most popular SEs(for example Google, Baidu, etc.) are ones that based on robot which always are called independent search engine or external search engine. This kind of SE has little intelligence, say nothing of the ability to study by itself. So, the different users will obtain the same results list to the same query. Furthermore, this kind of SE has the topic sensitivity. It means that one independent search engine has good performance in some topics but poor performance in other ones. Some researchers have done a statistical investigation based on the results set which contains 10316 different pages returned by the Yahoo, Google and Ask Jeeves. The research shows that, for each search engine, the recurrence rate in the first page of the results list is about 3.2%[3]. This experimental data proves the fact that the difference of results lists returned by the different SEs to the same query is very large. Although the traditional meta search engine can reduce the topic sensitivity of independent search engine efficiently, the massive data must be transferred between it and the independent search engines in the process of usage. In order to get better precision and recall rate, an intelligent mix meta search engine (IMMSE) based on the traditional meta search engine is proposed in the article.

The organization of this paper is as follows. In section 2, the architecture and functions of the main modules of IMMSE is introduced briefly. In section 3, some key technologies, for example, user interests mining, the dispatch strategy of independent SEs and A-PageRank algorithm etc. are demonstrated in detail. Finally, the conclusions are given in section 4.

2 The immse system

Contrasts with the traditional meta search engines, the IMMSE system has four advantages. Firstly, the abstract produced by the system is more comprehensive and more precise than the one that is produced by traditional SE; Secondly, through the establishment of user interest model, the different user will obtain the different results list to the same query. Furthermore, the IMMSE has local database and local spider which can not only decrease the scale of data transferring between IMMSE and the independent SEs which are connected to IMMSE but also help to make the Second-search process based on the Topic-clustering character of hyperlinks; Thirdly, IMMSE can call the most suitable independent SEs according to the user query by introducing the intelligent dispatch strategy of independent SEs; Lastly, an improved PageRank algorithm, which is called A-PageRank, is used in this system. The algorithm distributes the PageRank value of the source page to its Link-out pages based on the topic similarity. The architecture of IMMSE can be illustrated by the Figure.1

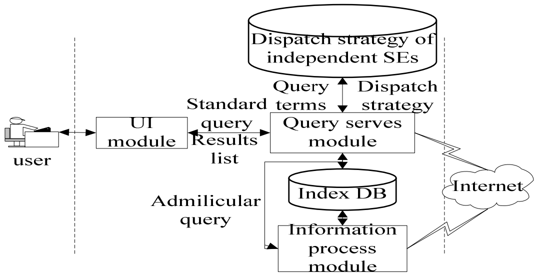


Figure1 the architecture of IMMSE

2.1 User Interface module

The main functions of the UI module can be generalized as follows. The first function is to extract the useful information by processing the user registration information and the user logs. These extractions can be regarded as the basis of the original user interest model; the second one is to define the set of independent SEs which connects to the IMMSE and the last one is to receive the user query and display the results list in a certain style. The architecture of UI module is shown as follows

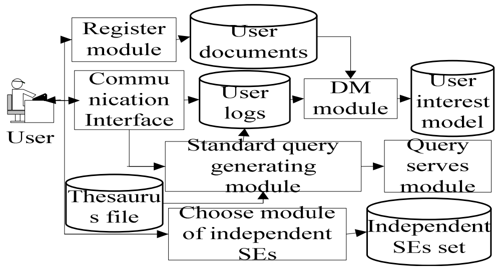


Figure2 the architecture of the UI module

IMMSE uses the user interest model to guide the task of web pages filtering and the user interest model can be improved automatically by consistently studying the user habits and interests by itself. The adjusted strategy of the user interest model can be divided into 3 aspects. Firstly, when one key term in the model is used by the user, the value of the key term should be adjusted; Secondly, when IMMSE gets a new key term which is not in the original user interest model, the new term should be given a original value and be added into the model so that the user interest model can be extended; Thirdly, the memory space of the user interest model is limited, so, the term of low value should be erased from the model when the memory space is full[4].

2.2 Information processing module

The information processing module can be used to call the local spider to fetch and process the web pages. There are three occasions to call the local spider in this system. One is selecting a URL from the URL seed bank as the starting point of searching when the server is in spare situation. The next is, when the number of the results list returned by the independent SEs is small or user is not satisfied with the results list, selects a URL from the results list as the starting point of searching. This method is called the Second-search method which bases on the fact that the hyperlinks between web pages are not disorderly but clustering according to the topic. It means the match degree on topic between a source web page and one of its Link-out pages is larger than one between it and a web page which is fetched randomly from the Internet [5]. The last is the sojourn time of one page in local database is larger than the predefined time, which is actually a problem of page refreshing. The main idea of the strategy of page refreshing can be generalized as follows. When the actual refresh rate of local database is larger than the ideal refresh rate, the refresh cycle should be shorted; On the contrary, when the actual refresh rate is smaller than the ideal one, the refresh cycle should be longed. Defines the original refresh time of local database is $T_0=a$, the ideal refresh rate is $p(0\leq p\leq 1)$, the number of web pages which are updated in the $(i\text{-}1)th$ refresh cycle is S. The ith refresh cycle, denoted by T_i, can be defined as (1)

$$T_i=\begin{cases} T_{i-1}\left[1+k\left(p-S/N\right)\right], p-S/N\leq -d \\ T_{i-1}\left[1+\left(p-S/N\right)\right]*k, p-S/N\geq d \\ T_{i-1}, -d<p-S/N<d \end{cases} \qquad (1)$$

where N denotes the sum of the pages in the local database, d is a diversity factor which denotes the max permissive diversity between ideal refresh rate and actual refresh rate and p is a damping factor which is used to amplify the influence of d. The architecture of the information processing module is described by Figure. 3

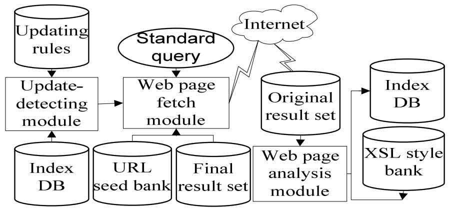


Figure3 architecture of the information processing module

2.3 Query service module

To the different independent SEs, they have different query expression forms and different display format. So, it is necessary to find a medium form between IMMSE and all the independent SEs which are connected to it. Considering the fact that the display format and the actual data are storied separately in XML, so, the XML can be used as a medium form. The result pages returned by different independent SEs can be shown in the same style which is selected by user from the XSL style sheet [6][7]. The architecture of this module can be shown as following figure

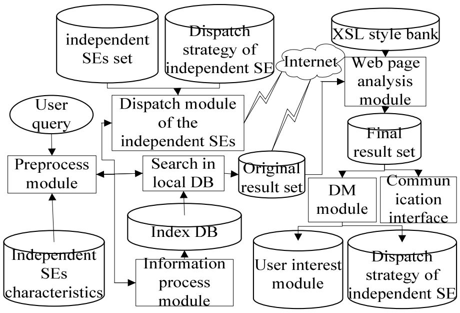


Figure 4 architecture of the query serves module

3 Some key technologies about the realization of the system of immse

For users, a nice SE means the high responsiveness, the friendly UI and the reasonable results list. In order to realize the goal, IMMSE introduces many advanced technologies and some ones will be demonstrated in detail in the following sections.

3.1 User interests mining[8]

One of the key questions of the SE is how to match the query information space with the web page information space so that a reasonable results list can be returned by the SE. However, in current web information retrieval applications, user query is always too short and fuzzy to describe user real request. There are many researches about user activities which have proved that, in web retrieval environment, user query is always 1~3 terms and users seldom consider how to express their requirements precisely. Furthermore,78 % of the users never adjust their original queries based on the results list returned by the SE [9][10][11]. So, the results list returned by the current SEs is always imprecise and seldom satisfies users. The article proposes a new conception, which uses the user interest model to adjust the results list.

How to build a suitable user interest model which can describe precisely the real interests and habits of users? User interests can be divided into major interests and minor interests according to the static distribution. They are also can be divided into stable interests and casual interests according to the dynamic distribution [12]. The major and stable interest can be regarded as the real interest of users. In this paper, a new algorithm based on the cluster analysis to find the real interest of users is proposed and the main idea of the algorithm is displayed as follows: regards each user operation to the SE as a record which is denoted by r_i and the set of records in a week is regarded as a semantic paragraph which is denoted by s. Supposes t_k is a key word in semantic paragraph of s_j and its value, which is denoted by $\mathrm{W}_j(t_k)$, can be defined using the following equation

$$\mathrm{W}_j(t_k) = \mathrm{TF}_j(t_k) * \mathrm{RF}_j(t_k) \qquad (2)$$

where $\mathrm{TF}_j(t_k) = \sum_{r_i \in s_j} tf_i(t_k) * f(TIME)$

$$\mathrm{RF}_j(t_k) = \sum_{r_i \in s_j} BOOL\ (r_i, t_k)$$

$$BOOL\ (r_i, t_k) = \begin{cases} 1, & \text{if } t_k \text{ occurses in } r_i \\ 0, & \text{otherwhise} \end{cases}$$

where $f(TIME)$ is a damping function about time. Then, the key terms in s_j should be ranked according to the descending order and the vector representation of s_j , which is denoted by $V(s_j)$, can be defined by the first num key terms. Viz. $V(s_j)=\{\mathrm{W}_j(t_1),\ldots\ldots, \mathrm{W}_j(t_{num})\}$. The purpose of compelling the vector representation of each semantic paragraph to have the same number of key terms is to simplify the computing complexity.

Algorithm 1 the user interest model construction

a). Defines the set of semantic paragraphs is $S=\{s_1,\ldots,s_N\}$ and each semantic paragraph, which is denoted by s_i, will be regarded as a class that have only one member viz. $c_i=\{s_i\}$. These classes form a cluster class of S, which is denoted by $C=\{c_1,\ldots,c_N\}$;

b). By computing the similarity of each pair of classes, the similarity matrix can be obtained.

$$s_{ij} = \mathrm{SIM}(c_i, c_j) = \mathrm{SIM}(s_i, s_j) = \frac{V(s_i) \bullet V(s_j)}{V(s_i) \times V(s_j)} \qquad (3)$$

where the s_{ij} denotes the similarity between c_i and c_i. The original similarity matrix is a symmetric matrix and a triangular matrix can be obtained by conserving only partial elements. The triangular matrix is displayed as follows

$$\begin{pmatrix} 1 & s_{12} & s_{13}\cdots & & s_{1N} \\ & 1 & s_{23}\cdots & & s_{2N} \\ & & \cdots\cdots\cdots & & \\ & & & 1 & s_{(N-1)N} \\ & & & & 1 \end{pmatrix}$$

The space complexity of this matrix is N*(N-1)/2 and the time complexity is O(N^2);

c). The pair of classes with the max similarity, which is denoted by arcmax$\sum$SIM(c_i,c_j),is chosen from the similarity matrix;

d). A new class $c_k=c_i \cup c_j$ will be added into C and c_i, c_j will be eliminated from C;

e). Selects the class of c_{max} which has the max number of records from C and computes the similarity between c_{max} and the other class.

If $\dfrac{\sum_{c_j \neq c_{\max}, c_j \in C} \mathrm{SIM}\left(c_{\max}, c_j\right)}{\left|C\right| - 1} \leq \theta$ 或 $\left|C\right| \leq 5$ then end the algorithm, else jump to the step of c), where θ is a threshold factor.

Algorithm 2 the adjustment of the interest model

Suppose $C=\{c_1,c_2,...,c_m\}$ when the adjustment will be executed and the new semantic paragraph is s_k

a) The semantic paragraph of s_k will be regarded as a class which have only one member viz. $c_k=\{s_k\}$;

b) Computes the similarity between c_i(i=1,...,m) and c_k;

c) Chooses the class of c_j which have the max similarity with c_k. Then, the class of c_k will be absorbed by c_j and the key terms of c_j will be adjusted;

d) Jump to the step of e) of algorithm 1.

3.2 The dispatch strategy of independent SEs

As mentioned in section 1, the different independent SE has a different performance to the same user query. If IMMSE calls the same set of the independent SEs to all different queries, the available information rate must be cut. The user logs of Tian Wang SE in April, 2000 have been studied. The statistics exposes the fact that, in the results list, the traffic volume of the first 5 pages runs up to 75% of the total traffic volume [13]. So, it is reasonable to use the proportion in the first 5 pages of the results list and the one in the latest 3 pages as a criterion to weigh the sensitivity of this independent SE to a certain user query. Defines the *jth* user query is $Q_j=\{t_{j1},t_{j2},\ldots,t_{jn}\}$ and its corresponding value vector is $Q_j'=\{w_{j1},w_{j2},\ldots,w_{jn}\}$ where the t_{jk} (k=1 to n) denotes the *kth* key term of Q_j. R_i is a certain independent SE. To the query of Q_j, the proportion of result pages returned by R_i in the first 5 pages of the results list is denoted by p_j and the proportion which is in the latest 3 pages of the results list is denoted by q_j. u_0 denotes the original adaptive factor of R_i to each key term of Q_j and v_0 denotes the original unadapted factor. So, after the *jth* search operation, the adaptive and unadapted factor should be adjusted as follows:

$$\begin{cases} v_i = v_0(1+\sum\limits_d w_{jk}^{(d)} * q_j) \\ u_i = u_0(1+\sum\limits_d w_{jk}^{(d)} * p_j) \end{cases} \tag{4}$$

where $w_{jk}^{(d)}$ denotes the value of t_{jk} in *jth* search operation. The sensitivity of R_i to t_{jk} can be denoted by the equation of $W(t_{jk})=u_i-v_i$. So, to the query of Q_j, the independent SEs which have large value to the expression of $\sum_{t_{jk}\in Q_j} W(t_{jk})$ should be chosen.

3.3 A-Page Rank algorithm

In traditional PageRank algorithm, the PageRank value of the source page is distributed evenly to its all Link-out pages. Does this distribution is reasonable? Of course, the answer is not. If all web pages have only one topic, PageRank algorithm undoubtedly is a perfect ranking algorithm. But web pages have millions of different topics or more. It means that the distribution method used in the traditional PageRank algorithm is unreasonable. According to some related researches, the source page has much larger influence to the Link-out pages which have the same topic than others which have the different topic. So, a new PageRank algorithm based on topic is proposed in this section, which is called A-PageRank.

The first task is to find the factor that can represent the topic of one page mostly. The anchor text is always regarded as the most proper representative of the topic because of the following characteristics. Firstly, anchor text has a simply structure; Secondly, one web page always has more than one anchor text, every anchor text is written by a different author of the Link-in page and presents the author opinion to this page. So, the set of anchor texts from the different authors can avoid the bias of one person and can objectively reflect the actual topic of a page. Not every term in anchor text has semantic meanings, so, the conceptions of the primary feature term (PFT), primary feature field (PFF) and primary feature space (PFS) are introduced.

Definition 1: In Internet environment, there are some words which are used by different authors of different pages to express the content and stand out the topic of their pages. These words are called PFT and the corresponding fields are called PFF. The set of all pages PFF are called PFS.

Because the tags of <b>, <i> are always used to express the main idea of a page and the tags of <title>, <h1>, <h2>, <h3> are always used to express the topic of a page, so, the set composed with these tags can be regarded as the PFS. Not every term in PFS can be regarded as a PFT, only the term used by the majority of authors can be regarded as a PFT. Thus, the value of a PFT can be substituted by the frequency that the PFT occurring in the PFS. Uses the vector of Anchor$_p$= $\{t_1,t_2,\ldots..,t_n\}$ denoting the anchor text set of page *p*, the value of t_i , denoted by V$_{ti}$, can be defined using the following equation$_{[14]}$

$$\mathrm{V}_{ti} = \sum_{k=1}^{N} tf_{ik} \qquad (5)$$

where t_i (*i*=1 to *n*) denotes a PFT that is contained by the anchor text set, *N* denotes the number of pages in page set, tf_{ik} denotes the frequency of the t_i occurring in the PFS of page *k*. Furthermore, the PFS is used to denote the common opinion about what terms can be used to express the topic of a page precisely. Therefore, it is necessary to make a unitary processing to limit the influence of the same author' opinion to a PFT. Viz. the definition of tf_{ik} is shown as follows

$$tf_{ik} = \begin{cases} 1, & \text{if term } ti \text{ occurs in the primary field of page } k \\ 0, & \text{otherwise} \end{cases} \qquad (6)$$

Then, the (5) can be simplified as follows

$$\mathrm{V}_{ti} = \sum_{k=1}^{N} tf_{ik} = DF \qquad (7)$$

where DF denotes the number of pages which t_i is contained in their PFS. By the smooth processing, the value of V$_{ti}$ can be defined using the following equation

$$\mathrm{V}_{ti} = \lg\left(DF+1\right) \qquad (8)$$

According to (8), the value vector about the anchor text set of page *A* can be denoted by Achor$_A$={ V$_{t1}^{(A)}$,……,V$_{tn}^{(A)}$ } and the value vector about anchor text set of page *B* can be denoted by Achor$_B$={ V$_{t1}^{(B)}$,……,V$_{tn}^{(B)}$ }. The topic similarity between *A* and *B*, denoted by SIM$_{topic}$(*A*,*B*), is defined by the following equation

$$SIM_{topic}\left(A,B\right) = \frac{\sum_i V_{ti}^{(A)} * V_{ti}^{(B)}}{\sqrt{\sum_i \left(V_{ti}^{(A)}\right)^2} * \sqrt{\sum_i \left(V_{ti}^{(B)}\right)^2}} \qquad (9)$$

So, the traditional equation to compute the PageRank value of one web page can be substituted by the (10)

$$\mathrm{PR}_A\left(A\right) = \left(1-d\right)/N + d * \sum_i \frac{SIM_{topic}\left(A,T_i\right)}{\sum_{j=1}^{C(T_i)} SIM_{topic}\left(Link\_out\left(T_i,j\right),T_i\right)} * \mathrm{PR}\left(T_i\right) \qquad (10)$$

where Link_out(T_i , *j*) denotes the j_{th} Link-out page of T_i.

4 Conclusion

In this article, the architecture of an intelligent mix meta search engine is proposed theoretically and its main modules are illustrated in detail. This system resolves the problem that the different users always obtain the same results list to the same query by using the algorithm 1 and algorithm 2 which are mentioned in section 3.1. Furthermore, in order to make good use of the topic sensitivity of the independent SE, an automatic dispatch strategy of independent SEs is introduced. Lastly, an improved PageRank algorithm, which is called A-PageRank, is demonstrated. This algorithm distributes the PageRank value of the source page based on the topic similarity but does not adopts the method of distributing evenly. So, IMMSE can guide users to locate the required information much precisely and quickly.

References

[1] http://www.cnnic.net.cn/html/Dir/2007/07/17/4722.htm

[2] Deng Changshou, Zhao Bingyan. Investigation on the Next Generation Web Search Engine. Information Science, Vol. 23, No.3, 2005, pp. 426-430

[3] Amanda Spink, Bernard J.Jansen, Chris Blakely. Overlap among major web search engines. Proceedings of the Third International Conference on Information Technology: New Generations(ITNG'06). IEEE COMPUTER SOCIETY, 2006

[4] Zhao Zhongmeng, Yuan Wei, He Shili, Shen Junyi. Reseach on the Intelligent Adjustive Algorithm for User Profile in Personalized Search Engine. Computer Engineering and Applications, Vol. 24, 2005, pp. 184-187

[5] Jeffrey Dean, Monika R. Henzinger. Finding related pages in the world wide web. Computer Networks, Vol. 31, 1999, pp. 1467-1479

[6] Robert W.P.Luk, Tharam S.Dillon, Vincent T.Y.Ng Supporting metasearch with XSL. Systems and Software, Vol. 73, 2004, pp. 159-168

[7] Yuan Fuyong, Chen Jinsen, Lin Haixia. A study on Intelligent Meta-Search Engine Based on Xml. New Technology of Library and Information Service, July 2006, pp. 29-33

[8] Shi Zhongzhi. The Discovery of Knowledge. Tsinghua University Press, 2002

[9] Anick PG. Adapting a full-text information retrieval system to computer the troubleshooting domain. In: Croft WB, van Rijsbergen CJ, eds. Proc. of the 17th Annual Int'l ACM-SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'94). ACM Press, 1994, pp. 349-358

[10] Stefan K, Armin H, Markus J, Andreas D. Improving document retrieval by automatic query expansion using collaborative learning of term-based concepts. Lecture Notes in Computer Science, 2002, pp. 376-387

[11] Croft WB, Cook R, Wilder D. Providing government information on the Internet: Experience with THOMAS. In: Proc. of the 2nd Int'l Conf. in Theory and Practice of Digital Libraries (DL'95). Texas, 1995.http://csdl.tamu.edu/DL95/ papers/croft/croft.html

[12] Guo Yan, Bai Shuo, Yang Zhi-feng, Zhang Kai. Analyzing Scale of Web Logs and Mining Users' Interests. Chinese Journal of Computers, Vol.28, No. 9, 2005, pp. 1483-1496

[13] Li Xiaoming, Yan Hongfei, Wang Jimin. Search Engine—theory, technology and system. Beijing: Science Press, 2005

[14] Zhang M, Ma SP, Song RH. DF or IDF? On the use of primary feature model for Web information retrieval. Journal of Software. Vol. 16, No. 5, 2005, pp. 1012-1020

A Research on J2EE Technology And Its Application

Wenfei Lan

College of Computer Science, South-Central University for Nationalities, Wuhan, Hubei, 430073, China

Email: lanwenfei1@163.com

Abstract

J2EE is a specification for developing enterprise Web-application. In this paper, a Web application development policy with the integration of Struts and Hibernate framework based-on J2EE platform is proposed according to the shortage of Struts. The persistence of data is implemented by Hibernate framework, view and controller are implemented by Struts. It has been applied in a CBC Foreign Credit Card trade Web system which has pass the examination and approval of Construction Bank of China (CBC). This system can meet the demand of long-term developments of trade under the Internet environment.

Keywords: J2EE; Design pattern; Struts; Hibernate1 Introduction

1 Introduction

J2EE is one of the most advanced web application development platforms. Struts, which has implemented the MVC design pattern, is applied in the Web tier. Struts has solved the problem of separation between user interface and business logic in a good way. It enables the hierarchy of Web application, making it easier to collaboration. However, it is not enough to construct J2EE architecture on with Struts. The design of Struts concentrate too much on the presentation logic tier, thus has no control of data tier. It is not reasonable to implement the J2EE model with one tier. It is a crucial problem as how to enhance the control on data tier by Struts.

Hibernate can implement the mapping between persistence object and table in database. We use Hibernate as an extension of Struts that is the framework of Web application development mentioned in this paper.

2 The Implementation of MVC With the Integration of Struts and Hibernate

The design pattern of integration struts with Hibernate takes advantage of the J2EE's MVC pattern. The view and controller part are implemented by Struts, but the model is implemented by Hibernate as well as the JavaBean component implements the business logic. Hibernate takes the responsibility for the persistence operation on entity field objects, while controller tier would not access the persistence tier directly, it responses the users' requests by invoking various business method providing by model layer. The integration framework of MVC is shown in Figure 1 below.

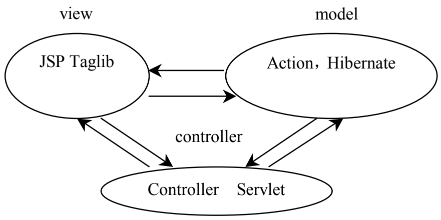


Figure 1 An implementation of MVC with the integration framework

3 The Design of Web Application Rchitecture With The Integration of Struts And Hibernate

As is shown in Figure2, we use Struts and

Hibernate as the development framework for Web application. It is clear that this Web application is a four-tiers architecture.

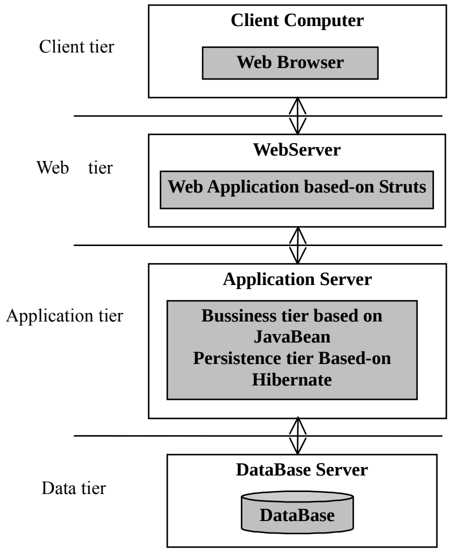


Figure 2 Web Application Architecture

The integration of Hibernate into Struts not only enables the Web application inherited merit from the Struts framework, making the presentation layer separated from business logic, but also enables the Web application to extend the control on persistence layer, thus improves the security of data operation and transaction. Because of the inteqration, Software developers may be freed from the presentation tier and database programming, and put more efforts on business tier development.

4 Design of the foreign Credit Card trade web system

Functions of the Foreign Credit Card Trade Web System

Foreign Credit Card trade Web system supports the online payment of Credit Card in China. The system consists of three function modules:

(1) Merchant Web Service Module

The merchant Web service module is mainly responsible for providing administrator with the administration of online exchange payment on Web, and maintainment of associated materials. Merchant operator performs operation via Internet. The specific business functions include operator management, merchant function management, Serial query, accounting process, certificate management and so on.

The merchant Web service module may be divided into bank management and merchant management in view of user interface and management level. They control permission strictly. The specific exchange functions include online payment, payment correcting, requisition for payment, recheck of requisition for payment(i.e. settlement), cancel payment and so on.

(2) Payment Gateway Module

It is the merchant validation module that verify the card holder by online process of VISA (MasterCard or JCB) directory server, card-holder verification server and other 3D-Secure system.

(3) Foreign Credit Card Payment Module

This module provides the online purchase payment mechanism for Foreign Credit Card holder. It includes such function as consumer online payment, merchant online registration and consumer payment message.

The three modules mentioned above communicate with each other through interfaces. The payment gateway module and merchant Web service module are responsible for sending information and they may work as an whole to perform the identity authentication process. Only those who have passed the authentication by International organizations can perform associated exchange operations, and communicate with Foreign Credit Card payment module by Foreign Credit Card online payment interface.

Design Schemes

Design Pattern: This system is based-on the J2EE's MVC design pattern. The client layer is the View, which is consisted of DHTML. Bussiness logic tier and persistence tier are the Model, which are the core parts of the system made up of JavaBean and Hibernate. Web tier is the Controller that controls the users' business logic operations, and display the operation result to client which is implemented by Servlet.

The MVC model of this system is shown in Figure 3. The PO, which is persistence object, is generated by invoking model tier, presented to client by JSP.

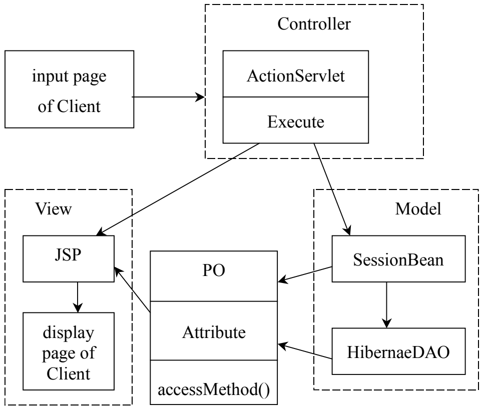


Figure 3 the MVC Model of Transaction System

Design of System Framework: The Foreign Credit Card online transaction system makes use of J2EE Framework with the integration of Struts and Hibernate. The system Framework is shown in Figure 4.

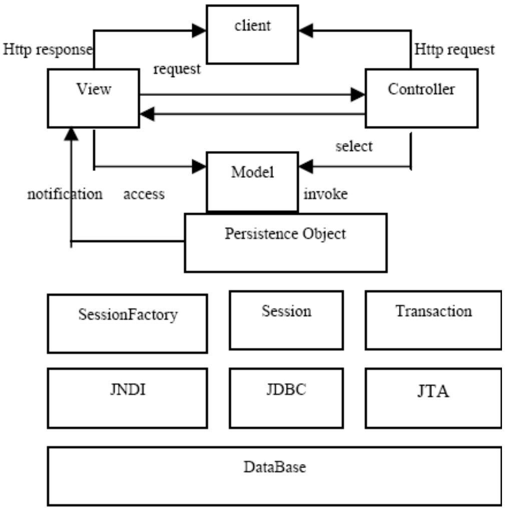


Figure 4 Application Framework of the Transaction System

When the client Http request arrives, ActionServlet would map the request to relevant Action, Action may invoke the interface of business logic model while executing its Execute() method, and the business logic model would operate the business object by HQL statement. Hibernate can transform HQL into SQL, implementing operation on tables of database by methods of Hibernate's Session object. Execute method will return an ActionForward object after finishing its execution, then ActionServlet will accept this ActionForward object and then forward to another named source, which may be a JSP Web page or another action object.

The application framework of this transaction system is integrated with Struts and Hibernate. On the hand, Struts framework is based-on MVC design pattern, it improves the maintainability and extensibility by dividing the system into business logic tier, control logic tier and presentation logic tier. On the other hand, the persistence layer of the model of Struts framework is implemented by Hibernate framework, View and Controller are implemented by Struts. By integrating Hibernate, not only enables the system inherit many merits from Struts framework, but also improves the shortages of the persistence layer of Struts.

database design: Database is responsible for data exchange between several modules, storing the input ,output and middle result of analysis and computation. Database design should make the tables loose coupling. Since the limitation of layout, we only list four tables related to operator management: tables of operator, role, Committee and OperatorLog, the relationship between these tables is shown in Figure 5.

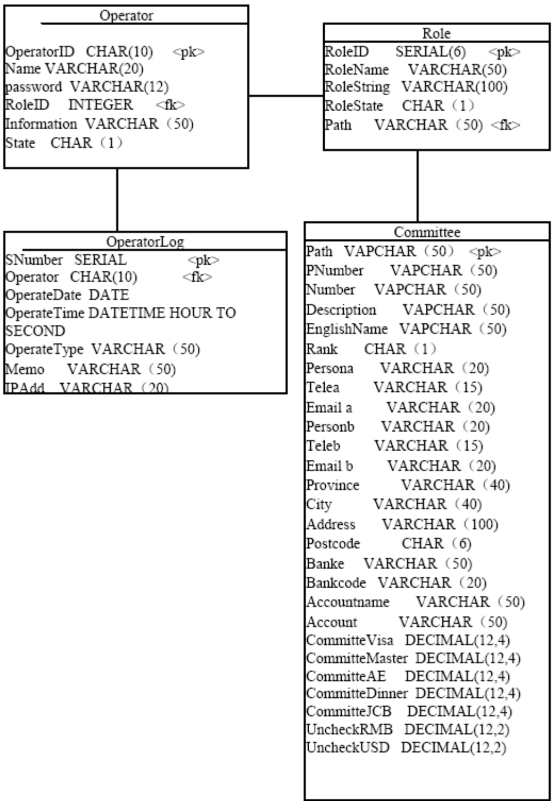


Figure 5 Relationship between Tables

5 Conclusion

J2EE is the mainstream Web application development platform. By integrating Struts with Hibernate as the Web application development policy, we may benefit from both of the two frameworks. We have developed a Web system of Foreign Credit Card transaction which is running normally. It demonstrates the validity and feasibility of the integrated framework.

References

[1] Xiaohua Liu, and Yaqiang Cheng, J2EE application developmentCheng, Beijing: Publishing House of Electronics Industry, 2004

[2] Tianhe Cheng, Struts, Hibernate, Spring integrated development bible, Beijing: Publishing House of Electronics Industry, 2007

[3] Yang Liu, Master Hibernate. Beijing: Publishing House of Electronics Industry, 2005

[4] Jun Zhang, “Research on Application Framework of Publishing Domain based-on J2EE”, Computer application and software, 24(5), 2007, pp. 102-105

[5] Xiaohua Liu. J2EE Enterprise Application Development, Beijing: Publishing House of Electronics Industry, 2003

System Simulation and Reliability Analysis of OFDM Based on System-View

Zhiyong Du[1] Xianfang Wang[2,3] Zhou Yu[2] Haiyan Zhang[2]

1 Henan Mechanical and Electrical Engineering College, Xinxiang, Henan, 453002, P.R.China
Email: zhiydu@126.com

2 Henan Institute of Science and Technology, Xinxiang, Henan, 453003, P.R.China

3 School of Information & Control Engineering, Jiangnan University, Wuxi, Jiangsu ,214122, P.R.China
Email: wangxianfang@sina.com

Abstract

Orthogonal Frequency Division Multiplex (OFDM) has been applied broadly in wideband wireless communication systems because it has ability in anti frequency selective shading. The basic principles of OFDM are described and then the schemes of transmitter and receiver are given in detail. According to the system architecture analysis, it builds the OFDM system model and simulations model through Matlab tools. The different computer simulation results and BER can be got at different channel conditions and modulation modes. A specific example is given and amore in totalistic result can be got from this example. The simulation results show that OFDM system has a good quality in anti channel interference.

Keywords: OFDM, Performance Analysis; System View; Simulation; Reliability.

1 Introduction

With the development of the digital signal processing technology and large-scale integrated circuit technology, allowing troubled OFDM (Orthogonal Frequency Division Multiplexing) technology to achieve the many problems are no longer existed, OFDM technology in the field of high-speed wireless communications is not only more and more attention, but also is one of the main techniques in 3G and 4 G [1-3]. It is very important to study its properties and characteristics

OFDM is a system of multi-carrier modulation. The main idea is: several orthogonal channel sub channels will be divided into high-speed data signals into parallel low-speed flow of data, to the modulation in each sub-channel on transmission. Quartered signal through the receiver can be used to separate related technologies, thus reducing the interference between the sub channels ICI[4,5]. Each of the signal bandwidth of sub channels is less than the relevant channel bandwidth, so every sub channels on the decline can be viewed as flatness, which can be eliminated ISI. And because each sub channel bandwidth is the only channel a fraction of the bandwidth, channel equalization become relatively easy.

In order to have an all-around knowing of quartered modulation and de-modulation, mainly research the implementing technology problem of quartered system. This paper introduced the basic principle of quartered system, given the architraves of system and derivation of expressions, and established a system simulation model. Through simulating and analyzing of the results, some conclusion can be obtained. These results lay the foundation for the subsequent development of communication technology.

2 The Base Theory of OFDM

The primary advantage of OFDM over single-carrier schemes is its ability to cope with severe channel conditions — for example, attenuation of high frequencies in a long copper wire, narrowband interference and frequency-selective fading due to multipath — without complex equalization filters. Channel equalization is simplified because OFDM may be viewed as using many slowly-modulated narrowband signals rather than one rapidly-modulated wideband signal. The low symbol rate makes the use of a guard interval between symbols affordable, making it possible to handle time-spreading and eliminate intersymbol interference (ISI). This mechanism also facilitates the design of single-frequency networks, where several adjacent transmitters send the same signal simultaneously at the same frequency, as the signals from multiple distant transmitters may be combined constructively, rather than interfering as would typically occur in a traditional single-carrier system.

2.1 OFDM algorithm [6]

Suppose a signal sequence (D_1, D_2,…, D_{N-1}), after IDFT transforming, OFDM symbols could be obtained {d_1, d_2 ,…, d_{N-1}}, where:

$$D = \sum_{n=0}^{N-1} De^{j2\pi n\frac{k}{N}} \tag{1}$$

These transformed OFDM symbol could become a base band transmission signal through the D / A converter with the transformation rate of f and a low-pass filter

$$\chi(t) = \sum_{n=0}^{N-1} De^{j2\pi nt\frac{f}{N}} \tag{2}$$

Assume a symbol cycle of the system is T, then width of the symbol OFDM system is T_s = NT, the n-th sub-carrier frequency is $f_n = f_0 + n / T_s$, f_0 is the minimum available frequency, when base band method: $f_0 = 0$. If each sub-carrier maintain this relationship, then the orthogonally of these sub-carriers would be satisfied.

Through the time-varying channel which the pulse response is h (t), the received signal could be expressed as:

$$R(t) = \int x(t-\tau)h(t,\tau)d\tau + n(t) \tag{4}$$

The received signal through converting of A / D and demodulating, the resumed signal could be obtained:

$$D = \frac{1}{N} \sum_{n=o}^{N-1} Re^{-j2\pi m\frac{k}{2N}} \tag{5}$$

2.2 The basic diagram of OFDM

Based on the above basic principles and algorithms, the basic block diagram of OFDM transmission system [7] could be designed as in Figure 1: after transformed by A / D (It is more suited for digital communications that a continuous signal would be transformed into a discrete signal), and string / parallel transformed, the sequence of transmission would be modulated parallel on many sub-carrier (orthogonal multiplexing), and then coding these channels (The aim is to reduce the error rate to the extent required by the system, the greater channel capacity, the lower error rate), then completive coding for the error correction package after be coded which could improve the correction ability of the coding. And restore the lost information of some sub-channels because of the frequency selective fading.

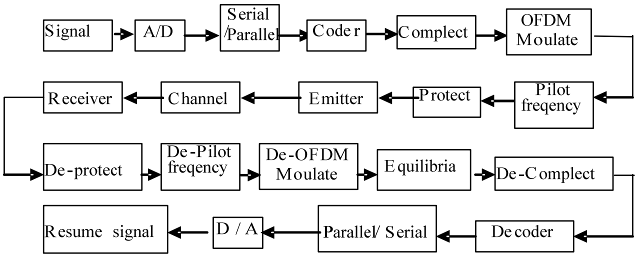


Figure 1 Basic diagram of tenets of OFDM

In OFDM, the sub-carrier frequencies are chosen so that the sub-carriers are orthogonal to each other, meaning that cross-talk between the sub-channels is eliminated and inter-carrier guard bands are not required. This greatly simplifies the design of both the transmitter and the receiver; unlike conventional FDM, a separate filter for each sub-channel is not required.

The orthogonally also allows high spectral efficiency, near the Nyquist rate. Almost the whole available frequency band can be utilized. OFDM generally has a nearly 'white' spectrum, giving it benign electromagnetic interference properties with respect to other co-channel users.

The orthogonally allows for efficient modulator and demodulator implementation using the FFT algorithm. Although the principles and some of the benefits have been known since the 1960s, OFDM is popular for wideband communications today by way of low-cost digital signal processing components that can efficiently calculate the FFT.

OFDM requires very accurate frequency synchronization between the receiver and the transmitter; with frequency deviation the sub-carriers will no longer be orthogonal, causing inter-carrier interference (ICI), i.e. cross-talk between the sub-carriers. Frequency offsets are typically caused by mismatched transmitter and receiver oscillators, or by Doppler shift due to movement. Whilst Doppler shift alone may be compensated for by the receiver, the situation is worsened when combined with multipath, as reflections will appear at various frequency offsets, which is much harder to correct. This effect typically worsens as speed increases, and is an important factor limiting the use of OFDM in high-speed vehicles. Several techniques for ICI suppression are suggested, but they may increase the receiver complexity.

According to the power and spectrum utilization requirements to select OFDM modulation method, the usually adopted modulation methods are 8 PSK, QPSK, 16QAM, and the other methods [8], which is to transform the original signal suitable for transmission channel signal. Retaining some of the insertion of sub channels Pilot to achieve the synchronization system, that is, the signal power of signal is detected in the receiver, and compared to with the threshold, determined whether the OFDM signal could be arrived at the receiver, then have a related operations between the received signal and local replication synchronization signals, controlled the timing error in the sample value, then according pilot signal to balance these sub channels and ensuring orthogonally. Then these protections are intervened among OFDM symbols. After a series of changing, the signal could be enter the channel and transform. The recovery phase is the anti- changes of the transmission phase. Balancing is used to eliminate residual ISI to ensure signal restoration accurately.

3 Simulation of OFDM Based on System View

3.1 Introduction of simulation tool of system view [9]

System View is designed by Elanix Company of United States, which is one of integrated visualization software of a complete dynamic system design, simulation and analysis, and is a very good platform with anglicizing, designing and researching. It runs on Windows, has a very user-friendly interface, users can use the mouse to complete all kinds of complex application processing, and can also interface and dialogue window, the function module parameters set. System View can be used to establish and rapid modification to the system, the system simulation, analysis and processing, and use the system to provide rapid development tools to create accurate models of dynamic systems. System view includes basic database and communications, DSP, logic, RF / analog, the user code and other professionals.

System view mainly has the following characteristics:

(1) Powerful simulation in designing;

(2) The rich resources;

(3) Open and friendly user interface;

(4) Intelligent aided design;

(5) Dynamic anglicizing and post-processing.

3.2 The establishment of model parameters and settings

According to the basic principles of block diagram about the OFDM system, utilizing the System View, the simulating model is designed as in Figure 2:

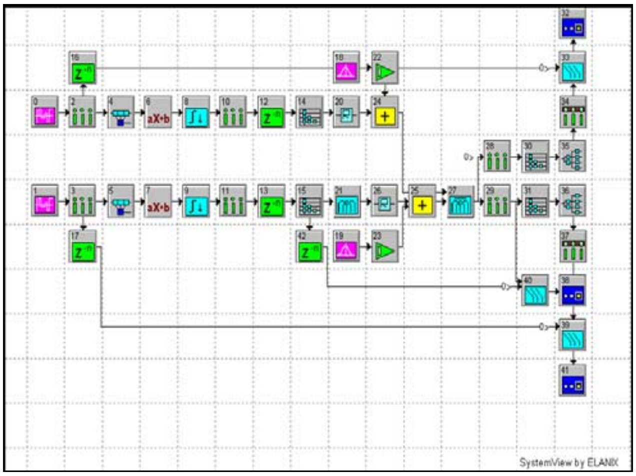


Figure 2 Simulation diagram of OFDM based on Systemview

Noting module and settings parameter are as follows:

Token 0, 1: pseudo-random sequence --- the rate is 512 bps, the amplitude is 0.5 V, two-level code;

Token 2, 3: sampling --- the sampling rate is 1024 Hz ;

Token 6, 7: polynomial ---y = -1+2 x;

Token 22, 23: gain --- the Gain value is - 2 × c_l, c_l is the cycling number of the system;

Token 11, 10: re-sampler ---the rate is 1024 Hz;

Token 12, 13, 16, 17, 42: delayed sampler ---the delayed time is 1μ s;

Token 4, 5: Convolution Encoder--- input bits n = 2, information bit k = 1, constraint length L = 7, generating polynomial (171,133);

Token 20, 26: Rice fading channel--- coherence time T = 0.01s, fading factor K = 0;

Token33, 39: bit error rate --- test bits No.= 512, Threshold = 0.5 V, time offset = 1579 it;

Token 40: bit error rate --- test bits No .= 512, Threshold = 0.5 V, time offset = 3072 bit;

Token14,15: convolution codes code matched---intertwined depth is 8 × 6;

Token 35: hardware judgment n = 2, k = 1, L = 40;

Token 36: bit software judgment, peacekeeping bit= 3, n = 2, k = 1, L = 7;

Token21,27: orthogonal frequency division multiplexing modulation; --- N = 512, Ts = 0.5, Δ = 0.01 s;

Token32, 38, 41: final observation window;

Token 8, 9: points token removal filter

4 Operating Results and the Reliability of the System

4.1 Results

Assuming some parameters of system: the sampling rate is 1280 Hz, the sampling points is 32768 points, the running time is 25.599218 s and the cycle number c_l = 5. After above operation, entering the analysis window, opening letter places calculator, selecting Style of the BER Plot Set Starting values and step length, respectively three BER curves, and then use Operator options in the Overlay Plots (superimposed graphics) function, with these three curves in a graph displayed. Figure 3 shows the following:

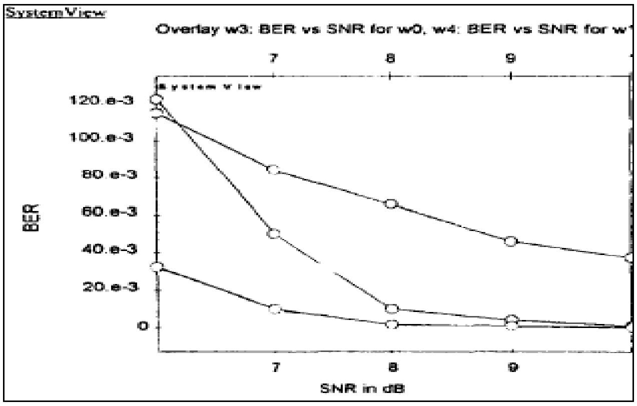


Figure 3 Comparing about BER of three systems

From top to bottom, three curve followed by single-carrier transmission, not coded OFDM system and the coded of OFDM system, which express the BER performance in a certain SNR of these system. Figure 3 shows that the three curves all have the downward trend with the increasing of signal to noise ratio, that is, the signal to noise ratio is downward. Therefore, by the look of SNR respect only, the system should obtain better one yard of performance by mistake, should improve the SNR of the input signal. So by the look of SNR respect only, the system should obtain better one yard of performance by mistake, should improve the SNR of the input signal. From the view point of analyzing by fixed position, when SNR is 7dB, one yard of performance by mistake of OFDM system encoded is higher by 4 dBs than one yard of performance by mistake of OFDM system that is not encoded, the performance transmitted more than the single signal carrier is higher by about 8dB. With increasing of SNR, code with OFDM system that does not encode one yard of performance close quite by mistake, the two all transmit and demonstrate better performance more than the single signal carrier. So in order to improve the performance of the whole system, adopt OFDM system to be transmitted, and it is an effective measure to carry on the channel code to the signal.

4.2 Reliability analysis

Error rate is important parameters of the performance in a communication system. Therefore, this paper adopt error rate to analyze the reliable performance of OFDM system. OFDM system adopts parallel transmission , these signals would be spread to every stature channel on average, and then forming a lot of parallel narrow band sub-channels, and only one code-element could be transmitted in every sub-channel at this moment, therefore when the alternative decline of frequency appears in the channel because of multi-path transmission, only these sub-carriers which leave frequency band sunken and its carried information would be influenced, and the other sub-carriers would not be damaged, and be still transmit correctly. In addition, adding Cyclic Prefix as protect interval in using OFDM. When the protect interval is bigger than the max-delay time, this method could overcome the Inter-symbol Interference (ISI), therefore the robust performance of the system become more better.

In OFDM system, the performance of the whole system can be improved through using sub-sets the channel characteristic and the joint sub carrier encoding and intertwined. Coding can be used various codes, such as block codes and convolution codes. Fading Channel would produce unexpected error data channel, to combat the channel is an effective method in the encoded data intertwined, and that will have unexpected errors Channel transform error independent channels and can be intertwined in time domain can be carried out in the frequency domain.

5 Conclusions

Orthogonal frequency-division multiplexing (OFDM) — essentially identical to Coded OFDM (COFDM) — is a frequency-division multiplexing (FDM) scheme utilized as a digital multi-carrier modulation method. A large number of closely-spaced orthogonal sub-carriers are used to carry data[10-12]. The data is divided into several parallel data streams or channels, one for each sub-carrier. Each sub-carrier is modulated with a conventional modulation scheme (such as quadrature amplitude modulation or phase shift keying) at a low symbol rate, maintaining total data rates similar to conventional single-carrier modulation schemes in the same bandwidth. OFDM has developed into a popular scheme for wideband digital communication, whether wireless or over copper wires, used in applications such as digital television and audio broadcasting, wireless networking and broadband internet access[13].

This paper introduced the basic principle of OFDM and key technologies, and including system simulation and reliable analysis. Because of the advantages of broad prospects in wireless access and mobile high-speed transmission, OFDM has applied to the improving core technology in the next generation mobile communication. System simulation is necessary before OFDM systems have been designed, because it can optimize the whole system parameters and indicators, shorten the development cycle and save manpower, financial and material resources. With the development of OFDM technology, it is foreseeable that in the future, the application of the OFDM technology would expand in the field of wireless high-speed communications, and its excellent performance will also greatly facilitate high-speed wireless communications technology development.

References

[1] Erik Dahlman, Stefan Parkvall, Johan Sköld, Per Beming. “ OFDM transmission”[J]. 3G Evolution, pp: 45-66 2007

[2] Shinsuke Hara, Ramjee Prasad. “Multicarrier Techniques for 4G.Mobile Communication”[M]. London: ArtechHouse, 2003

[3] S. Hara and R. Prasad, “Multicarrier techniques for 4G Mobile Communications”, Artech House, Norwood, 2003

[4] Pierre Siohan, Cyrille Siclet. “Analysis and Design of OFDM/OQAM Systems Based on Filterbank Theory”[J]. IEEE Transac-tions on SignalProcessing. 2002, 50(5) pp: 1170-1183

[5] H. M. Mourad, “Reducing ICI in OFDM systems using a proposed pulse shape”, Wireless Personal Communications, vol. 40, pp. 41–48, 2006

[6] Hamid R. Sadjadpour. “Orthogonal Frequency Division Multiplexing(OFDM)”. Handbook of RF and Wireless Technologies, pp.333-353, 2004

[7] ZigangYang, XiaodongWang. “A Sequential Mont Carlo Blind Receiver for OFDM System in Frequency-Selective Fading Channles”[J]. IEEE Transactions on SignalProcessing. pp.271-280. 2002, 50(2)

[8] Yin Chang-chuan, Luohaichao, “Orthogonal Frequency Division Multiplexing”, Communication Technology of Zhongxing[J], pp.10-15, 2003. 2

[9] Qing Song, et al. “System view emulation and analyzing of the digital communication system”[M]. Beijing: Beihang University press, 2001

[10] K. Fazel, S. Kaiser, “Multi-Carrier and Spread Spectrum Systems”, John Wiley & Sons, 2003, ISBN 0-470-84899-5

[11] Bahai, A. R. S., Saltzberg, B. R., Ergen, M. “Multi Carrier Digital Communications: Theory and Applications of OFDM”, Springer, 2004

[12] "The how and why of COFDM" Jonathan Stott. EBU: EBU Technical Review 278 (winter 1998)

[13] Leeper, David G (2007)., "WiFi - The Nimble Musician in Your Laptop", IEEE Computer Magazine -- How Things Work, April, 2007, pp 108-110

Application Research of CORBA/Web in the Network Management*

Yujian Wang[1]   Chen Liu[2]   Hong Bao[3]

1 Institute of Information Technology, Beijing Union University, Beijing 100101, China
Email: vcplus@163.com

2 College of Computer Science & Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
Email: Lchen@bupt.edu.cn

3 Institute of Information Technology, Beijing Union University, Beijing 100101, China
Email: Baohong@buu.com.cn

Abstract

The technical characteristics of COBRA and WEB in the application of network management are discussed. The feasibility for adopting CORBA techniques to construct telecommunication network management is analyzed A network management model, which is based on TMN logic layering and CORBA interoperable mechanism, is proposed. The model adopts three-layer architecture of “application layer + management layer + network element layer”, implements the function design of network management which includes performance management, fault management, configuration management, account management and security management. The System based on this model adopts B/S mode, so there’s no regional limit and can be extended easily.

Keywords: network management, distributed computing, CORBA, Web

1 Introduction

The network is developing towards the direction of large-scale and isomeric, the traditional centralized network management doesn’t meet the demands of growing network due to its inherent limitations[1]. The technique of CORBA[2] distributed object presented by OMG has become mainstream in distributed application environment. The network management technique based on CORBA also has matured increasingly. Systems designed by using distributed object technique and by exploring the characteristics of CORBA’s ORB, IIOP and IDL[3], have the advantages of transplantation, extensibility and setting-independent and be able to support effectually the exploitation of isomeric distributed system.

While the network management based on CORBA has the distributing advantages, it is not easy to support interactivity. With the advent of internet, how to apply the distributed technique to the internet has been researched. Thus, emerges the network management technique based on the Web[4]. There are two advantages for designing network management system if web technique is combined with CORBA. One is that the system can use the mature web technology (for example CGI and Applet) to realize interactivity, the other is that the large-scale, complex network management system can be realized by adopting the distributed architecture of CORBA which the traditional web technique is unable to achieve.

2 The Network Management System Based on CORBA/Web

The goal of network management is to allocate and control network resources (hardware and software)

* This work is supported by the science-technology development plan of Education Committee of Beijing of China. No. KM200811417001.

properly, ensure the network runs efficiently and reliably, so as to meet the demands of service providers/users. Generally, the tasks of network management includes collecting, analyzing, handling data and other management action. Among popular architectures of network management is SNMP, CMIP, TMN, etc.

2.1 Integrating CORBA with web

CORBA realizes distributed characteristic by distributing management logic to network nodes. Management logic is one CORBA object, which is fixed in one node. The manager needs simply to get the object reference, and uses the reference to call object operation, thus realizes remote management. The network management system based on CORBA is independent of protocols or programming languages. The designer concerns mostly with the interface definition and implementation of distributed object, he/she does not concern as to how the manager communicates with the agent, and the communication is almost always the most complicated task in the designing of the network management.

The biggest advantage to adopt Web network management is its interactivity. The manager can complete network management by using the web browser in any site, his work is not limited to region and platform, while in traditional network management, the manager can only finish managing network in the console. Of course there will be higher request to system security in web based network management[5]. Additionally, making use of web server can reduce the cost of maintenance, because the system can be modified simply by modifying the web server, there is no need to modify any client.

Integrating CORBA with Web enables web client to access the far-ranging CORBA object services which are defined by IDL, and to communicate with other objects which are programmed in different languages. Moreover, their communication is no limited by address space and network area. There are two methods for integrating CORBA with Web. The simplest method is to make use of CGI. But in order to implement bidirectional interaction for the client and the server, the second method has to be used.

This method is depicted in Figure 1. There is an applet in the web page which is made by the request for client to web server, when the browser interprets the applet, the applet which acts as a CORBA client program will access method of remote object. When remote object receives method request, the response is made, and the result is returned to browser.

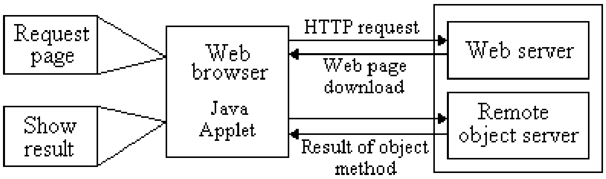


Figure 1 Applet method principle

2.2 TMN architecture

Telecommunications Management Network (TMN) is the configurable network architecture which is defined for managing telecom equipment and operation. At present, TMN has already defined and designed a great number of information models for resource management, which are applied widely for SDH net, ATM net, mobile net and join net. The network management techniques based on OSI and TMN are suitable to manage many simple network objects and network element objects, which do not have complicated operations and mutuality. TMN adopts object oriented modeling technique, and introduces Manager/ Agent mode. The communication between manager and agent is based on their common comprehension for protocols stack (for example SNMP and CMIP protocol) and management information base (MIB).

The existing technique of TMN can't solve problems in operation management completely. For example in commerce business, complicated operation objects and management functions are difficult to be modeled by Guidelines for the Definition of Managed Objects (GDMO). When the actual system conforming to TMN rules is designed, the traditional TMN criterion pays attention mostly to management functions in the layers below network layer and constituting standard information models, rarely consider supporting distributed handling. Besides, TMN pays attention mostly to communications in management information when defines management frame interfaces, and rarely considers development of software techniques

CORBA distributed techniques have advantages TMN does not have in modeling complicated operation objects[6]. It has been becoming a trend to use CORBA to build extensible, distributed and setting-independent telecom network management system of TMN. CORBA is independent of protocols in application layer, it can exert the advantages in network management. For example, if service businesses need to expand their services, the new service can be opened after registered to the CORBA platform, and it doesn't affect the existing services. When the client needing new service submits request, the network management system based on CORBA is able to activate new service. After client request is executed, the result will be returned to client.

Using CORBA technique to build network management system of TMN, the questions need most consideration are how CORBA entities correspond with 1). Manager/Agent mode in OSI management, and 2). criterion definitions in GDOM language, in order to achieve the interoperable with existing TMN system as well as system integration.

2.3 Network management model

The network management model is constructed based on OSI architecture, that includes information model, organization model, communication model and function model. Information model is the abstract denotation for network resources and its supporting management activities, its main function is to describe physical and logistic network resources. Organization model prescribes the definition of every entity and the mutuality of entities in the network management system, some examples of organization models are client/server, manager/agent mode and so on. Communication model prescribes the communication mechanism of entities in the network management system, most communication models adopt OSI protocols. Function model prescribes the functional set in the network management system.

TMN logic layering architecture is a successful organizational mode of network management, it makes complicated network management simple. Consulting the logic layering architecture of TMN, the system adopts the network management model based on TMN logic layering and CORBA interoperable mechanism. Its primary principle is to distribute traditional network management and CORBA network management to different logic layer. The traditional network management (SNMP and CMIP) is used in network element layer, and that the CORBA network management is used above management layer, the main reason is that network management probably runs in isomeric networks, while the managed objects of operation and affair are more suitable to be described by IDL. Moreover the relation between management layer and network element layer is the relation between manager and agent, it's completely standardized. Information model in network element layer is described by MIB or GDOM, and information model in management layer is described by IDL. These are completely different information models of two layers, and there is no need to introduce switch gateway.

Using CORBA techniques to construct TMN management layer needs to find correspondences in CORBA with entities of Manager/Agent mode in OSI. Manager/Agent mode includes three kinds of entities: Manager, Agent and Managed Object (MO). The direct approach is to model Manager and MO by IDL interface. One MO instance corresponds to one CORBA object supporting relevant IDL interface, the communications between Manager and MO is realized by management operation and notify interface.

Differing from mode of traditional network management, the mode this system finally adopts is not two layers architecture of Manage/Agent, but the three layers architecture of “application layer + management layer + network element layer”, which is depicted in Figure 2. Application layer takes on communications with client and puts in network management instructions of manager. Management layer takes charge of data pick-up, which includes cyclic data pick-up and real-time data pick-up, and returns gathered data to upper layer. Management layer also takes charge of tasks which includes data analysis, report referring and real-time monitor. Network element layer consists of communication devices managed by network (for example router, exchanger, etc.) and network management Agent residing in devices, it takes charge of initialization of network management agent proceeding and operation of device MIB, in addition to examining and controlling network devices, in order to get varieties of network information.

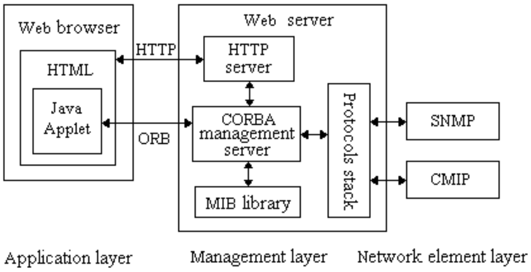


Figure 2 Network management architecture based on CORBA/Web

In the case of CORBA/Web network management, when the web browser user starts network management in a computer, the user must log on the server first to download the applet program in order to get permission to perform management assignment. After the manager inputs the instructions, the instructions are analyzed first by the web browser, and then are send to the web server. The CORBA management server in the web server gathers information of network management and complete varieties of management operation. Finally the HTTP server in the web server returns the result to the web browser user with HTTP form. The system also sets up a platform of transform protocols between web server and network element layer, so that it can convert management protocols in the different network (for example SNMP, CMIP, etc.) into HTTP protocol.

2.4 Design of network management system

When designing a network management system, not only we need to consider the architecture of network management system in level, we should also consider the module composition of network management system in software construction. The software construction of network management system includes commonly three kinds: interface software, special software of network management, support software of network management. Users associate with special software via interface software, generally interface software are in the form of GUI. Special software locates between interface software and support software, it consists of some application elements which are used to realize basic management functions. These application elements can be invoked by user program. The lowest tier of special software provides the service of data transmission in network management, so as to exchange management message. The management information accessing modules and communication protocols stack are support software. MIB includes device information of configuration and action, and parameters of controlling operation. The communication protocols stack supports communications among nodes.

The system adopts B/S structure, the manager can uses WWW browser to manage network from anyplace. Figure 3 shows the system architecture, HTTP server in Interface Module is used to support Web browser, CORBA server communicate with Java applet via ORB. Data Disposal Module, Data Pick-up Module and Authority Management Module in Figure 3 are the special software of network management which needs to be designed, and they are the main body of CORBA management server. Interface Module provides HTTP interface, CORBA interface and third party application interface (for example FTP, SNMP, CMIP, etc.).

Network management includes performance management, fault management, configuration management, charge management and security management, the most important part is performance management. Performance management first sends assignment of data pick-up and finishes data pick-up, and then stores performance data and makes analyze. The basic analyze methods include statistics, summary, log and so on.

Based on these basic analyze, various performance management models are established, and higher level integrated analysis are made, and thus final analysis are formed, for example capability, performance abnormity (alarm), performance forecast and history performance.

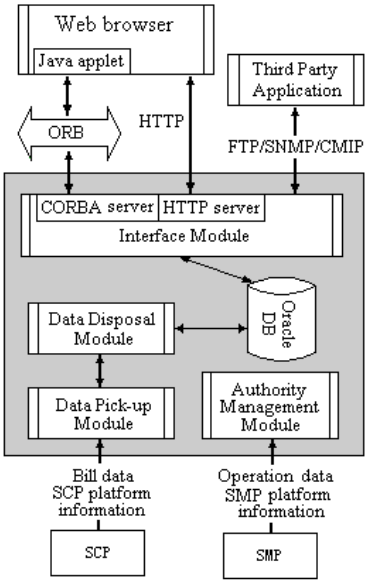


Figure 3 System architecture

Network performance is showed by B/S structure. Using browser, the manager can get expediently analyze report according to different requirements, and monitor network performance in real time. Network performance is viewed with dynamic linear graph. When network performance is exceptional, the manager can get real-time performance alarm. According to different user’s requirements, XML file can be reconfigured, so as to implement dynamic performance data statistics. Statistics data seen in the client end is confirmed by XML configuration file in the server, so needn’t to program repeatedly in the client.

Data Pick-up Module possesses important status in performance management. Manager gets assignment of performance data pick-up from Oracle database in upper layer via calling interface of Oracle, and according to communication protocols, communicates with agent processes residing in devices in network element layer. Manager gets network performance data by means of periodic operation, and sends the data to Oracle database via calling interface.

The system fault management is implemented by polling drive. From the view of real time, adopting the approach of initiative notification can speed up the processing of manager, but increase workload. The service end for mode of CORBA/Web is CORBA object, its connection with the client end may be permanent, thus to ensure that failure alarm and response are real-time in the client end.

Configuration management completes mostly the function of device status examine, polling response, system configuration and configuration query. The scope of change on device status is wide, there is status change of device entirety as well as device mutuality pattern. The system uses database to realize resource configurations, these mainly are network element configuration of SCP/SMP, subsystem configuration, memory configuration, fixed disk configuration and port configuration, include saved configuration value, default value and current configuration value.

For security management, higher security is requested as the system adopts B/S structure. Security service provided by the system includes identity validate, accessing control and safety audit. Accessing control in security service introduces the notion of privilege and authority, which constitutes a set of accessing control strategy and model, separates logically user from accessing authority, distinguishes layer relations among roles, ensures effectually security of accessing control. When someone attempts to acquire password, he is locked by session control till overtime session is cleared up.

The system uses a PC to implement the function of Web server, which connects with SCP/SMP server and client via TCP/IP protocol. The communication model based on CORBA is depicted in Figure 4. Two background programs run in the PC, one takes charge of data pick-up via SCP server and stores in the local database, reads information from SMP server database according to CORBA mode. The other one deals with commands sent by client, accomplishes database query. Another background program also runs in SCP/SMP server, collects system configuration information of CPU, memory and fixed disk. These information is written in a file which is downloaded and stored in local database by system.

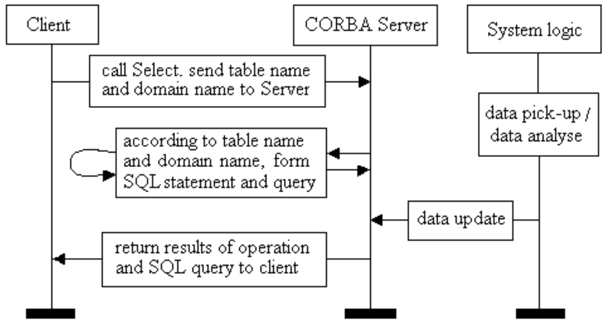

Client
CORBA Server
System logic
call Select. send table name and domain name to Server
data pick-up / data analyse
according to table name and domain name, form SQL statement and query
data update
return results of operation and SQL query to client

Figure 4 Implementation of communication model

3 Conclusion

CORBA technology has been applied to many area in telecommunication industry in recent years[7~8]. This paper presents a CORBA network management model based on TMN logic layering, and we have successfully developed the telecommunication network management system integrating Web techniques. It is shown in practice, compared with traditional network management, CORBA/Web network management based on TMN logic layering not only possesses the characteristics of traditional network management, for example higher efficiency and being suitable to network management, but also makes the best of the characteristics of distributing object, for example computing agilely, being independent of protocols, developing simply, extending easy, etc. The network management model based on CORBA/Web technique expands Client/Server mode[9], and makes manager operation not be restricted by area and platform.

The variety of user demands result in mutability of network management functions[10~11], it puts forward higher requests for software framework of network management. The system is divided into functional modules according to the certain rules, its devise extracts the idea of class factory. The main thread class is defined as base class of all other thread classes, and the others may be loaded dynamically via XML configuration file. The users can add or delete a function by merely adding or deleting the sentence THREAD in XML configuration file. Adopting such function partition is beneficial to extend system, maintain system and upgrade system. In addition, using XML configuration file to load dynamically a thread class, separates operations from concrete implementations. Various sub- threads run independently, and they don't interfere each other generally, thus the system fulfils the demands of high cohesion and low coupling for software engineering.

The system utilizes synthetically CORBA, Web, XML and multithreading technique, implements management for devices and operation of telecom net, and it runs well in practice. For further work, the system will be improved more to meet UNIX or Linux operating system.

References

[1] Huo M, Majumdar S, "Performance of parallel architectures for CORBA-based systems", ACM SIGSOFT Software Engineering Notes, Vol.29, No.1, 2004, pp.249-253

[2] Object Management Group, The Common Object Request Broker: Architecture and Specification, 3.0 edition, OMG document: ptc/2002-06-01 3.0 edition, June 2002

[3] Object Management Group, The Common Object Request Broker: CORBA Component Model Specification, 3.0 edition, OMG document: ptc/2002-06-01 3.0 edition, June 2002

[4] Satoh I, "Building Reusable Mobile Agents for Network Management", IEEE Transactions on Systems, Man, and Cybernetics, Part C, Vol.33, No.3, 2003, pp.350-357

[5] Obelherio R, Fraga S, "Role-based access control for CORBA distributed object systems", Proc. of the seventh International Workshop on Object-Oriented Real-Time Dependable Systems (WORDS 2002), San Diego, IEEE Computer Society, 2002, pp.1443-1530

[6] Bohoris C, Pavlou G, Liotta A, "A Hybrid Approach to Network Performance Monitoring Based on Mobile Agents

and CORBA", Proc. of the 4th International Workshop on Mobile Agents for Telecommunication Applications, Barcelona, Spain: Springer, 2002, pp.151-162

[7] Sakurai CA, Junior MM, "An open system architecture for operation support system at telecommunications service providers", Proc. of the 1st International Symposium on Information and Communication Technologies, Dublin, ACM Press, 2003, pp.524-529

[8] Li J, Zheng FB, Chen ZG, "esearch and implementation of network management based on CORBAR", Application Research of Computers, Vol.21, No.11, 2004, pp.210-214

[9] Abdul-Fatah I, Majumdar S, "Performance of CORBA-based client-server architectures", IEEE Transactions on Parallel and Distributed Systems, Vol.13, No.2, 2002, pp. 111-127

[10] Bohoris C, Pavlou G, Cruickshank H. "Using Mobile Agents for Network Performance Management", Proc. of the IFIP/IEEE Network Operations and Management Symposium (NOMS'00), USA, Hawaii, IEEE Computer Society Press, 2000, pp. 637-652

[11] Dittrich A, Rasmussen S, O'Sullivan D. "Co-existence of TMN and CORBA for service management", Proc. of the third International Symposium on Autonomous Decentralized Systems (ISADS '97), Berlin, IEEE Computer Society, 1997, pp. 35-42

Design and Implement of Autonomic Software Maturity Evaluation

Haitao Zhang[1] Huiqiang Wang[1] Honggang Liu[2]

1 College of Computer Science and Technology, Harbin Engineering University , Heilongjiang 150001, China
Email: zhanghaitao@hrbeu.edu.cn

2 Agricultural Bank of China Heilongjiang Branch, Harbin, Heilongjiang 150001, China
Email: lhg1970@sohu.com

Abstract

Based on the properties of autonomic computing, software quality characteristics and complexity of software, we propose a autonomic software multi-dimension evaluation model which is called ASMDEM. It has three levels: aim layer (autonomic maturity), characteristics layer (autonomic characteristics) and factor layer (quality factor). The model can evaluate the autonomic maturity of the complex software by AHP. At present there is no common autonomic evaluation criterion and method, so this research is important. In the future, we will optimize the model which ensures the valid and reliable.

Keywords: Autonomic Computing, Autonomic Software, QOS, Autonomic Maturity, Autonomic Evaluation

1 Introduction

Autonomic computing is generally considered to be a term first used by IBM in 2001 to describe computing systems that are said to be self-managing [1]. An autonomic systems must have the capability to self-configuration, self -optimization, self-healing and self-protection. Actually, autonomic computing is inspired by biology of neural system which can feel the internal and the external environments, and adjust its state or behaviors to the new condition automatically, so that it can be adapted to changes. There are some a certain amount of artificial intelligence may be required in autonomic computing system.

Currently, some research results of autonomic computing in commercial, academic and military area are being, there is some achievements, for example: The Trustworthy Computing of Microsoft, Self Adaptive Computing of HP, Proactive Computing of Intel and VALUMO of NEC; Sun's network allow Systems experiencing failure of a component could find components on the network with the required functionality that are still functioning and then reallocate resources to them the network with the required functionality that are still autonomously; Enterprise Edition Deadline policies can be used to make sure that project nearing the deadline receive a greater share of resources. Then share based policies consider the accumulated resource usage of a user, so that if a user "over-uses" resources, then the grid will lower the entitlement of that user to resources for a certain period of time [2]. In 2004, IBM proposed DB2 with autonomic computing tools. In 2005, University of Waterloo in Canada constructed infrastructure and applying platform of autonomic computing by intelligent agents.

In China, the representative results include: Institute of Computing Technology (ICT), Chinese Academy of Sciences researched autonomic computing from the angle of software engineering, defining an autonomic computing model [3]. National University of Defense and Technology designed a server system of computer gird resource backup, which has the autonomic characteristics in some extent. Northwestern Polytechnical University proposed a virtualized technology, [4] implementing the self managing of software and declining the whole cost of system.[5] input forward a common framework of SAN storage management system based on autonomic computing, implementing a autonomic FC-SAN storage system.

Autonomic computing concerns many domains, has a wide research scope. At present, it has a lot of achievement. But in initial stage, there is no general evaluation criteria and valid evaluation system. The research and material is few. But autonomic evaluation is very import. In this paper we propose a complex autonomic software multi-dimension evaluation model called ASMDEM, in the next section discussion on performance of the evaluate model.

2 Autonomic Software Multi-Dimension Model

2.1 Characteristics of autonomic software

According to the definition of autonomic computing, autonomic characteristics of software include four major and four minor characteristics that can be described as follows:

(1) *Self-configuration*: an autonomic computing system configures itself according to high-level goals by specifying what is desired, not necessarily how to accomplish it.

(2) *Self-optimization:* an autonomic computing system optimizes its use of resources. It may decide to initiate a change to the system proactively in an attempt to improve performance.

(3) *Self-healing*: an autonomic computing system detects and diagnoses problems. What kinds of problems are detected can be interpreted broadly: they can be as low-level as a bit-error in a memory chip (hardware failure) or as high-level as an erroneous entry in a directory service (software problem) [6].

(4) *Self-protection*: an autonomic system protects itself from malicious attacks but also from end users who inadvertently make software changes, e.g. by deleting an important file. The system autonomously tunes itself to achieve security, privacy and data protection. Thus, security lows: is an important aspect of self-protection, not just in software, but also in hardware (TCPA [7]).

(5) *Self-awareness*: it should be aware of its state and control its behavior, cooperating with other systems.

(6) *Open*: it must operate in a heterogeneous environment and is portable across multiple platforms.

(7) *Context-awareness*: it should be aware of execution environment and react to environment changes.

(8) *Anticipatory*: it must anticipate the optimized resources needed while keeping its complexity hidden.

Just like organism, autonomic software can be aware of the dynamic changing of environment and predict the potential danger. Once it finds abnormity, it will respond rapidly. The core of software running mechanism is the autonomic manager that forms an intelligent control circle and implements the autonomicity of executive activity, similar to the brain of organism.

2.2 Complexity of software

Complexity of software is the major cause of software errors, and the essence of software reliability is complexity. When the complexity exceeds a certain limit, the flaws or errors in software will rapidly ascend. Thus, autonomic characteristics of software are close to its complexity. When a simple system is integrated into a large-scale and complex system, its autonomic characteristics will be lost possibly. The autonomic characteristics should consider complexity of software adequately. Software complexity is the main reason of software fault, the essential of software reliability is complexity. When the complexity exceeds certain limitation, the fault and defect of software go up rapidly. Autonomicity is closed to complexity, so complexity should be fully considered in the assessment autonomic software autonomicity. The evaluation of software autonomicity include two aspects i) the assess autonomicity of system; ii) assess autonomicity of system environment.

Software complexity has two aspects: one is the complexity of software construction, another is the complexity of software environment. Considering of these elements, the complexity of software is decomposed in business domain complexity (BC), management complexity (MC) , and development complexity (DC).

BC, MC and DC, the three parts are not independent, which are influencing on each other. For example, the complexity of business domain induces the complexity of development, leading to the complexity of management at last; the counterforce of management will increase the complexity of management. So, the analysis of soft complexity should not divide the three parts, and the integrated evaluation results will be accurate.

2.3 QOS

Software autonomic characteristics belong to the category of software quality evaluation, so the autonomic maturity (autonomic degree) should be validated by software quality metrics. The mainstream software quality models can be divided into two types: hierarchy model and relational model. The famous hierarchy models include McCall model, Boehm model and ISO9126 quality model. The famous relational models include Perry model and Gillies model. This paper adopts 6 characteristics of ISO9126 quality model to evaluate the software autonomic characteristics, including reliability, efficiency, maintainability, usability, functionality and portability [8].

IBM has created an autonomic assessment software tool to measure the level of autonomic function against each of the six operational areas within and I/T environment. McCann et al. list a set of metrics by which the autonomic systems can be evaluated and compared, namely: Quality of Service (QoS), cost, granularity/flexibility, robustness, degree of autonomy, adaptivity, reaction time, sensitivity, and stabilization [9-10]. According to the research above, software complexity has great influence on autonomic characteristics, while the two methods mentioned don't deeply analyze the complexity. Autonomic evaluation based on software complexity will have actual meaning because the complexity is the frontier science in this century.

Software quality is composed by a set of quality factors which is composed by a group measure criteria. Every criterion can be analyzed and measured. Barry W. Boehm first proposed software quality levels model. MaCall proposed FCM three-levels Model: factor, criteria and metrics. ISO 9126 summarize quality into six characteristics which include a set of vice-characteristics, this quality model include three levels, the top level is SQRC, the middle level is SQDC, and the bottom level is SQMC.[9] Based on the ISO9126 model, constructing an autonomic evaluation model which has autonomic characteristics , quality factor and software complexity multi-dimension ,namely ASMDEM.

2.4 Architecture of ASMDEM

The model defines three levels. Aim layer is autonomic software maturity; Characteristics layer is describe autonomic software characteristics. Metrics layer is quality and complex factors of certain element. The constraint relations between the three parts are shown in Table 1.

Table 1 Quality Factors, Complexity Factors and Autonomic Characteristics

Autonomic Characteristics Factors	Major Characteristics				Minor Characteristics			
	Self configuring	Self ealing	Self optimizing	Self protecting	Self awareness	Open	Context aware	Anticipatory
Quality Factors								
Reliability		√		√				
Efficiency			√					√
Maintainability	√	√	√					√
Usability	√							
Functionality	√		√	√	√		√	
Portability	√					√		
Complexity factors								
BC				√			√	
MC					√	√		
DC	√	√	√	√	√	√	√	√

In the ASMDEM, autonomic characteristics acts as element set, the model divide factors set U according to characteristics. The principle of autonomic evaluation model can be described as follows:

(1) According to software feature and applying environment, implement the first class factor change and compose autonomic characteristic parameter.

(2) The characteristic parameters are divided into effective factors, implement the second class change. Effective factors has quality factors and complexity factors.

(3) Adopting AHP, if fuzzy relation matrix of lowest layer is given, the power matrix of all layer provided, the model can acquire comprehensive evaluation results of all layer.

(4) The final comprehensive evaluation maturity can be obtained.The hierarchy structure of model is showed in Fig1.

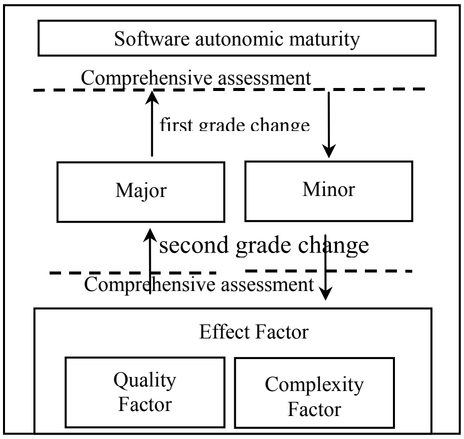


Figure 1 Autonomic Software Evaluation Construction

3 Evaluation Process

AHP is proposed by American operation researcher Satie in 1970s, AHP is a hierarchy weight decision-making analysis method applying network system theory and multi-objective comprehensive evaluation method. By dividing decision-making factors into goal, rule and scheme, AHP can implement qualitative and quantitative analysis [10]. Multi-level space model is a hierarchy structure satisfying the comprehensive evaluation metrics of complex software. The factors in lower hierarchy will be evaluated firstly, then the factors in higher hierarchy will be evaluated until the integrated evaluation result is got in the highest hierarchy [11].

Here, software complexity mainly includes business domain complexity, system development complexity and system management complexity. To software with different complexity, the definitions of autonomic characteristics are different. For example, if the system management complexity of software is higher, autonomic characteristics defined will be more and the autonomic maturity will be higher.

Each type of autonomic characteristic component is defined as an s-dimension vector [12], such as self-monitoring, self-configuration, self-optimization, self-healing, self-awareness, open, context-awareness, anticipatory, etc. The function of an autonomic characteristic component in software with different complexity is uncertainly same. Each autonomic characteristic component is decided by *n* relative quality factors. We suppose that T_l expresses autonomic maturity of the l^{th} complex software, S_{il} expresses the important degree (weight) of the i^{th} autonomic characteristic component in the l^{th} complex software, and S_i expresses autonomic value of the i^{th} autonomic characteristic component. We can calculate *T* in formula $T = \sum_{i=1}^{s} S_i S_{il}$ .In the same way, we suppose that N_i expresses autonomic value of the i^{th} quality factor and complexity factor, and N_{im} expresses the important degree of the i^{th} factor in the m^{th} autonomic characteristic component. The autonomic value S_m of the m^{th} autonomic characteristic component can be expressed as in formula $S_m = \sum_{i=1}^{n} N_i N_{im}$.

Complex software has the features of uncertainty and unstability. To quantitatively analyze the factors that are difficult to be quantified, multi-level fuzzy comprehensive evaluation method is adopted. Based on fuzzy math, this method applies the fuzzy relation composite theory to implement comprehensive evaluation.

4 Conclusions

At present, there is no standerd definition of what an Autonomic system is. So the evaluation or comparison about autonomicity is more difficult. Furthermore the very emergent nature of such systems adds further complexity to the evaluation of such systems. This paper is an attempt to build a useful model to evaluate the maturity of autonomic software. Though it is not quite accurate, because there is no the effective evaluation criteria. With the deeply research and full development, accurate and quantitative evaluation is more important. The model based on three-levels, that can evaluate the autonomic characteristics of complex software by the hierarchy method and fuzzy comprehensive evaluation method, this model calculates the autonomic maturity of each level in complex software. It can distribute multi factors of complex system into each level equally. The more the level is, the better the evaluation result will be. Because this model has generality function, it should be perfected and validated by simulation in the future.

References

[1] Kephart J. O., Chess D.M., "*The Vision of Autonomic Computing,"* Computer, IEEE, Volume 36,Issue 1, pp. 41-50, January 2003

[2] *How Sun™ Grid Engine, Enterprise Edition 5.3 Works*.URL: http://wwws.sun.com/software/gridware/sgeee53/wp-sgeee/wp-sgeee.pdf

[3] H.J.Zhang, Z.Z.Shi. "Software Engineering for Autonomic Computing," *Mini-Micro system*, 2007.6

[4] W.J.Li, Z.H.Li, "Application of Virtual Mechanism in High Availavility System Based on Autonomic Computing," Journal of Computer Appliacation, 2006.2

[5] J.Yao,J.W.Shu,W.M.Zheng, "Distributed storage cluster design for remote minoring based on storage area Network," Journal of Computer Science and Technology, 2007.4

[6] Dailey Paulson L., "*Computer System," Heal Thyself com*puter, IEEE, Volume 35, Issue 8, pp. 20-22, January 2002.8

[7] TCPA – The Trusted Computing Platform Alliance, URL: http://www.trustedcomputing.org /

[8] J.A. McCann and M.C., "Huebscher. Evaluation Issues Autonomic Computing," In PROCEEDINGS OF Grid and Cooperative Computing Workshops(GCC), 2004, pp. 597-608

[9] R.Barrent, P.P.Maglin, E.Kandogan, and J.Bailey, "Usable Autonomic Computing System: the Administrator's Perspective," In Proceedings of the First International Conference on Autonomic Computing, IEEE Computer Society,2004, pp. 18-25

[10] J.A.McCann, M.C.Huebscher, "Evaluation Issues in Autonomic Computing," *In Proc. of Grid and Cooperative Computing Workshops(GCC)*,2004, pp.597-608

[11] F.Z.Li,G.D.Hu, "Model of Network Security Comprehensive Evaluation Based on Analytic Hierarchy Process And Fuzzing Mathematics," *Ningxia Engineering Technology*, 2006.12

[12] X.Z.Wang,W.Z.Shi, and S.L.Wang , "Fuzzy Space Information Process," *Wuhan: Wuhan University Publisher*, 2003.10, pp.122-143

Making E-education Effective through Visual Communication Design

Xiaoyuan Ren[1 2] M ingxuan Chen[1]

1 School of Education, Jiangnan University, Wuxi, Jiangsu 214122, China

2 School of Design, Jiangnan University, Wuxi, Jiangsu 214122, China
Email: chaixy@126.com

Abstract

E-education is an important approach to provide the education to people regardless of time and place. With the development of the information technology, multimedia becomes a popular technique to advance traditional forms of E-education. To merge multiple media into E-education seamlessly, visual communication design becomes a necessity to pay more attention to. This paper depicts the current status of visual communication used in E-education area, and discusses its probable development trend in future.

Keywords: E-education, Visual Communication, Multimedia Education, Educational Multimedia

1 Introduction

E-education is an important approach to provide the education to people regardless of time and place. One of the oldest e-education is 'Distance Education' that is mainly based on the TV technology. It extends the classes offered at regular universities. Nowadays, with the rapid development of the information technology, many new approaches (e.g., Digital Slide Show, World Wide Web and E-mail etc) are appeared in e-education area. E-education comes to the multimedia education and educational multimedia phase. Multimedia data is becoming ubiquitous on any computing device from small, handheld devices like PDAs and mobile phones, to medium sized devices such as traditional desktop PCs and laptops, to very large devices such as public information systems with big screens. As mentioned in reference [1], today, almost every university claims to have a strategy to utilize the opportunities provided by the Internet or digital media in order to improve and advance traditional education. Together with the advent of the World Wide Web in the mid 1990s, the term e-learning was coined (along with other terms such as e-commerce or e-government) and created. Some people predicted dramatic changes in the educational environment or the end of traditional education in general. Like any other hype, disillusion started to spread a few years later and today, we face ubiquitous criticism claiming that many excellent ideas have been shadowed by the mass of mediocre and uninteresting work in this area. In fact, especially in relation to new media, it seems that the question how multimedia technology can really make learning more exploratory and enjoyable has still not been answered. For example, do the various web sites and lecture videos produced as part of the e-learning hype really exploit the full potential of multimedia-based teaching?

Visual communication is the conveyance of ideas and information in forms that can be read or looked upon. Primarily associated with two dimensional images, it includes: art, signs, photography, typography, drawing, graphic design, illustration, colors and electronic resources. Recent research in the field has focused on web design and graphically oriented usability. Graphic designers use methods of visual communication in their professional practice. Visual Communication Design is never stable but always in the process of evolving in response to current demands of technology, society and culture. It is concerned with the transfer of information, by embedding digital technologies and traditional media in relation to each other, for which methodological thinking, creativity, theoretical knowledge and practical skills are all essentially required. Through design, our visual surroundings are analytically and creatively processed so as to be rendered persuasive, descriptive or informative. As mentioned in reference [2], visual communication on the World Wide Web is perhaps the most important form of communication taking place when users are surfing the Internet. When experiencing the web, one uses the eyes as the primary sense and therefore the visual display of a website is important for the users understanding of the communication taking place.

Hence, the e-education in multimedia phase needs visual communication design to improve its effectiveness. This paper concludes the related work in this field and discusses how to employ visual communication design in e-education.

This paper is organized as follows. In section 2, The current status and related research of visual communication in e-education is discussed; in section 3, some key elements should be pay much attention to when employ visual communication design in e-education is discussed; section 4 concludes the paper and gives some preview for using visual communication design in e-education area.

2 Previous Related Work

The previous related works in this field can be mainly classified into two categories: related to e-education, Educational Multimedia and Multimedia education; and related to visual communication. There are some works related to the previous category as follows: Reference [3] focuses on slide-based presentation systems and presents a new categorization schema for educational presentation systems. The schema itself arose from their experiences and observation of work of other research groups using presentation systems in multimedia enabled classrooms. The basic criteria are illustrated using existing systems and real-world examples. The criteria have been extended and applied to distinguish the intentions of existing presentation systems. Reference [4] argues the current evolutionary changes in educational technology and pedagogy will be seen, 50 years from now, as revolutionary changes in the nature of higher education as a process and as an institution. We are in the process of moving: From: face-to-face courses using objectivist, teacher-centered pedagogy and offered by tens of thousands of local, regional, and national universities; To: online and hybrid courses using digital technologies to support constructivist, collaborative, student-centered pedagogy, offered by a few hundred “mega-universities” that operate on a global scale. Reference [1] depicts that the progress in multimedia capture, analysis, and delivery, combined with the rapid adoption of broadband communication, has resulted in educational multimedia systems that have advanced traditional forms of teaching and learning. In addition, new trends in multimedia technology, such as multimedia on handheld devices or advanced approaches for the automatic analysis of multimodal signals, offer novel and exciting opportunities for teaching and learning. However, the reference proposes that the question about how multimedia can really make education more exploratory and enjoyable is as yet unanswered, and we are just beginning to understand the real contribution of multimedia to education. They argue that e-education blended with information technology provides a motivation for the ACM Workshop on Educational Multimedia and Multimedia Education. Reference [5] introduces a panel for the ACM Workshop on Educational Multimedia and Multimedia Education (ACM EMME 2007) held in conjunction with ACM Multimedia 2007. Fleming Lampi et al [6] describe the design and implementation of an automatic cameraman for lecture recording. They argue that a major problem with traditional lecture recordings is that they tend to be boring for the students, especially if only the slides and the audio of the lecturer are pre-sent. In a first step, they determine the tasks a real cameraman would have, in particular with respect to liveliness of the video.

They then adapt these tasks to a computer system and show in de-tail how they can be implemented. In a second step, they describe how our algorithms support the virtual director system into which the automatic cameraman is integrated. They conclude that lecture recordings can be much more lively and interesting using their approach. Nalin Sharda [7] presents a framework for creating innovative New Media programs. He analyzes demand for traditional Computer Science and Information Technology (CS/IT) programs and concludes that demand has gone down in recent years, while new multimedia applications have grown exponentially over the same period. Additionally, teaching and learning paradigms are changing to cater for different learning styles. In today's global economy, creativity and innovation are tools for gaining completive advantage, and therefore, need to be included in educational programs. Developing innovative New Media programs is being recognized as a pathway to revitalize CS/IT programs. The course development framework presented in this reference incorporates new teaching and learning paradigms and provides the opportunity to inculcate creativity and innovation.

Besides those works related to e-education, Educational Multimedia and Multimedia education, there are lots of research focusing on visual communication as well: T. Kamae et al [8] learn that Digital motion video is getting important in telecommunications, computers and packaged media as well as in broadcasting. Requirements for digital motion video are versatile in each of those. However the evolution of digital technologies will make it possible to integrate these requirements into an integrated digital video. They focused on the technology of digital video to improve visual communication. Masami Kato et al [9] present a visual communication system for an apartment house, constructed of the Home Bus and a recently developed network, which can be used in an apartment complex. This system uses only twisted paired cable, and offers a variety of services. Delbert Hart thought that formal methods hold the promise for high dependability in the design of critical software. However, software engineers who employ formal methods need to communicate their design decisions to those who may not be in a position to acquire a full understanding of the formal notation being used. He concludes that visualizations might be able convey the required information precisely and reliably without the use of formal notation. Then, he presents a case study on how to employ visualization to communicate information about successive refinements involved in the formal derivation of a message router. The ultimate goal is to identify issues fundamental to this particular use of visualization and to outline a methodology which achieves effective visual communication without compromising formal reasoning. James L. Mohler [11] thought visual communication is vitally important to a variety of disciplines. Generally information is communicated in a way deemed appropriate for the discipline, governed by established rules based on prior research. However, when communicating to general audiences' enmasse, it is imperative that information be presented as simply as possible - the rules governing specialized scientific communication may or may not be applicable. In short, accessing information should be straightforward, allowing the user to efficiently gain the knowledge for which they are searching. Visual content that is presented should be clear, concise and accurate. With this in mind, this contribution discusses issues related to delivering spatially oriented data to the general populace. Communication through computer graphics should be for the masses as much as it is for individual disciplines. As such, he firstly examines issues related to visual communication for the mass population, then examines one such resource being used for visualization and communication at Purdue University. Janusz Konrad [12] overviews the state of visual communication at the end of the 20th century, discuss today's challenges, and outline some future directions. Yuyu LIU et al [13] propose a new visual communication system where eye contact is possible by using a virtual image. The virtual image is obtained by view synthesis with stereo matching from two real camera views. They developed a region based Dynamic Programming (DP) approach with improved matching cost, occlusion cost and vertical smoothness constraint. They also proposed a fast view interpolation method. To achieve real time performance, they developed a hardware system. Furthermore, to avoid the reordering problem in the foreground region, a view change approach with disorder detection was adopted.

Furthermore, some research related to employing visual communication design into multimedia design, TV program, news website etc began recently. Zhongge [14] thought that multimedia is the phenomenon and result led by Information Technology. Comparing with the traditional visual communication, the multimedia has its own special properties. The designer should pay much attention to the relationship between information and its appearance in order to do the better visual communication to people through the multimedia approach. Xiuhui Dong [15] introduces how to utilize the visual communication design in weather cast TV program. He proposes some aspects we should pay attention to when create a weather cast TV program, such as words' alignment, images' layout and the color's matches etc. Jing Yu et al [16] introduce some key elements to improve the quality of news-oriented website through visual communication designing.

As mentioned above, many works have been done related to e-education, Educational Multimedia and Multimedia education, visual communication, visual communication design for multimedia etc. However, few work related to using visual communication into e-education is introduced currently. This paper will discuss some key elements of using visual communication design in e-education in section 3.

3 Key Elements of Visual Communication in E-Education

Other than some general properties similar to visual communication in other fields, visual communication design in e-education has its own properties. In this paper, we discuss some of the key elements of visual communication in E-education as follows:

3.1 Trade off the requirements between different audiences

E-education is an important approach to provide the education to people regardless of time and place. Many different kinds of people probably become its audience. The visual communication design should provide special layout and content to different audience. For general purpose e-education, visual communication design should pay much attention to the trade off between teaching methods, communication approaches, user interface etc in order to meet any kinds of requirements.

3.2 Trade off the image quality, information content and network bandwidth

Generally, e-education is utilized in the campus. The visual communication design should prepare and provide educational multimedia with different image quality, information content, and screen resolution etc based on the network bandwidth.

3.3 Meet the requirements for different receiving and displaying devices

Besides desktop computer, slide player, TV set etc, there are many other receiving and displaying devices (e.g., PDAs, laptop, and big screen) can be used in e-education. The visual communication design should provide proper information content, playing format etc for different devices.

4 Our Practice

Currently, we provide a TV program of campus news suited to be played at TV set with high image quality. At the same time, we provide another replication of the program modified for internet player allow for the campus network bandwidth. The web page the video can be linked from is shown as Figure 1.

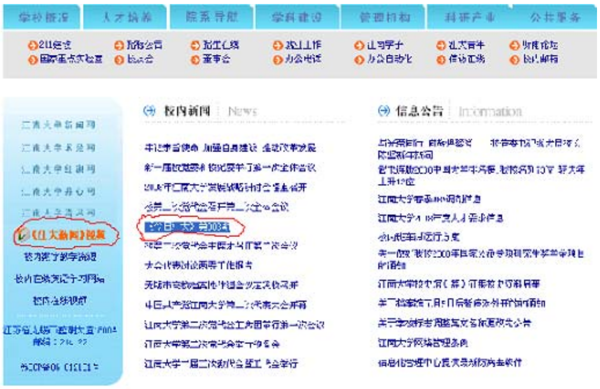


Figure 1 Website of our campus news video

5 Summary

In this paper, we take a glimpse at research work focused on e-education and visual communication. And discuss the trend to employ the visual communication design in e-education field when it steps into multimedia education and educational multimedia phase. Some key elements were introduced when using the visual communication design in e-education. Furthermore, our practice to use the visual communication design in e-education was introduced as an example.

References

[1] Gerald Friedland, Wolfgang Hürst and Lars Knipping, "Educational Multimedia Systems:The Past, the Present, and a Glimpse into the Future", the ACM Workshop on Educational Multimedia and Multimedia Education(EMME'07), 2007

[2] http://en.wikipedia.org/wiki/Visual_communication

[3] Georg Turban, "Categorization of Educational Presentation Systems", the ACM Workshop on Educational Multimedia and Multimedia Education(EMME'07), 2007

[4] STARR ROXANNE HILTZ and MURRAY TUROFF, "The Evolution of Online Learning and the Revolution in Higher Education", COMMUNICATIONS OF THE ACM, October 2005, Vol. 48, No. 10, pp.59-64

[5] Gerald Friedland, Wolfgang Hürst and Lars Knipping, "The Future of Multimedia Education and Educational Multimedia", the ACM Workshop on Educational Multimedia and Multimedia Education(EMME'07), 2007

[6] Fleming Lampi, Stephan Kopf, Manuel Benz, Wolfgang Effelsberg, "An Automatic Cameraman in a Lecture Recording System", the ACM Workshop on Educational Multimedia and Multimedia Education(EMME'07), 2007

[7] Nalin Sharda, "Creating Innovative New Media Programs: Need, Challenges, and Development Framework", the ACM Workshop on Educational Multimedia and Multimedia Education(EMME'07), 2007

[8] T. Kamae and T. Kishimoto, "Visual Communication and Digital Video", IEEE, 1993

[9] Masami Kato, Touru Kamimura, and Akio Kumagai, "VISUAL COMMUNICATION SYSTEM IN AN APARTMENT HOUSE USING ONLY TWISTED PAIRED CABLE", IEEE Transactions on Consumer Electronics, Vol. 40, No. 3, AUGUST 1994, pp418-427

[10] Delbert Hart, "Visual Communication of Formal Design Properties -- A Case Study", IEEE, 1994

[11] James L. Mohler, "Visual Communication for the Masses", IEEE, 2000

[12] Janusz Konrad, "Visual Communications of Tomorrow: Natural, Efficient, and Flexible", IEEE Communications Magazine, 2001

[13] Yuyu LIU, Keisuke YAMAOKA, Hiroyuki SATO*, Akira NAKAMURA, Yoshiaki IWAI, Ken-ichiro OOI, Weiguo WU and Takayuki YOSHIGAHARA, "Eye-Contact Visual Communication with Virtual View Synthesis", IEEE, 2004

[14] Zhong Ge, "Multimedia Design and Visual Communication", KAOSHI weekly, 2007,27

[15] Xiuhui Dong, "Visual Communication in weather cast TV program", 2007

[16] Jing Yu and Junhan, "Using Visual Communication to design news website"

A Parallel Platform for QPSO's High Performance Computing

Yinghui He   Wenbo Xu   Zhilei Chai

College of Information, Jiangnan University, Wuxi 214122
E-mail: heyinghui1983@163.com

Abstract

Quantum-behaved particle swarm optimization (QPSO)is an algorithm, which has good global optimization effects, and simple calculations. And it has inherent parallelism. Field Programmable Gate Array (FPGA) with a fine-grained parallel computing capabilities, is suitable as QPSO high computing platform. This paper designed a high-performance computing platform for QPSO, and realized in XILINX company's SPARTAN-III chips. The whole design structure of the system is modular, and use pipeline technology to optimize the system. By testing some reliable benchmark functions, the computing platform that can be effective in reducing the running time and improve QPSO practical value.

Keywords: QPSO algorithm, VHDL, FPGA

1 Introduction

QPSO which is derived based on the quantum theory, is a new model of particle swarm[8]. Because of the hidden inherent parallelism, it is easy to be realized. Compared to PSO[1], it does not have many parameters need to be adjusted, and for most of the issues, the general parameter settings can effectively complete the optimization, so QPSO algorithm is suitable for hardware implementation.

There are two ways to improve the QPSO algorithm's operating speed: one method is a software approach; another one is the hardware approach[10]. However, because of its sizes and costs, the QPSO algorithm which is realized by LAN or other parallel technologies has great limitations in its practical applications. By comparison, the QPSO algorithm that is realized by hardware can integrate the complicated algorithm onto a small chip.

FPGA[3] is a further development product, on the basis of PAL, GAL, EPLD and other programmable devices[9]. It is a semi-custom circuit as a application-specific integrated circuit(ASIC), not only solved the shortcomings of the custom circuits, overcame the limited number of the original programmable device gate circuits, provided a wealth of resources for realizing a more complex design.

2 Algorithm Thinking

QPSO also is a particle swarm evolutionary algorithm [2]. It using "Group" and "evolution" concepts, also based on the individual (particulate) fitness values to operating. QPSO look each individual as a particle with no weight and size in N-dimension searching space, and in a certain flight speed. The flight speed is dynamic adjusted by the individual and group's flying experience. Each particle represents a position that is in N-dimension space, towards the following two directions to adjust the particle's position:

(1) So far, the optimal position of each particle i is found:

(2) The optimal position of the particle swarm..

Each particle i contains the following information:

(1) $x_i = \left(x_{i1}, x_{i2}, \cdots, x_{id}\right)$:Particle's current position;

(2) $v_i = (v_{i1}, v_{i2}, \cdots, v_{id})$:Particle's current velocity;

(3) $p_i = (p_{i1}, p_{i2}, \cdots, p_{id})$:Represents the individual's best fitness value(pbest);

(4) $p_g = (p_{g1}, p_{g2}, \cdots, p_{gd})$:Represents the group's best fitness value (gbest);

QPSO particle's evolutionary expressions:

$$mbest = \frac{1}{M}\sum_{i=1}^{M} P_I = (\frac{1}{M}\sum_{i=1}^{M} P_{i1}, \cdots, \frac{1}{M}\sum_{i=1}^{M} P_{id}) \qquad (1)$$

$$P = \varphi * P_{id} + (1-\varphi) * P_{gd} \qquad \varphi = rand(0,1) \qquad (2)$$

$$x_{id} = p_{id} \pm \alpha * \left| mbest - x_{id} \right| * \ln(\frac{1}{u}) \qquad (3)$$

Here mbest is pbest's middle position, P is a random point between P_{id} and p_{gd} . α is called Contraction-Expansion Coefficient[7], which can be tuned to control the convergence speed of the algorithms.

QPSO algorithm's general processes:

1) Initialize particle swarm;

2) According expression (1) to computing mbest's value;

3) Computing the fitness values of each individual, and compare them to get P_{id} ;

4) Comparing each individual's P_{id} , and get p_{gd} ;

5) Update p_{gd} ;

6) For each dimension of the individual, according expression (2) to get a random point between P_{id} and p_{gd} ;

7) According expression (3) to get a new position ;

8) Repeat 2)-7) until meeting conditions, iterative ended.

3 Hardware Process

3.1 The overall structure of hardware

This paper realized 16 particles as a group. The particles are 3-dimension, and iterative 2000 times. According to the generic framework for population algorithms [4,5], the whole design use modular design structure. Overall, the population is divided into three independent parallel one-dimensional QPSO hardware structure, its internal uses the same modular design structure. There are no synchronization and communication between each independent module.

3.2 The parallelism and pipeline of the hardware

By analyzing the QPSO algorithm which is realized on software, we can see that: Comparing select, update particles, as well as computing fitness functions take the entire algorithm time 80% to 90%. If population size is a, iterative few is b, so these operations will be carried out a*b times. On a complex issue, a and b will be very high. So, improve the efficiency of the QPSO algorithm is to improve the efficiency of operations. Obviously, because of the limitations of the QPSO software algorithm(procedure can only be executed in order), there is not much improving space in this respect ; In contrast, when use the hardware to realize the QPSO, it can give full play to the characteristics of parallel algorithm. For example, in QPSO algorithm, if you want to provide a individual which is needed in comparing select operation, you need at least two fitness values. However, when use hardware to realize it, these two fitness values can compute at the same time, then the time of the comparing select operation can be cut by half. This paper, in each dimension makes the two pipelines evolving at the same time. Each pipeline is responsible for half of the population. There are no synchronizations and communications between the two pipelines. Because of the randomicity of the evolution, there may be an evolutionary pipeline has been completed, in the idle state, and another one has not yet completed. Although this method may sacrifice some efficiency, it saves much time in Communication and synchronization between two pipelines, and greatly reduce the workload of the programming. Figure 1 is a diagram of one-dimensional system modules.

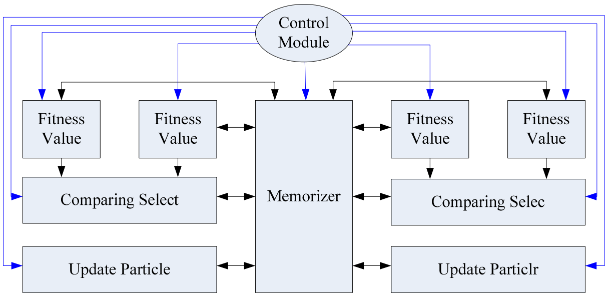


Figure 1 diagram of one-dimensional system modules

In figure 1, one-dimensional system has two parallel pipelines. Control module controls each operation, fitness value sends computing results to comparing module, then comparing module will send a shake hand signal to fitness value module after receiving the results to notifying the fitness value module to do next operation; Other modules are similar to the above operation.

3.3 Internal module design of one-dimensional system

3.3.1 Control Module

Control module can control the whole system by sending control signals. It sheds three control signals to memory, separately read, write and chip-select signal, at the same time it sends different signals to other modules. The control module use microprogramming[6], whose intention is to design complex control unit to achieve complicated instructions by simple hardware organ. Firstly, the write signal of control memory is effectual, data is read and sent to fitness value module to operation (here the sent signals to fitness value module is effective).Then, save the results to relevant address of memorizer, now the write signal is effective. Set the read signal effective, read data, and send the data to compare module by multi-channel selector, send the result back to memorizer after comparing. Finally send the data in memorizer to update particle module to operation, the operated result will be sent back to memorizer. The process only is once iteration. Performing the jump instruction then the iteration will be carried on. The following VHDL codes are some jump instructions:

```
...
if   MPC_en ='0'   then
if   count<2048   then
MPC <= NextMPC;
if   MPC="00011101"   then
count:=count+1;
end if;
else   MPC<="11111110";
end   if;
end   if;
…
```

MPC_en is enable signal of controller, count represents counter, MPC is microinstruction, NextMPC is next microinstruction.

3.3.2 Memorizer

Memorizer includes read-only memory (ROM), random read-write memory (RAM), and many registers (REG). Random numbers and sixteen particles initial positions are mainly stored in ROM. RAM is used to story the final results and send data to each module. REG is data transfer station, it can make clock cycle shorter and speed up the process speed.

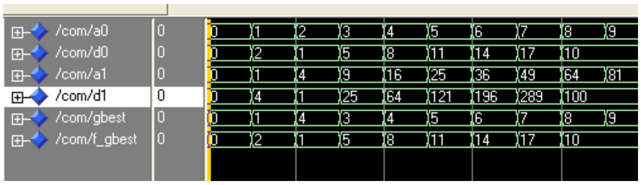
3.3.3 Fitness Value Module

Fitness value module gets data from memorizer, then computes fitness value according to the standard function. After computing, send the results back to memorizer in current address. These operations are all controlled by controller.

Xilinx is offering a lot of nuclear function, use of the nuclear can facilitate the construction of different standards function.

3.3.4 Comparing Select Module

Comparing select module compares data that is from memorizer and fitness value module. Input four data, two come from fitness value module (f_pbest), the other two come from memorizer(pbest),then output f_gbest and gbest after comparing. Figure 2 is simulating result of comparing select module:



```
/com/a0       0   0  1  2  3   4   5    6    7    8    9
/com/d0       0   0  2  1  5   8   11   14   17   10
/com/a1       0   0  1  4  9   16  25   36   49   64   81
/com/d1       0   0  4  1  25  64  121  196  289  100
/com/gbest    0   0  1  4  3   4   5    6    7    8    9
/com/f_gbest  0   0  2  1  5   8   11   14   17   10
```

Figure 2 simulate result of comparing select module

3.3.5 Update Particle Module

This module requires a lot of computing, mainly updates the particles positions. There is a nesting cycle in the operation of updating particle position. External cycle depends on jump instructions which are in control module, and internal cycle depends on adding counter in the module. Counter start counting from 0, when completed once every computing, counters count plus 1, 16 particles are stopped counting after 16 operations, then outputs the result. This module uses 30 16-bit adders, two 32-bit adder, three 32 subtraction, 12 32-bit registers, two 32-bit comparator.

4 Hardware Implementation and Testing

The whole design uses VHDL describing, and realized on FPGA chips of Xilinx's SPARTAN-III. To verify the correctness and performance of the system, in this paper, the two benchmark functions are tested.

$$f_1(x)=\sum_{i=1}^{3} x_i^2 \qquad 50\le x\le 100 \qquad (4)$$

$$f_2(x)=\sum_{i=1}^{3}100(x_i^2-x_{i+1})^2+(1-x_i)^2 \quad 15\le x\le 30 \qquad (5)$$

Hardware implementation for the 40 MHz system clock, 16 three-dimensional particles, 2000 times are iterative, two benchmark functions' results of the simulation are respectively shown in Figure 3 and Figure 4

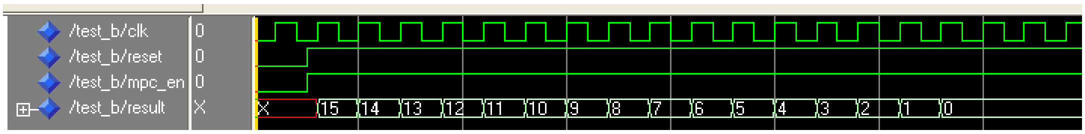


Figure 3 The simulating result of the first function

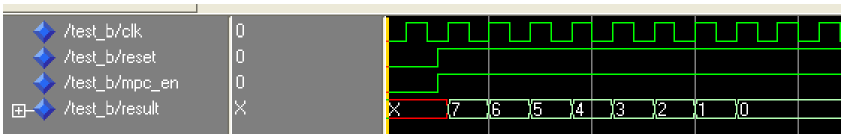


Figure 4 The simulating result of the second function

Testing results are shown in Table 1

Table 1 Testing results

Functions	Implementation	Fitness value	Cpu time	The number of clock cycles
$f_1(x)$	software	6.69E-5	1.27s	3048
	hardware	0	20.34ms	814
$f_2(x)$	software	4.74E-4	2.44s	5856
	hardware	0	23.53ms	941

The software program is written in MATLAB, running on a frequency of 2.4G Pentium PC.

From the testing results we can see that, the number of software system clock cycles is much more than the number of hardware system clock cycles, and its running time is far more than hardware.

5 Concluding Remarks

This paper uses FPGA as computing platform for the realization of QPSO algorithm. The whole design with FPGA characteristics, realized a parallel evolution not only in the overall structure, at the same time, also in the internal parallel processing, and using a pipeline structure, which greatly enhance the speed of the algorithm. Testing results show that the hardware implementation of QPSO algorithm can effectively shorten the running time, enhance the possibility for real-time applications.

References

[1] J. Kennedy, R. C. Eberhart, "Particle Swarm Optimization," Proc. IEEE Int'l Conference on Neural Networks, IV. Piscataway, NJ: IEEE Service Center, 1995, pp. 1942-1948

[2] J.Sun,B.Feng and W.Xu, "Particle Swarm Optimization with Particles Having Quantum Behavior,"IEEE Proc. of Congress on Evolutionary Computation, 2004

[3] C.Wang, X. Xue,FPGA/CPLD Design Tools, POSTS & TELECOM PRESS,2005.1

[4] John Newborough ,Susan Stepney. A Generic Framework for Population-Based Algorithms,Implemented on Multiple FPGAs. C. Jacob et al. (Eds.): ICARIS 2005, LNCS 3627,2005, pp. 43–55

[5] Xilinx, Inc., XACT Development System Reference Guide, Jan 1993

[6] David A .Patterson, John L.Hennessy, Write. W. Zheng , Translate, Computer Organization and Design-The Hardware/Software Interface, Third Edition, China Machine Press, 2007.4

[7] J.Sun,B.Feng and W.Xu, “Adaptive Parameter Control for Quantum-behaved Particle Swarm Optimization on Individual Level,” IEEE Proc. of Conference on Systems, Man and Cybernetics,2005

[8] J.Sun, B.Feng and W.Xu, “A Global Search Strategy of Quantum-Behaved Particle Swarm Optimization ,”IEEE Proc. of Conference on Cybernetics and Intelligent System, 2004

[9] Lynn Abbott, Peter M. Athanas, Luna Chen, and Robert L. Elliot. Finding lines and building pyramids with Splash 2, Proceedings of the IEEE Workshop on FPGAs for Custom Computing Machines, p155-163, Apr 1994

[10] Stocia, D. Keymeulen, V. Duong, and C. Salazar-Lazaro. Automatic synthesis and fault-tolerant experiments on an evolvable hardware paltform. IEEE Aerospace Conference Porceedings, Vol. 5, 2000, Pags 465-471

De-noising of THz Image based on Wavelet Threshold Methods

Wenquan Liu[1]   Shuangchen Ruan[2]

1 College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, Guangdong 518060, China

Email: liuwqbz @163.com

2 College of Electronic Science and Technology, Shenzhen University, Shenzhen, Guangdong 518060, China

Email：scruan@szu.edu.cn

Abstract

Terahertz (THz) wave imaging wave imaging has attracted considerable attention since its first demonstration by Hu and Nuss[1]. Noisy THz images are often a problem especially at poor visibility. Therefore, image de-noising could improve the quality of the THz images. Wavelet image de-noising method becomes an important method for image processing. With the MATLAB Toolbox wavelet analysis algorithm, the basic principle of wavelet threshold de-noising and the application of the three kinds of wavelet threshold de-noising methods to terahertz (THz) wave images are presented. Experiment results have shown that it is better than the traditional methods. And the further development of algorithm is pointed out.

Keywords：THz Image; De-noising; Image processing; Wavelet Threshold; Wavelet Transform

1 Introduction

Terahertz (THz) wave imaging is attractive due to its potential applications in diverse areas such as packaging inspection, quality control of plastic parts, chemical composition analysis, and biomedical diagnostics[2]. However, terahertz wave images are often corrupted by noise in its acquisition and transmission. Improving the signal-to-noise ratio of the image by removing noise from the original image is still a challenging problem for researchers. Wavelet de-nosing has become more and more popular for image de-noising, because additive noise can be removed while preserving important features of the image. As shown by Donoho and Johnstone’s threshold the coefficients is the most important task in wavelet de-noising [3-6]. Thus this paper has used three kinds of wavelet threshold de-nosing methods to de-noise THz images as following: the de-noising mandatory method, the default threshold de-noising method and the given soft or hard threshold de-noising method.

This paper is structured as follows. Firstly Section 2 provides the review on discrete dyadic wavelet transform, and then Section 3 gives the strategy of de-noising of THz image by wavelet threshold methods. Section 4 presents and discusses the results of the experiments. Finally the conclusions are given in the Section 5.

2 Discrete dyadic wavelet transform

A wavelet is a family of functions derived from a basis function $\psi(x)$ defined in terms of two parameters: a, dilation (scale) and b, translation (time).And the definition of the function $f(x)$ ’s CWT (Continuous Wavelet Transform) is

$$WT_f(a,b)=|a|^{-\frac{1}{2}}\int_{-\infty}^{\infty} f(x)\psi^*(\frac{x-b}{a})dx \qquad （1）$$

Where $\psi(x)$ is the basis wavelet function, which meets the conditions: $\psi(x)\in L^2(R)$, $\int_{-\infty}^{\infty}\psi(x)dx=0$ .and

$\psi^*(\frac{x-b}{a})$ is the conjugated function of $\psi(\frac{x-b}{a})$.

In order to adapt to digital signal processing, the continuous wavelet transform is needed to get discrete. The operation flow of discrete process can be summarized as follows: firstly, the scaling factor *a* should be discrete for dyadic wavelet transform, and secondly the translation factor *b* also should be discrete for the discrete wavelet transform.

If ψ is the basis dyadic wavelet function, the scaling factor *a* can be ordered to be equal to 2^k*, or a* = 2^k. After that, the translation factor *b* can be ordered to be equal to *n*, or *b*=*n*, the discrete dyadic wavelet transform of $f(x)$ at scale 2^k is

$$W_{2^k} f(n) = \frac{1}{\sqrt{2^k}} \int_{-\infty}^{\infty} f(x) \psi^*(\frac{x-n}{2^k}) dx \qquad （2）$$

The wavelet can be designed to display local maxima for sharp variations of *f*(*x*), which is important for analyzing the properties of signals and images. The most effective way to implementation of discrete wavelet transform is to use filters. This method, developed by Mallat in 1988, is called Mallat algorithm[7]. This method, known as dual-channel sub-band coding in the field of digital signal processing is actually a signal decomposition method.

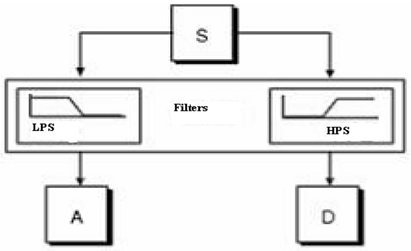


Figure1 signal decomposition with dual-channel filters

As shown in Fig. 1, S, the original input signal, passed through two complementary filters, have decomposed into two signals A and D, A signal is the approximation of signal, D signal that the details of signal. In many applications, the low-frequency signal is the most important part, while the high-frequency signal is an "additive" part.

Evidently, DWT can be expressed by a tree consisting of the low-pass filters and the high-pass filters. The decomposition of original signal through dual-channel filters called the first level decomposition. The signal decomposition can be in iterative processing, that is multi-level decomposition. If the signals in the high-frequency component is no longer in decomposition after the first level decomposition, and the low-frequency component is in the consecutive decomposition, the low-frequency component can make up of a big tree called wavelet decomposition tree as shown in Figure2.

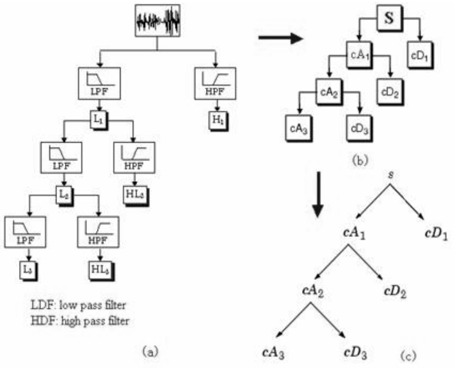


Figure2 wavelet decomposition tree

3 Wavelet denoising

Actually, image noise is generally mixed with all kinds of noise. According to the distribution of noise obeying their classification, the noise can be divided into Gaussian noise, Poisson noise, Salt and Pepper Noise and grain noise. Most approaches assume that the images are corrupted by independent and identically distributed (I.I.D.) zero mean, white Gaussian noise$\varepsilon$with standard deviationσ.Generally a noisy image can be modeled by (3) as follows[8].

$$g(x, y)=f(x, y)+\varepsilon(x, y) \qquad （3）$$

where $g(x, y)$ is noisy image, $f(x ,y)$ is signal image and $\varepsilon(x, y)$ is noise.

Wavelet de-noising method can be roughly divided into wavelet threshold method, projection method and related methodologies. Wavelet threshold method is the most extensive study of the method. It is based on the principle of multi-resolution analysis. Because the signal and the noise have different singularity, the signal wavelet transform modulus maxima will increase as the scale increases. However, the noise is to the contrary. It can effectively remove the noise using the threshold processing through the scale transform.

The main idea of wavelet de-noising is to transform the data into the wavelet basis, where the large coefficients mainly contain the useful information and the smaller ones represent noise. By suitably modifying the coefficients in the new basis, noise can be directly removed from the data. More details on the wavelet-based de-noising techniques can be found in [9].The general de-noising procedure involves three steps:

1)Decomposition: compute the wavelet decomposition of the original data;

2)Threshold wavelet coefficients: select a threshold and apply threshold to the wavelet coefficients;

3)Reconstruct: compute wavelet reconstruction using the modified wavelet coefficients.

Following above three steps, the flow of wavelet de-noising of the image is shown as Fig. 3. The three main steps should be divided into four parts: Wavelet transform, Threshold processing, and Reverse Wavelet transform, and Image reconstruction. And Threshold processing can be used to remove the noise from the image.

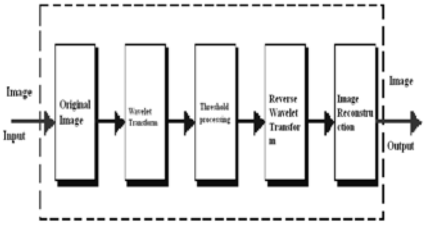


Figure3 the Sketch of Wavelet de-noising of the image

4 Experimental results and discussion

Based on the above analysis, the selection of wavelet de-noising threshold is the most critical. It can be illustrated by Fig. 4. The reconstructed image can be observed that the image quality will change as the threshold value changes. With the increase of threshold values, the image quality has also decreased.

If the threshold value is too small, the variance will get too large and image will be less smoothing. Vice versa. In this paper, with the MATLAB Toolbox wavelet analysis algorithm, it can de-noise the THz image using the de-noising mandatory method, the default threshold de-noising method and the given soft or hard threshold de-noising method.

De-noising mandatory method is that the high-frequency part of the wavelet decomposition of the image is eliminated, and then the signal is reconstructed. This method is simple and the signal will be relatively smooth after de-noising, but it can easily lose the useful signal.

The default threshold de-noising use ‘ddencmp’ function in the MATLAB Toolbox to make a signal default threshold value, and then use ‘wdencmp’ function(also in the MATLAB Toolbox) to de-noise.

The given soft or hard threshold de-noising method is that the threshold value can be obtained through experience formula during the process of the actual noise elimination. So this threshold has more credibility than the default threshold.

For the experimentation it have used a 394×393 grayscale THz image, shown in Figure4.

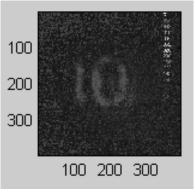


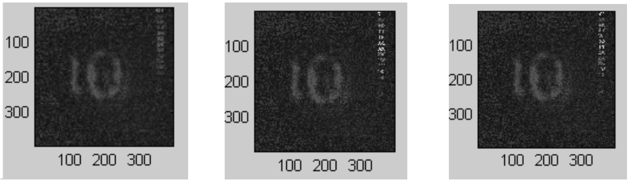
Figure4 Original THz image

By the three de-noising methods with MATLAB 6.5 programming, the result of experimentation is illustrated by Figure5.

The performance of the estimators was measured by using the classical signal to noise ratio, defined as (4) and (5).

$$Ems = \sqrt{\frac{\sum_{i=1}^{M}\sum_{j=1}^{N}[I_{i,j} - I_{i,j}^{\wedge}]^2}{M \times N}} \qquad (4)$$

$$Snr = 20\log_{10}\frac{\sum_{i=1}^{M}\sum^{N} I_{i,j}^{2}}{\sum_{i=1}^{M}\sum^{N}[I_{i,j} - I_{i,j}^{\wedge}]^2} \qquad (5)$$



(a) (b) (c)

Figure5 the results of de-noising image. (a)De-noising mandatory method (Snr =30.945);(b)The default threshold de-noising (Snr=29.732);(c)The given soft or hard threshold de-noising (Snr=29.736)

Among the results, the performance of the de-noising methods can be measured. The De-noising mandatory method has higher Snr than the other two methods, but evidently the quality of image has decreased. The default threshold de-noising and the given soft or hard threshold de-noising method has common Snr. But the given soft or hard threshold de-noising threshold has more credibility than the default threshold.

5 Conclusions

This paper presents three Wavelet Image De-noising methods using MATLAB toolbox algorithm, while classical algorithm for image de-noising depended on noise type and size. Wavelet de-noising is applicable to many types of noise. Experiment results have shown that it is better than the traditional methods. Wavelet transform in the image processing for de-noising will be more and more advantages.

It is an effective method for de-noising THz images with GCV (Generalized Cross Validation) [10]selected threshold value close to an ideal threshold under certain conditions, which only relies on input and output data without estimating the noise energy. This algorithm is the next step in the further development.

References

[1] B. B. Hu and M. C. Nuss, "Imaging with terahertz waves," *Opt. Lett.*, vol.20, 1995, pp. 1716–1718

[2] D. Mittleman, R. Jacobsen, and M. C. Nuss, "T-Ray imaging," *IEEE J.Select. Topics Quantum Electron.*, vol. 2, 1996, pp. 679–692

[3] D.L. Donoho and I.M. Johnstone, "Ideal spatial adaption via wavelet shrinkage" *Biometrica*, Vol. 81, 1994, pp. 425-455

[4] D.L. Donoho and I.M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage" *Journal of American Statistical Association*, Vol. 90, No.432, 1995,pp. 1200-1224

[5] D.L. Donoho and I.M. Johnstone, "Wavelet shrinkage: Asymptopia?" *Journal of the Royal Statistical Society Series B 57*, 1995, pp. 301-369

[6] D.L. Donoho. "Wavelet shrinkage and W.V.D.: A 10-minute tour" *Proceedings International Conference on Wavelets and Applications*,Toulouse, France , 1992

[7] S. Mallat and S. Zhong: "Characterization of Signals from Multiscale Edges"*IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, No. 7, 1992,pp. 710-732

[8] Dietmar Wippig, Bernd Klauer and Hans Christoph Zeidle, "Denoising of Infrared Images by Wavelet Thresholding" *Advances in Computer, Information, and Systems Sciences, and Engineering*, 2006,pp.103–108

[9] D.L.Donoho, "De-Noising by Soft-Thresholding" *IEEE Transactions on Information Theory*, Vol. 41, No. 3, 1995, pp. 613-627

[10] J.T.Kent, M..Mohammadzadeh, "Global Optimization of the Generalized Cross Validation Criterion" *Statistics and Computing,* No.10 ,2000 ,pp.231-236

Image Pre-processing of Harmful Stored Grain Insects Based on Mathematics Morphology

Long Zhou[1] Yi Mou[1] Mianyun Chen[2]

1 Wuhan Polytechnic University, Wuhan, 430023, P. R. China
Email：zhoulong@whpu.edu.cn

2 Huazhong University of Science and Technology, Wuhan, 430074, P. R. China

Abstract

The detection method of pests in stored grain is always investigated. The method of based on image recognition is often discussed. The importance of image pre-processing based on mathematical morphology in the grain insect image processing is introduced. The functions of Matlab image processing are used in the image processing. The image pre-processing and analysis techniques based on mathematical morphology are realized, and gain the goal of image denoising and smoothening, from which concusions are drawn. The examples show that the method can obtain better edge detection.

Keywords：image of grain insect; mathematical morphology; image pre-processing

1 Introduction

Our country is the biggest grain production, preserves and consumption nation in the world, No doubt, it is of great significance to make good grain preserve. making the grain preserve. 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]. But much grain suffers heavy loss because of harmful stored grain insects. At present, at home and abroad, the examination aspect of harmful stored grain insects mainly has the following several methods: spots-check the law[2]], tempts the collection law[3], near-infrared method[4]], the sound signal law and the pattern recognition and so on[5-6]. With rapid development of computer technology, pattern recognition, intelligent examination based on the machine vision, pattern recognition examination method becomes the main development direction of grain insect recognition. According to the material demonstrated, if doing the harmful stored grain insects on-line examination well, taking the reasonable preventing and controlling measure promptly, We can reduce the grain preserve loss by 0.05% again, and recall the economic loss 250 million yuan every year for the country, this numeral is extremely considerable[7]. Therefore, researching the harmful stored grain insects on-line examination system, not only has the important academic value, but also has the broad application prospect.

2 Image pre-processing based on mathematics morphology

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.

Mathematics morphology is a subject established on the set theory, its basic thought is using the certain shape structural element to measure and withdraw the corresponding shape in the image, by achieves the goal of analysis and recognition the image. Mathematics morphology has direct-viewing on simplicity and mathematics rigorousness in digital signal processing, and has the unique advantage in the description digital signal shape characteristic, its algorithm has natural parallel realization structure, and is used widespread application in recent years[8]. Thus, using mathematics morphology in the image pre-processing, one can effectively eliminate all kinds of disturbance and noises, and also can retain plenty of important information, such as the outline and the edge of the image and so on, connecting all small interrupted of the goal and making the image view better[9].

Mathematics morphology is developed based on the set theory. The viewpoint of mathematics morphology based on set is very important, this has also decided that its operation must be defined by the set operation (sum aggregate, occurs together, complementary set), all images must be transformed into a set in a reasonable way. Mathematics morphology uses the certain structural and characteristic structure element (actually a set) to measure the shape in the image, then solves the problem[10]. Considering from the set theory, mathematics morphology contains a set operational method transformed from one to another. The purpose of this kind of transformation is to find the specific set structure of the primary set, but the transformed set contains information of this kind of specific structure. Certainly, this kind of transformation is depended on the structural element with the certain characteristic, therefore we obtain the result related to the structural element characteristics[11-12].

2.1 The of basic operations in morphology

Mathematics morphology has defined four kind of basic operations, namely the inflation, the corrosion, opens and closes. Through these basic operations we can also infer and combine every kind of mathematics morphology practical algorithm. In two values morphology, operational object is a set, usually giving an image set and a structural element set, using the structure to deal with the image. But in actual operations, the two sets do not be regarded as mutually coordinated.

Supposing A is the image matrix, B is the structural element matrix, assigning a central image element to each structural element matrix, it is the reference point of structural element and morphology operation, usually expressing the image element of user's expectation, then the central image element is defined as :

floor ((size (SE) +1) /2)

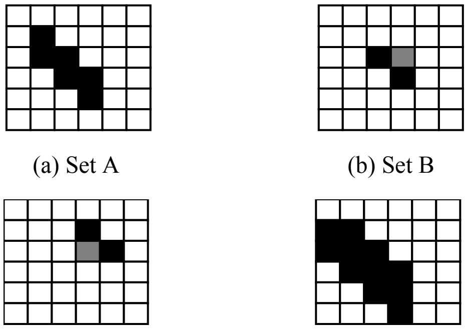
Where SE is a matrix which is defined with the structural element, for example: SE= $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

2.1.1 inflation

Inflation operator is $\oplus$, A inflates with B writing as $A \oplus B$, defined as:

$$A \oplus B= \{x|[(\hat{B})_x \cap A] \neq \Phi\}; \qquad (1)$$

From the figure 1 we may see inflation operation expands the original region.


(a) Set A (b) Set B

(c) Reflection of B (d)The result of inflation

Figure1 Inflation operation process

2.1.2 Corrosions

The corrosion operator is Θ, A corrodes with B writing as $A \Theta B$, defined as:

$$A \Theta B=\{x|(\hat{B})_x \subseteq A\}; \qquad (2)$$

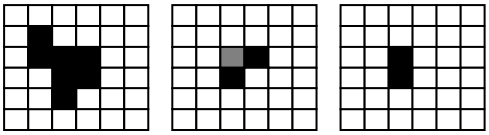

(a) Set A (b) Set B (c) The result of corrosion

Figure2 Corrosion operation process

2.1.3 Opened and Closed

Because the inflation and the corrosion is not an inverse operation mutually, we may unite them.

Opens is corroding the image first, then inflating its result. The operator of opens is $\circ$, A opens with B writing as $A \circ B$, defined

$$A \circ B = (A \ominus B) \oplus B; \qquad (3)$$

Opened A with B is selecting some spots which matched with B in A, these spots might be obtained by translating the structural element B which contained completely in A.

Closed is inflating the image first, then corrodes its result. The closed operator is ·, A closed with B writing as A·B, defined as:

$$A \cdot B = (A \oplus B) \ominus B; \qquad (4)$$

Closed A with B is the spot set which satisfies the following condition, namely when this spot may be covered with the mapping and displacement structural element, the occurs together of A with the mapping and displacement structural element is not zero.

2.2 The application in image pre-processing of Mathematics morphology

These two kinds of operations: opens and closes, all can eliminate specific image detail which is smaller than the structural element, at the same time, guaranteeing that does not have the overall situation geometry distortion. Opens can filter out sudden thorn which is smaller than the structural element, shut off the tall and slender joining, then separate them. Closed can make up the gap or the hole which is smaller than the structural element small.

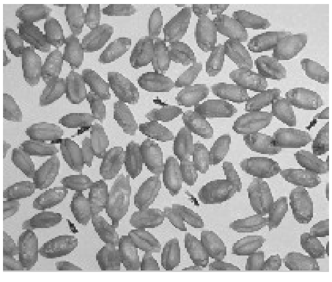


Figure3 inflation processing result

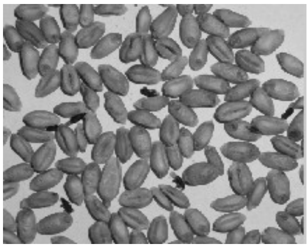


Figure4 corrosion processing result

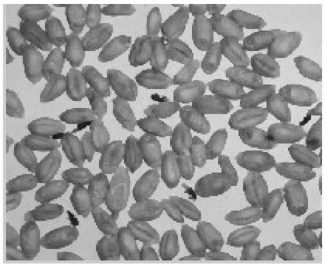


Figure5 open and close result

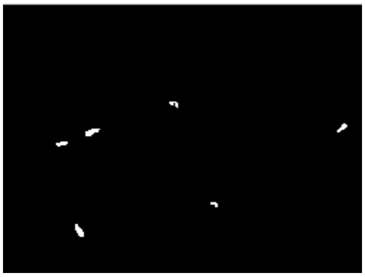


Figure 6 the division processing result

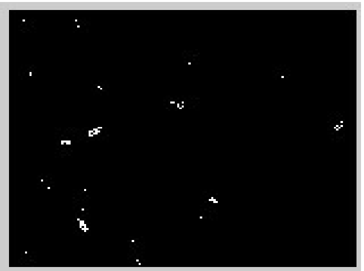


Figure7 the division result of the definite and threshold

Figure 5 is the image which inflates the original Figure 3, Figure 6 is the image which corrodes the original Figure 3, Figure 7 is the image which opens and closes the original Figure 3, Figure 7 is the image which divides Figure 6, The realization source program and the correlation function is from MATLAB..

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. This article begins with the mathematics morphology basic thought, and applies it in image pre-processing of harmful stored the grain insect, after the research, we can find this method is a simple, practical, and effective image pre-processing method, and provides a good foundation for the following image analysis such as image division, the characteristic formed, withdraws, compression and so on.

Acknowledgements

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] The national grain bureau "15" the grain profession science and technology development plans. 2001.1

[2] Wilkin D.R.,Fleural-Lessard F.,The detection of insects in grain using conventional samplings spears [C].Proc,5th Int.Wkg,Conf.OnStored-ProductProtection, Bordeaux, France. 1990

[3] Yao Wei ,the grain insect trap examination engineering research and applies, the Chinese cooking oil academic society first session academic annual meeting paper anthology (preserve specialized volume). 2000

[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

[5] Wick K.W.,Webb J.C.,Weaver B.A.Et al.,Sound detection of stored-product insects that feed insede kernels of grain [J].Econ Entomol.1988

[6] Xu Fang, Qiu Daoyin, pattern recognition in granary harmful insect examination aspect applied research, Zhengzhou engineering college journal. 2001.2

[7] Wan Zhengqun, the current our country science guarantees the grain,The preserve of grain. 1996.2

[8] Mr. Dai, Mathematics morphology in imagery processing application ,Foshan science and technology university (natural sciences version). 1998.16

[9] Shen Tingzhi, Fang Ziwen, digital image processing and the pattern cognisition, Beijing Technology College publishing house. 1998

[10] Hu Xiaofeng, Visual C++/The MATLAB imagery processing and the recognition practical case select,The people's posts and telecommunications publishing house. 2004.9

[11] Pal S K,King Pa.On edge detection of X-ray images using fuzzy set.IEEE Trans. Pattern Anal. Machine Intell., 1983, 5(1): 69-77

[12] I. Y. Zayas, Y. Inna, Detection of Insects in Bulk Wheat Samples with Machine Vision[J].Trans. Of the ASES, 1998(3): 883-888

A Novel Semi-fuzzy Clustering Algorithm with Application in Image Texture Segmentation*

Suqun Cao[1 2] Shitong Wang[1] Yunfeng Bo[2] Xiaofeng Chen[1]

1 School of Information, Jiangnan University, Wuxi, 214122, China

2 Department of Mechanical Engineering, Huaiyin Institute of Technology Huaian 223001, China
Email: caosuqun@126.com

Abstract

Image texture segmentation is one of the important branches in image pattern recognition, which provides usefulness in many applications. Until now, how to find an effective way for accomplishing texture segmentation in practical applications is still a major task. In this paper, a novel semi-fuzzy clustering algorithm is presented. The basic idea is to extend Fisher discrimination method with fuzzy theory and define fuzzy Fisher criterion as the objective function of the proposed clustering algorithm. By iteratively optimizing this function, the final clustering results are obtained. Experimental results on its application in image texture segmentation show that the proposed algorithm has better performance when noise is present than the standard algorithm.

Keywords: Fisher Criterion; Optimal Discriminant Vector; Semi-fuzzy clustering; Fuzzy C-means; Image Texture Segmentation

1 Introduction

Cluster analysis is an important pattern recognition tool used in diverse fields such as image process, computer vision and data mining. It aims to cluster a data set into most similar groups in the same cluster and most dissimilar groups in different clusters. Many clustering methods have been proposed and exhibited their extensive applications [1-7]. K-means and Fuzzy c-means (FCM) [1, 2] are two of the most well known methods. In general, FCM outperforms K-means due to the introduction of the fuzzy concept. Most of these methods like FCM are essentially rooted at the within-cluster scatter matrix as a compactness measure, which means that the assumption that the clusters are hyperspheroidal is taken. In fact, real data sets seldom accommodate such an assumption. Except for the compactness measure, the concept of the separation measure such as the between-cluster scatter matrix should also be involved in the design of clustering methods such that well separated cluster can be obtained.

In this paper, we extend Fisher linear discriminant to its fuzzified version and define the fuzzy Fisher criterion function. Then a novel semi-fuzzy clustering method is presented. Compared with FCM, the proposed algorithm has the following characteristics: (1). It directly uses the fuzzy FLD as its objective function, therefore, it is an unsupervised fuzzy partition clustering method. Its objective function integrates the fuzzy between-cluster scatter matrix well with the fuzzy within-cluster scatter matrix. (2). It incorporates the discriminating vector into its update equations such that the obtained update equations do not take commonly-used FCM-like forms. (3). It is more robust to noise and outliers than FCM.

The remainder of this paper is organized as follows. Section 2 introduces some concepts of Fisher criterion and Fisher linear discriminant. Section 3 introduces fuzzy Fisher criterion and then a novel semi-fuzzy

* This work was supported by Youth Foundation of Huaiyin Institute of Technology(No.351707039)

clustering algorithm FFC is presented. Section 4 performs some experiments on image texture segmentation.

2 Fisher Criterion and Fisher Linear Discriminant

Given c pattern classes $\mathrm{X}^{(i)}=\left[x_i^1,x_i^2,...,x_i^{N_i}\right]$ in the pattern set which contains N d-dimensional patterns, where i =1, 2, … , c, N_i is the number of all the patterns in the ith class, thus, $N=N_1+N_2+\cdots+N_c$. The between-class scatter matrix S_b and the within-class scatter matrix S_w are determined by the following formulae:

$$S_b=\sum_{i=1}^{c}\frac{N_i}{N}(m_i-\overline{x})(m_i-\overline{x})^T \tag{1}$$

$$S_w=\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N_i}(x_j^i-m_i)(x_j^i-m_i)^T \tag{2}$$

where m_i denotes the mean of the ith class, $\overline{x}$ denotes the mean of all the patterns in the pattern set.

According to the scatter matrices, the Fisher criterion function can be defined as follows:

$$J_{FC}(\omega)=\frac{\omega^T S_b\omega}{\omega^T S_w\omega} \tag{3}$$

where ω is an arbitrary vector in d-dimensional space. The Fisher optimal dicriminant vector is ω^* corresponding to maximum of $J_{FC}(\omega)$, which is the eigenvector corresponding to maximum eigenvalue of the the following eigen-system equation:

$$S_b\omega^*=\lambda S_w\omega^* \tag{4}$$

Where λ is diagonal and consists of the corresponding eigenvalues.

3 Fuzzy Fisher Criterion and the Proposed Algorithm

First, let us fuzzify the concept of the above Fisher linear discriminant(FLD).

Suppose that the membership function $u_{ij}\in[0,1]$ with $\sum_{i=1}^{c}u_{ij}=1$ for all j and the fuzzy index $m>1$ is a given real value, where u_{ij} denotes the degree of the jth d-dimensional pattern belonging to the ith class, we can define the following fuzzy within-class scatter matrix S_{fw} [5]:

$$S_{fw}=\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}^m(x_j-m_i)(x_j-m_i)^T \tag{5}$$

and the following fuzzy between-class scatter matrix S_{fb}:

$$S_{fb}=\sum_{i=1}^{c}\sum_{j=1}^{N}u_{ij}^m(m_i-\overline{x})(m_i-\overline{x})^T \tag{6}$$

Thus, we can derive a novel fuzzy Fisher criterion called fuzzy FLD as follows:

$$J_{FFC}=\frac{\omega^T S_{fb}\omega}{\omega^T S_{fw}\omega} \tag{7}$$

In terms of the fuzzy FLD as above, we will derive a novel semi-fuzzy clustering algorithm based on fuzzy Fisher criterion. Maximizing J_{FFC} directly in Eq.(7) is not a trivial mathematical derivation task due to the existence of its denominator. However, we can reasonably relax this problem by applying the following Lagrange multipliers λ and $\beta_j(j=1,2,\cdots n)$ together with the constraint $\sum_{i=1}^{c}u_{ij}=1$ to Eq.(7):

$$F=\omega^T S_{fb}\omega-\lambda\omega^T S_{fw}\omega+\sum_{j=1}^{N}\beta_j(\sum_{i=1}^{c}u_{ij}-1) \tag{8}$$

Setting $\frac{\partial F}{\partial\omega},\frac{\partial F}{\partial m_i},\frac{\partial F}{\partial u_{ij}}$ to be zero, we respectively have

$$S_{fb}\omega=\lambda S_{fw}\omega \tag{9}$$

where λ may be taken as the largest eigenvalue.

$$m_i=\frac{\sum_{j=1}^{N}u_{ij}^m(x_j-\frac{1}{\lambda}\overline{x})}{\sum_{j=1}^{N}u_{ij}^m(1-\frac{1}{\lambda})} \tag{10}$$

$$u_{ij}=F_1/F_2 \tag{11}$$

where

$$F_1 = (\omega^T (x_j - m_i)(x_j - m_i)^T \omega - \frac{1}{\lambda} \omega^T (m_i - \bar{x})(m_i - \bar{x})^T \omega)^{-\frac{1}{m-1}}$$

$$F_2 = \sum_{k=1}^{c} (\omega^T (x_j - m_k)(x_j - m_k)^T \omega - \frac{1}{\lambda} \omega^T (m_k - \bar{x})(m_k - \bar{x})^T \omega)^{-\frac{1}{m-1}}$$

When Eq.(11) is used, as stated in the above, u_{ij} should satisfy $u_{ij} \in [0,1]$, hence, in order to satisfy this constraint, we let

$u_{ij} = 1$ and $u_{i'j} = 0$ for all $i' \neq i$, if

$$\omega^T (x_j - m_i)(x_j - m_i)^T \omega \leq \frac{1}{\lambda} \omega^T (m_i - \bar{x})(m_i - \bar{x})^T \omega \tag{12}$$

That is to say, if Eq.(12) holds, we take a hard partition for the pattern. The rational can be intuitively explained from a geometric viewpoint as shown in Figure 1. In Figure 1, all the patterns with one class □ and the other class ○, the two clusters center with * and the total mean point (i.e. the average point of all samples.) with ☆ are projected along dotted lines onto the optimal discriminating vector. Obviously, if the Euclidean distance between the projection of a pattern and the projection of certain cluster is equal to or less than multiplied by the Euclidean distance between the projection of this cluster and ☆, then we should take a hard partition for this pattern.

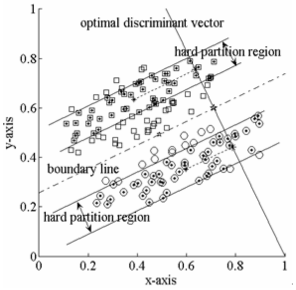


Figure 1 FFC hard partition regions

From the above analysis, we can obtain a novel semi-fuzzy algorithm based on fuzzy fisher criterion (FFC).

Algorithm FFC:

Step1. Set the given threshold ε *, initialize* $U = [\mu_{ij}]_{c \times N}$ *and* $m = (m_1, m_2, ..., m_c)$ *using K-means;*

Step2. Compute S_{fw}, S_{fb} *using Eq.(5), Eq.(6) respectively;*

Step3. Compute the largest eigenvalue λ *and the corresponding* ω *using Eq.(9);*

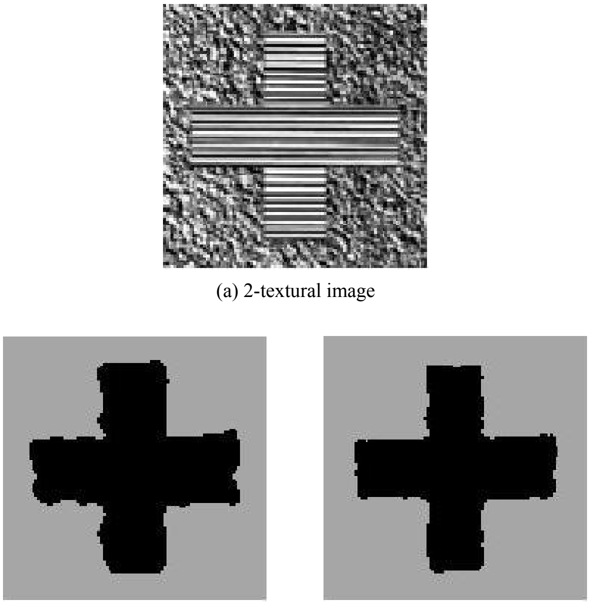
Step4. Update m_i *and* μ_{ij} *using Eq.(10), Eq.(11) and Eq.(12) respectively;*

Step5. Compute J_{FFC} *using Eq.(7);*

Step6. If $J_{FFC} < \varepsilon$ *or the number of iteration* ≥ *the given value, output the clustering result and then terminate, otherwise back to Step 2.*

4 Experimental results

This experiment is arranged to examine FFC's clustering performance and robust capability for texture segmentation for artificial images without noise and especially noisy images. The texture images which contain the textures taken from Brodatz texture base [8] are used here. The 2-textural image in Figure 2 (a) consists of textures (D4, D49), the 5-textural image in Figure 4 (a) consists of textures (D21, D22, D49, D53, D55), respectively. Noisy images Figure 3 (a) and Figure 5 (a) were obtained by adding up Gaussian noise with mean 0 and variance 0.05 to Figure 2 (a) and Figure 4 (a).

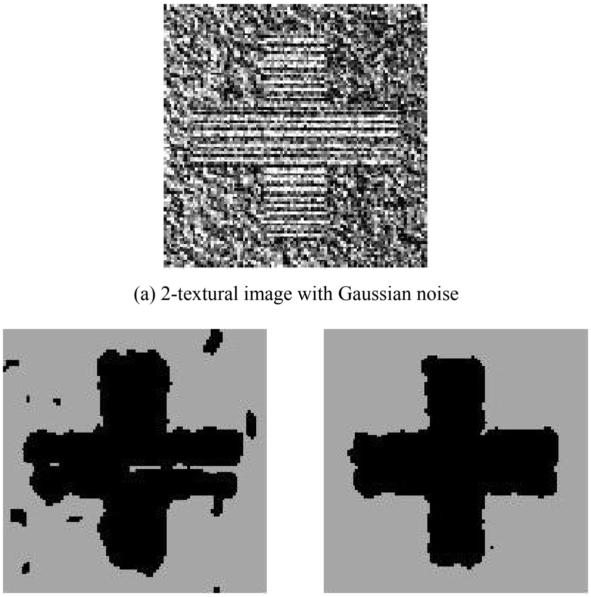


(a) 2-textural image

(b) FCM

(c) FFC

Figure 2 Segmentation results of two algorithms for 2-textural image

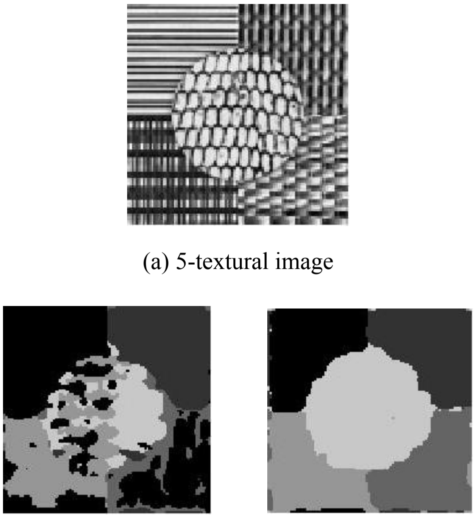


(a) 2-textural image with Gaussian noise

(b) FCM

(c) FFC

Figure 3 Segmentation results of three algorithms for Gaussian noisy 2-textural image



(a) 5-textural image

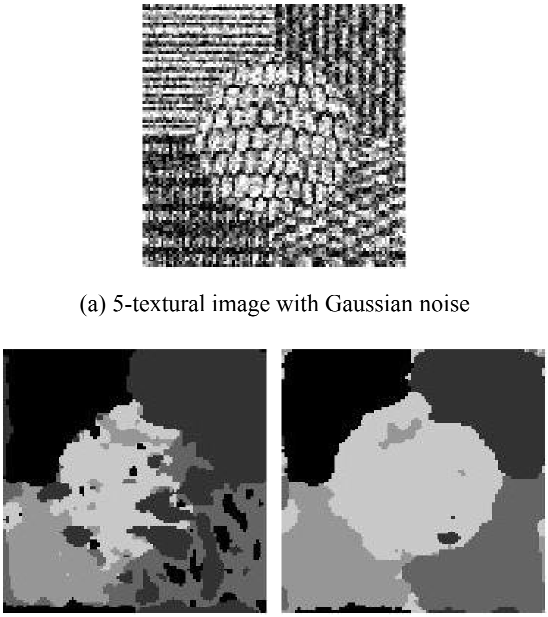
(b) FCM

(c) FFC

Figure 4 Segmentation results of two algorithms for 5-textural image

To perform image texture segmentation, texture features can be collected on a pixel by pixel basis. Pattern recognition methods are then used to group these features into appropriate classes to achieve the segmentation. In this section, examples using Gabor filters [9, 10] for the purpose of performing texture segmentation are provided.

After the texture features are extracted using the same Gabor filters, we use FCM and FFC to segment these texture images. Figs. 2-5 (b-c) illustrate the corresponding unsupervised segmentation results using those two algorithms.



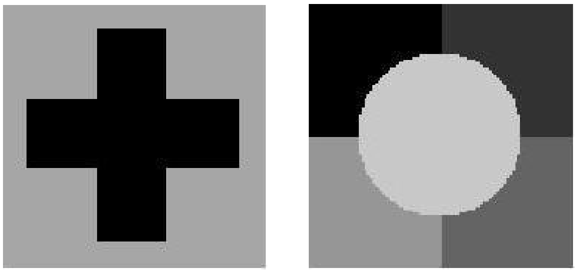
(a) 5-textural image with Gaussian noise

(b) FCM

(c) FFC

Figure 5 Segmentation results of two algorithms for Gaussian noisy 5-textural image

In order to quantitatively evaluate the segmentation and robustness capabilities of these two algorithms, Figure 6 shows perfect segmentation results for the above images with 2 textures and 5 textures, respectively. After counting the number of wrongly assigned pixels by comparing all assigned pixels with the corresponding perfect segmentation result, we can obtain the corresponding segmentation accuracy. Table 1 lists segmentation accuracies of two algorithms for all the above texture images. From these figures and Table 1, we can see that FFC has the best segmentation accuracies in most of cases, and is comparatively competitive to FCM.



(a)

(b)

Figure 6 Perfect segmentation results for the above 2(or 5)-textural images

Table 1 Segmentation accuracies of two algorithms for (noisy) texture images

Figure	Noise Type	FCM Accuracy (%)	FFC Accuracy (%)
Fgure 2	——	96.4	93.7
Figure 3	Gaussian	93.2	94.3
Figure 4	——	63.2	94.0
Figure 5	Gaussian	69.2	90.8

References

[1] Bezdek J C, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981

[2] Rouseeuw P J, Kaufman L, Trauwaert E, “Fuzzy clustering using scatter matrices”, Computational Statistics & Data Analysis, 23(7), 1996, pp. 135-151

[3] Krishnapuram R, Kim J, “Clustering algorithms based on volume criteria”, IEEE Transactions on Fuzzy Systems, 8(2), 2000, pp. 228-236

[4] Gath I,Geva A B, “Unsupervised optimal fuzzy clustering”, IEEE Trans on Pattern Anal. & Machine Intell., 11(7), 1989, pp. 773-781

[5] Kuo-Lung Wu, Jian Yu, Miin-Shen Yang, “A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests”, Pattern Recognition Letters, 26(4), 2005, pp. 639-652

[6] Zhonghang Yin, Yuangang Tang, Fuchun Sun, et al, “Fuzzy Clustering with Novel Separable Criterion”, Chinese J. Tsinghua University, 11(2), 2006, pp. 50-53

[7] ZaoQi Bian, XueGong Zhang, Pattern Recognition, Beijing: TsingHua University Press, 2001, pp. 87-90

[8] Trygve Randen, Brodatz Textures. http://www.ux.uis.no/~tranden/brodatz.html, 2006

[9] D.A.Clausi, “K-means iterative Fisher (KIF) unsupervised clustering algorithm applied to image texture segmentation”, Pattern Recognition, 35(9), 2002, pp. 1959-1972

[10] Kyrki, V., Kamarainen, J.-K., Kalviainen, “H. Simple Gabor Feature Space for Invariant Object Recognition”, Pattern Recognition Letters, 25(3), 2004, pp. 311-318

A Novel Image Compression Method Based on Fuzzy Theory and Neural Networks

Xingliang Zhu[1] Xin Tan[2] Shilian Xu[3]

1 School of Management, Chongqing Jiaotong University Chongqing, 400074, China
Email: manaduona@yahoo.com.cn

2 School of Communication, Chongqing University of Posts and Telecommunications Chongqing, 400065, China
Email: weisite2008@yahoo.com.cn

3 Department of Logistics Information Engineering, Logistical Engineering University Chongqing, 400016, China
Email: swaol2008@yahoo.com.cn

Abstract

A image compression algorithm using fuzzy competitive learning neural network is proposed in this paper. The proposed scheme is based on vector quantization. Then, competitive learning neural network and fuzzy control system are included in the scheme to train the codebook and encode the source image by using the membership function and control rules. Finally, to demonstrate the effectiveness of this scheme, experiment results are presented. According to the experimental results, comparing with conventional vector quantization(LBG), the proposed scheme could greatly improve the quality of codebook.

Keywords: Image Compression; Fuzzy; Competitive Learning; Neural Network; Vector Quantization; Codebook

1 Important Information

The scalar version of K-mean algorithm was published by Lioyd in 1957. And Forgy proposed turning it into vector quantization in 1965. It is also called LBG algorithm [1]. Since it was presented by Linde, Buzo and Gray in 1980 [1].

LBG vector quantization [1][2][3][4][5] is a well established scheme of image compression. Under this method, the input image is divided into several small rectangular blocks known as source vectors.

A codebook contains a number of available codewords of the same size as the source vectors, and the quantizer replaces each source vector by the codeword that is most similar to it. The objective of the codebook design is to find a codebook that contains the most representative codewords. This codebook is then used to encode the image for transmission and decode the image for reproduction.

In order to avoid excessive computation time by generating a new codebook for each image, and also to increase the compression ratio by removing the need to transmit the codebook with each image, a universal codebook is created by applying the LBG algorithm to predefined set of training images. This codebook can then be used to encode and decode images outside of the training set. The reproduction image, encodes using such a codebook, exhibits considerable distortion. An improvement in codebook quality which will result in less distortion reproduced images is thus desirable.

Neural networks [6][7][8][9] provide various schemes to optimize the design of vector quantization. Fuzzy classification [10][11][12] is also a well-known technique of designing codebook. In this paper, We propose a scheme called Fuzzy Competitive Learning Neural Network(FCLNN) to optimize the codebook design by using competitive learning neural network and

* This paper is supported by the Project of the Chongqing Municipal Education Commission (No. KJ070409) and the Natural Science Foundation of Chongqing University of Posts and Telecommunications (No. A2007-20)

fuzzy control system [11][12][13], which could make codebook quality superior and more representative.

After a brief view of competitive learning algorithms in section 2, the proposed scheme is described in detail in section 3. The experimental results are given in section 4. Finally, section 5 draws conclusions.

2 Competitive Learning Algorithms

The basic competitive learning algorithms steps are as follow:

Step 1. Initialization of neuron weights: The *M* neuron coupling weights are initialized as the starting codebook:

$$W_j(0), j = 1, 2, 3, \ldots\ldots, M \qquad (1)$$

Step 2. Competition: Computing the Euclidean distance:

$$D_{ij} = d(X_i, W_j(t)) \text{ , } j = 1,2,\ldots\ldots, M \qquad (2)$$

Where, X_i is the input vector, the winning neuron *k* is selected with:

$$D_{ik} = \min_{j} D_{ij} \qquad (3)$$

Step 3. Learning and updating:

$$W_k(t+1) = W_k(t) + \alpha(t) I_k(t) \{X_i - W_k(t)\} \qquad (4)$$

Where $\alpha(t)$ is the learning rate at iteration *t*; $I_k(t)$ is the scaling function specifying the sign and magnitude of the difference vector which is updated for the winning neuron k. Both updating factors $\alpha(t)$ and $I_k(t)$ can be designed or replaced by fuzzy membership functions to become variations of fuzzy competitive learning neural networks.

Step 4. Termination: Repeat steps 2-3 until the terminating criterion is fit.

3 Codebook Design of FCLNN

3.1 Initialization

Given the input training sequence of vectors $X = \left\{X_1, X_2 \cdots\cdots X_n\right\}$, and X_j is current input; initial codebook (neurone coupling weight) $W = \{W_1, W_2 \cdots\cdots W_m\}$; scaling function $I_k(t) = 1$ [3],[6]and learning rate $\alpha(t) = 1/\left\|S_i^j\right\|$ [3],[6]. $\left\|S_i^j\right\|$ is the total number of vectors which belong to the ith cluster (S_j) up to the input of vector X_j, where the cluster (S_j) is associated with the ith neuron. The superscript *j* indicates that the cluster relates to X_j.

Initial inputs of fuzzy control system are $E_1(t) = D(t)$ and $E_2(t) = D(t) - D(t-1)$. $D(t)$ represents average distortion between original image and reconstruction image here. The symbol " *t* " represents current iteration condition.

3.2 Fuzzy control rules

Define fuzzy sets on the universes of $E_1(t)$, $E_2(t)$ and ΔU as:

PE_1: positive process error1; NE_1 : negative process error1,

PE_2: positive process error2; NE_2 : negative process error2,

$P\Delta U$: positive change in control output.

$N\Delta U$: negative change in control output.

$Z\Delta U$: zero change in control output.

Rule1: If E_1 is PE_1 and E_2 is PE_2, then ΔU is $P\Delta U$.

Rule2: If E_1 is NE_1 and E_2 is PE_2 , then ΔU is $Z\Delta U$.

Rule3: If E_1 is PE_1 and E_2 is NE_2, then ΔU is $Z\Delta U$.

Rule4: If E_1 is NE_1 and E_2 is NE_2 , then ΔU is $N\Delta U$.

The membership functions of these fuzzy sets are shown in Figure 1 and Figure 2.

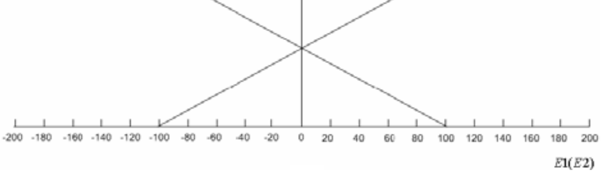


Figure 1 Membership function for input of fuzzy control system

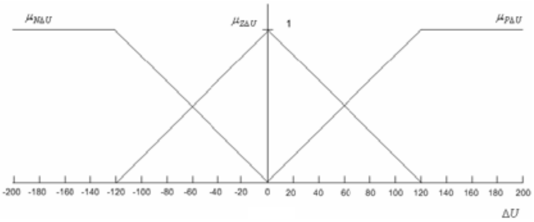


Figure 2 Membership function for output of fuzzy control system

3.3 Fuzzy Inference

According to[11],[12], we can get the center of gravity defuzzifier by

$$\Delta U(t)=\frac{\sum_{j=1}^{m}\Delta\overline{U}_j\left[\mu_{B'}(\Delta\overline{U}_j)\right]}{\sum_{j=1}^{m}\left[\mu_{B'}(\Delta\overline{U}_j)\right]} \tag{5}$$

To compare (4), we can obtain updating equation by replace $I_k(t)$ with $\Delta U(t)$ which is shown in (5). And the codebook design of fuzzy competitive learning neural network is formed. The algorithm is described as follows:

Step 1. Initialization of neuron weights: The M neuron coupling weights are initialized as the starting codebook:

$$W_j(0), j=1,2,3,......,M \tag{6}$$

Step 2. Competition: Computing the Euclidean distance:

$$D_{ij}=d\left(X_i,W_j\left(t\right)\right),\quad j=1,2,......M \tag{7}$$

Where, X_i is the input vector, the winning neuron k is selected with:

$$D_{ik}=\min_{j} D_{ij} \tag{8}$$

Step 3. Learning and updating:

$$W_i^N(j)=W_i^N(j-1)+\Delta U(t)\cdot\frac{1}{\left\|S_i^j\right\|}\cdot\left(X_j-W_i^N\left(j-1\right)\right) \tag{9}$$

Where the superscript N is used to indicate that $W_i^N(j-1)$represents the ith neuron coupling weight before X_j is partitioned.

Step 4. Termination: Repeat steps 2-3 until the terminating criterion is fit.

At last, the block diagram of fuzzy competitive learning neural network is shown in Figure 3.

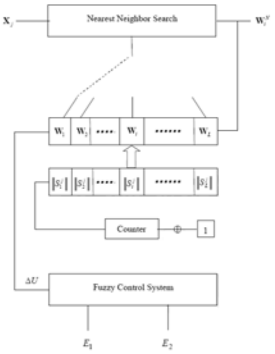


Figure3 Block diagram of Fuzzy Competitive Learning Neural Network

The figure showed the process of FCLNN. The fuzzy control system is included to improve PSNR. The input of fuzzy control system is E_1 and E_2, and output is ΔU. The block diagram means that image blocks are used to make the nearest neighbor search [5] from the neurons firstly. Then, the winning neuron use fuzzy control system to update and training the codebook until the last image block.

4 Experimental Results

In the experiment, for comparison purpose, we apply proposed algorithm and LBG algorithm to process the input image based on Visual C++ 6.0 programming language and Matlab 6.5 package software.

In the experiment, there are nine still 2-D images(512×512,256×256,128×128 pixels)with different gray-level contained in the training set to design a large static codebook of the compression system. Meanwhile, we train a series of codebook by the proposed algorithm and LBG algorithm. Their sizes are “128×4”, “128×16”, “256×4”, “256×6”, “512×4”, “512×16”, “1024×4”, “1024×16”, in which first element represents items of codebook—row size of codebook, second element represents dimension of codebook— column size of codebook.

For each size of codebook, we get the PSNR (Peak Signal-to-Noise Ratio)[1],[2] of each image. Then, the average PSNR of the proposed algorithm and LBG algorithm are shown in Table 1 and Table 2 respectively. Finally, the comparison of two algorithms is shown in Figure 4.

Table1 Average PSNR of all LBG(Linde, Buzo and Gray) codebooks

Codebook size	Average PSNR(dB)
1024×4	32.9321
512×4	31.4792
256×4	30.0985
128×4	28.4301
1024×16	25.3402
512×16	24.5962
256×16	23.5600
128×16	22.3410

Table 2　Average PSNR of all Fuzzy Competitive Learning Neural Network codebooks

Codebook size	Average PSNR(dB)
1024×4	36.1994
512×4	34.6301
256×4	32.2497
128×4	30.4815
1024×16	27.8702
512×16	26.3989
256×16	25.9534
128×16	24.7626

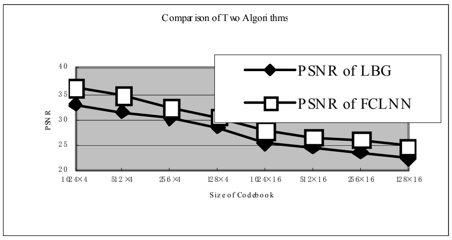


Figure 4　Comparison of two algorithms. It obviously shows that the FCLNN method has better PSNR than LBG algorithm.

5　Conclusions

In this paper, an image compression scheme for still 2-D image compression based on competitive learning neural network and fuzzy control system is proposed, which can improve the codebook design to reach higher image compression quality. According to the experiment results, with various codebooks, the proposed scheme has higher PSNR values that are around 10% better than LBG's. All in all, all of the experimental results demonstrated that the proposed scheme is feasible and effective.

References

[1] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design", IEEE Transactions on Communications, Vol. 28, No. 1, 1988, pp.84-95

[2] N. M. Nasrabadi, and R. A. King, "Image Coding Using Vector Quantization: a Review", IEEE Transactions on Communications, Vol. 36, No. 8, 1989, pp.57-971

[3] C. M. Huang, and R. W. Harris, "A Comparison of Several Vector Quantization Codebook Generation Approaches", IEEE Transactions on Image Processing, Vol. 2, No. 1, 1993, pp.108-112

[4] I. Jee, and R. A. Haddad, "Optimum design of vector-quantized subband codecs", IEEE Transactions on Signal Processing, Vol. 46, No. 8, 2004, pp.2239-2243

[5] M. R. Soleymani, and S. D. Morgera, "An Efficient Nearest Neighbor Search Method", IEEE Transactions on Communications, Vol. 35, No. 6, 1987, pp.677-679

[6] R. D. Dony, and S. Haykin, "Neural Network Approaches to Image Compression", Proceedings of the IEEE, Vol. 83, No. 2, 2003, pp.288-302

[7] J. H. Wang, C. Y. Peng, and J. D. Rau, "Harmonic Neural Networks for On-line Learning Vector Quantization", IEEE Proceedings, Image Signal Process, Vol. 147, No. 52, 2005, pp.485-492

[8] R. Lancini, and S. Tubaro, "Adaptive Vector Quantization for Picture Coding Using Neural Networks", IEEE Transactions on Communications, Vol. 43, No. 234, 1995, pp. 534-544

[9] A. Namphol, S. Chin, and M. Arozullah, "Image Compression with a Hierarchical Neural Network", IEEE Transactions on Aerospace and Electronic Systems, Vol. 32, No. 1, 1996, pp. 326-337

[10] C. Amerijckx et al, "Image Compression by Self-organized Kohonen Maps", IEEE Trans. Neural Networks, vol. 9, 1998, pp.503–507

[11] 11. N. B. Karayiannis, "A Methodology for Constructing Fuzzy Algorithms for Learning Vector Quantization", IEEE Transactions on Neural Networks, Vol. 8, No. 3, 1997, pp.505-518

[12] N. B. Karayiannis, and P. I. Pai, "Fuzzy Vector Quantization Algorithms and Their Application in Image Compression", IEEE Transactions on Image Processing, Vol. 4, No. 9, 1995, pp.1193-1201

An Image Compression Algorithm Based on Neural Networks

Zejun Hu[1] Xin Tan[2] Xingliang Zhu[3]

1 Department of Computer and Information Engineering, Guangdong Vocational College of Mechanical and
Electrical Technology, Guangzhou, 510515, China
Email: meixiagent@yahoo.com.cn

2 School of Communication, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
Email: weisite2008@yahoo.com.cn

3 School of Management, Chongqing Jiaotong University, Chongqing, 400074, China
Email: manaduona@yahoo.com.cnAbstract

Abstract

A novel image compression algorithm using competitive learning neural network is proposed to improve the original LBG algorithm, by using which higher image compression quality could be reached. The proposed scheme is based on vector quantization and modifies the learning rate and scaling function of updating equation, which is used to train the codebook. Finally, to demonstrate the feasibility of this scheme, experimental results are presented, which show the proposed scheme can improve the quality of codebook greatly.

Keywords: Compression, Image, Neural Network, Training, Learning

1 Introduction

LBG(Linde, Buzo and Gray) vector quantization [1,2,3] is a well established method of image compression. It is a kind of application of vector quantization [1,2,3,4,5]. Under this scheme, the input image is divided some square blocks called as input vectors. A codebook contains several codewords of the same size as the input vectors and the quantizer take the place of each input vector by the most similar codeword. The objective of codebook design is to train a codebook that contains the most representative codewords so that codebook can adapt itself to all kinds of image. Then, this codebook is used to encode the image for transmission and decode the image for reproduction. The LBG algorithm minimizes iteratively the total image distortion to produce the useful codebook.

In order to increase the compression ratio by removing the need to transmit the codebook with each image, and also to avoid excessive computation time by generating a new codebook for each image, a universal codebook is created by applying the LBG algorithm to pre-defined set of training images. This codebook could be used to encode and decode images outside of the training set. And the reproduced image which is encoded by using the trained codebook represents a great deal of distortion. An improvement in codebook quality which be going to leads less distortion is thus desirable.

However, one of shortcomings of original LBG algorithm is a sort of off-line training technique [2,3,4]. Neural networks [6,7,8,9] provide various schemes to optimize the design of vector quantization. In this paper, the competitive learning neural network [10,11,12,13] is adopted to improve the codebook quality, which designs a kind of on-line training and learning scheme to achieve the objective of improvement. This is because as each vector is processed, the corresponding coupling weight is updated accordingly.

After a brief view of competitive learning neural

* This paper is supported by the Project of the Chongqing Municipal Education Commission (No. KJ070409) and the Natural Science Foundation of Chongqing University of Posts and Telecommunications (No. A2007-20).

network in section 2, the proposed scheme is described in detail in section 3. The experimental results are given in section 4. Finally, section 5 draws conclusions.

2 Competitive Learning Neural Network

In order to avoid excessive computation time by generating a new codebook for each image, and also to increase the compression ratio by removing the need to transmit the codebook with each image, a universal codebook is created by applying the LBG algorithm to predefined set of training images. This codebook can then be used to encode and decode images outside of the training set. The reproduction image, encodes using such a codebook, exhibits considerable distortion. An improvement in codebook quality which will result in less distortion reproduced images is thus desirable. The updating of the codebook using the conventional LBG algorithm is viewed as off line. All the vectors in the input image are processed against the existing codebook before the codebook is next updated.

As a result, a number of outstanding researches have been done on neural network vector quantization in recent years. Many successful neural networks [6,7,8,9,10] are proposed to provide alternative techniques in optimizing the vector quantization codebook design, one of the representative neural networks is competitive learning [10,11,12,13]. Competitive learning networks employ competition among a number of output units or neurons in the network to represent the input vectors. In standard or hard competitive learning, known as winner-take-all (WTA) networks, an input vector is simply approximated by the closest data cluster center. These centers are represented by the weight vectors of the competing units after learning is completed. This provides a vector quantization (VQ) of the inputs, which serves as a compressed and localized approximation of the input data vectors. Alternatively, in soft versions of competitive learning, multiple units are excited simultaneously, creating a distributed representation or encoding of the data. Competitive networks may also be only the first layer of a multilayer network, in which supervised learning is employed in the layers above. In this case, the use of unsupervised learning in the first layer may substantially reduce learning time for the network, as compared to a fully-supervised learning algorithm such as backpropagation [12].

3 Recommended Point Sizes

3.1 Competitive learning algorithm

The basic competitive learning algorithm steps are as follow:

Step1. Initialization of neuron weights: The M neuron coupling weights are initialized as the starting codebook:

$$W_j(0), j = 1,2,3,\ldots\ldots,M \tag{1}$$

Step 2. Competition: Computing the Euclidean distance:

$$\mathrm{D}_{ij} = d(X_i,W_j(t))\ , j = 1,2,\ldots\ldots,M \tag{2}$$

Where, Xi is the input vector, the winning neuron k is selected with:

$$D_{ik} = \min_j D_{ij} \tag{3}$$

Step 3. Learning and updating:

$$W_k(t+1) = W_k(t) + \alpha(t) I_k(t)\{X_i - W_k(t)\} \tag{4}$$

Where $\alpha(t)$ is the learning rate at iteration t ; $I_k(t)$ is the scaling function specifying the sign and magnitude of the difference vector which is updated for the winning neuron k.

Step 4. Termination: Repeat steps 2-3 until the terminating criterion is fit.

3.2 Proposed scheme

Codebook design based on neural network can be separated into two different ways. One is that neural network production of the codebook can be viewed as an on- line learning in comparison with LBG. This is because that as each input vector Xi is being processed, neural network tends to learning from the vector the information about the input sequence and update its codebook on line rather than classify all the vectors and produce the next partition for further classification in the next cycle. The other is we can view it as traditional off- line learning. That is so called LBG vector quantization.

On-line learning makes it possible to implement the LBG algorithm by modifying the competitive learning algorithm. Above all, two events need to be paid attention here. Firstly, it is very definite to use the termination criterion $(D_{m-1} - D_m)/D_m \leq \varepsilon$ for the competitive learning neural network. Secondly, we ought to make a conversion of off-line updating codebook by LBG to an on-line competitive learning rule. We can start from LBG off- line training of its codebook that is described by [14] to achieve the conversion. LBG algorithm could be converted to competitive learning neural network with the following equation:

$$W_i^{LBG}(t+1) = \frac{1}{\left\|S_i^j\right\|}\sum_{X \in S_i^j} X_k = \frac{\sum\limits_{X_k \in S_i^j} X_k + X_j}{\left\|S_i^j\right\|} \tag{5}$$

Where we assume that n = j and Xj is current input for an input training sequence of vectors $X = \{X_1, X_2, ..., X_n\}$. Besides, $W_i^{LBG}(t+1)$ is the next weight for the ith winning neuron. $\left\|S_i^j\right\|$ denotes the total number of vectors which belong to the ith cluster (S_j) up to the input of vector X_j, where the cluster (S_j) is associated with the ith neuron. The superscript j indicates that the cluster relates to X_j.

Now we select n = j-1. It means the input training sequence is X = {X_1, X_2,……, X_j-1}. The ith winning neuron would have been updated by following equation:

$$W_i^{LBG}(t+1) = \frac{1}{\left\|S_i^{j-1}\right\|}\sum_{X \in S_i^{j-1} \& k \neq j} X_k \tag{6}$$

Furthermore, considering the on-line training with the competitive learning neural network, Eq. (5) actually represents the neuron weight updated before X_j is arrived.

This is exactly the same as the old weight for the ith neuron when the vector Xj is being processed inside the competitive learning neural network. As a result, we have:

$$W_i^{LBG}(t+1) = W_i^N(j-1) \tag{7}$$

And reword Eq. (5) and Eq. (6) as:

$$W_i^N(j) = \frac{1}{\left\|S_i^j\right\|}\sum_{X \in S_i^j} X_k = \frac{\sum\limits_{X_k \in S_i^j} X_k + X_j}{\left\|S_i^j\right\|} \tag{8}$$

$$W_i^N(j-1) = \frac{1}{\left\|S_i^{j-1}\right\|}\sum_{X \in S_i^{j-1} \& k \neq j} X_k \tag{9}$$

Where the superscript N is used to indicate that $W_i^N(j-1)$ represents the ith neuron coupling weight before X_j is partitioned. Hence, substituting Eq. (8) and Eq. (9) into Eq. (5) and, after rearranging the terms, we obtains:

$$W_i^{LBG}(t+1) = W_i^N(j) = \frac{\left\|S_i^{j-1}\right\| + 1}{\left\|S_i^j\right\|} W_i^N(j-1) + \frac{1}{\left\|S_i^j\right\|}\left(X_j - W_i^N(j-1)\right) \tag{10}$$

Due to $\left\|S_i^j\right\|$ is the total number of vectors which belong to S_i up to the input of vector(X_j), the equation is obtained as follow:

$$\frac{\left\|S_i^{j-1}\right\| + 1}{\left\|S_i^j\right\|} = 1 \tag{11}$$

Consequently, Eq. (10) can be reworded as Eq. (12):

$$W_i^N(j) = W_i^N(j-1) + \frac{1}{\left\|S_i^j\right\|}\left(X_j - W_i^N(j-1)\right) \tag{12}$$

Comparing Eq. (12) with Eq. (4) could find that learning rate $\alpha(t) = 1/\left\|S_i^j\right\|$ and the scaling function $I_k(t) = 1$.

At last, we can realize the whole system by referencing Figure 1 which is the block diagram of competitive learning neural network.

The figure showed the process of competitive learning neural network. It means that image blocks are used to make the nearest neighbor search [5] from the neurons firstly. Then, find the winning neuron and add one to the counter which belongs to the winner. After all image blocks are used, the training processes are finished.

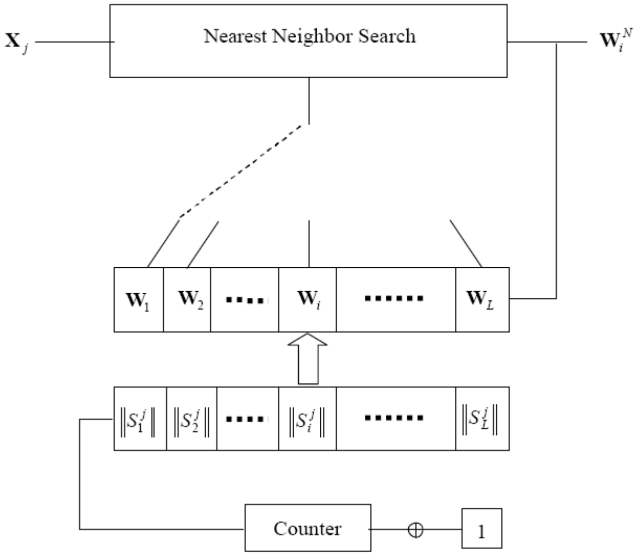


Figure 1 Block diagram of competitive learning neural network

4 Experimental Results

In the experiment, for comparison purpose, we apply proposed algorithm and LBG algorithm to process the input image based on Visual C++ 6.0 programming language and Matlab 6.5 package software.

In the experiment, there are nine still 2-D images(512×512,256×256,128×128 pixels)with different gray-level contained in the training set to design a large static codebook of the compression system. Meanwhile, we train a series of codebook by the proposed algorithm and LBG algorithm. Their sizes are “128×4”, “128×16”, “256×4”, “256×6”, “512×4”, “512×16”, “1024×4”, “1024×16”, in which first element represents items of codebook—row size of codebook, second element represents dimension of codebook—column size of codebook.

For each size of codebook, we get the PSNR(Peak Signal-to-Noise Ratio) [1,2,5] of each image. Then, the average PSNR of the proposed algorithm and LBG algorithm are shown in Table 1 and Table 2 respectively. Finally, the comparison of two algorithms is shown in Figure 2.

Table 1 Average PSNR of all LBG (Linde, Buzo and Gray) codebooks

Codebook size	Average PSNR(dB)
1024×4	32.9321
512×4	31.4792
256×4	30.0985
128×4	28.4301
1024×16	25.3402
512×16	24.5962
256×16	23.5600
128×16	22.3410

Table 2 Average PSNR of all Competitive Learning Neural Network codebooks

Codebook size	Average PSNR(dB)
1024×4	35.7927
512×4	34.2137
256×4	31.6115
128×4	30.8505
1024×16	27.9089
512×16	26.6257
256×16	25.4715
128×16	24.7422

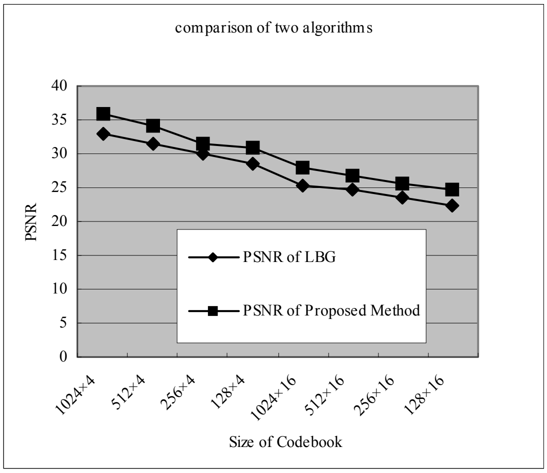


Figure 2 Comparison of two algorithms. It obviously shows that the proposed method has better PSNR than the LBG algorithm

5 Conclusions

The traditional LBG image compression algorithm has a shortcoming: a sort of off-line training technique [2,3,4]. In this paper, a image compression scheme for still 2-D image compression based on competitive learning neural network is proposed to improve the codebook design, which designs a kind of on-line training and learning scheme to reach higher image compression quality.

According to the experiment results, with various codebook size, the scheme based on competitive learning neural network has higher PSNR values, which are around 8% better than LBG's. In a word, all of the experimental results demonstrated and verified the proposed scheme practicable and satisfactory.

References

[1] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design", IEEE Transactions on Communications, Vol. 28, No. 1, 1988, pp.84-95

[2] N. M. Nasrabadi, and R. A. King, "Image coding using vector quantization: a review", IEEE Transactions on Communications, Vol. 36, No. 8, 1988, pp.957-971

[3] C. M. Huang, and R. W. Harris, "A comparison of several vector quantization codebook generation approaches", IEEE Transactions on Image Processing, Vol. 2, No. 1, 1993, pp.108-112

[4] I. Jee, and R. A. Haddad, "Optimum design of vector-quantized subband codecs", IEEE Transactions on Signal Processing, Vol. 46, No. 8, 2003, pp.2239-2243

[5] M. R. Soleymani, and S. D. Morgera, "An efficient nearest neighbor search method", IEEE Transactions on Communications, Vol. 35, No. 6, 1987, pp.677-679

[6] R. D. Dony, and S. Haykin, "Neural network approaches to image compression", Proceedings of the IEEE, Vol. 83, No. 2, 1995, pp.288-302

[7] J. H. Wang, C. Y. Peng, and J. D. Rau, "Harmonic neural networks for on-line learning vector quantization", IEEE Proceedings, Image Signal Process, Vol. 147, No. 5, 2000, pp.485-492

[8] R. Lancini, and S. Tubaro, "Adaptive vector quantization for picture coding using neural networks", IEEE Transactions on Communications, Vol. 43, No. 234, 1995, pp.534-544

[9] A. Namphol, S. Chin, and M. Arozullah, "Image compression with a hierarchical neural network", IEEE Transactions on Aerospace and Electronic Systems, Vol. 32, No. 1, 1996, pp.326-337

[10] S. C. Ahalt, and A. K. Krishnamurthy, "Competitive learning algorithms for vector quantization", Neural Networks, Vol. 3, No. 3, 1990, pp.277-290

[11] D. C. Park, "Centroid neural network for unsupervised competitive learning", IEEE Transactions on Neural Networks, Vol. 11, No. 2, 2001, pp.1134-1146

[12] H. C. Card, G. K. Rosendahl, D. K. McNeill, and R. D. McLeod, "Competitive learning algorithms and neurocomputer architecture", IEEE Transactions on Computers, Vol. 47, No. 8, 2005, pp.847-858

[13] L. Wang, "On competitive learning", IEEE Transactions on Neural Networks, Vol. 8, No. 5, 1997, pp.1214-1217

[14] G. Basil, and J. Jiang, "An improvement on competitive learning neural network by LBG vector quantization", IEEE International Conference on Multimedia Computing and Systems, Vol. 1, 1999, pp.244-249.

Efficient Path-Reconstruction Operator

Jinhui Cai   Hui Cai*   Guangxin Zhang   Zekui Zhou

1 College of Metrological Technology and Engineering. China Jiliang University, Hangzhou 310018, China

2 Department of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
Email: caihui1982@hotmail.com

Abstract

In order to efficiently filter noise by length and orientation characters of objects, path-reconstruction operator was brought forward based on mathematical morphological reconstruction. And a novel First-In-First-Out queue-based efficient algorithm for path-reconstruction operator was proposed. In this algorithm, Pixels in marker image were processed in decreasing grey leve1 order to label connected components, and some connected components in mask image were selectively reconstructed according to the δ-path length of corresponding regions in marker image. Results show that the rapid algorithm dramatically reduces running time, and the efficient path-reconstruction operator can effectively extract narrow and long objects from noises.

Keywords: Mathematical Morphology, Morphological Filter, Path Opening, Reconstruction Operator

1 Introduction

Mathematical morphological filtering is one of the most interesting subjects of research in mathematical morphology. Morphological filters are nonlinear transformations that locally modify geometric features of images [1]. The basic Morphological filters are morphological openings and closings with given structuring elements. By using basic filters, we can build others with different filtering characteristics. But these filters present several inconveniences. In general, if the undesirable features are eliminated, the remaining structures will be changed. Recently, connected operators have become powerful tools for their particular characteristics. These operators do not remove some frequency components like linear filters or some shapes like median filters or morphological opening and closing. They can only remove connected components of the sets or fill connected components of the background[2,3]. Connected operators being able to simplify while preserving the contour information are very attractive for segmentation purpose. Intensive work has been done on the characterization of these transformations [4,5,6,7,8].

One of the most common filters with connected operator is filter by reconstruction. From a practical point of view, filters by reconstruction are built by means of reference image (mask) and a marker image include in the reference image. But in general, operation by reconstruction only thinks about grads or width, and often ignores orientation and shape information. While path openings, with more flexible structuring elements, can be adequate in the common situation where there exist narrow, locally oriented features in an image of interest[9,10]. Thus in this paper, a new connected operator that combined reconstruction with path openings was proposed. Based on the definition of the proposed operator, the properties were researched and discussed. Then an efficient algorithm was presented and an example was illustrated to confirm its effectiveness.

We conclude this section with an overview of this paper. In the next section we present a brief discussion of connected operator and path opening. In Section 3 a

* Corresponding author Email: caihui1982@hotmail.com

new connected operator named path-reconstruction operator was proposed. Section 4 is concerned with the efficient algorithm and application of the proposed filter. Section 5 gives the conclusion.

2 Connected Operator and Path Opening

Definition 1 *Flat Zones*: The set of *flat zones* of a gray-level function *f* is the set of the largest connected components of the space where *f* is constant.

The set of flat zones of a function constitutes a partition of the space. In the following, this partition will be called the partition of flat zones of a function.

Definition 2 *Connected Operators*: An operator acting on gray-level functions is said to be connected if, for any function *f*, the partition of flat zones of ψ(*f*) is less fine than the partition of flat zones of *f*.

Definition 3 *reconstruction*: the reconstruction of mask *I* from marker *J* is the union of the connected components of *I* which contain at least a pixel of *J*.

Definition 4 *Adjacencies*: Let *E* be a given set of points representing pixel locations. Define a *directed graph* on these points via a binary adjacency relation '→'. Specifically, *x*→*y* means that that there is an edge going from *x* to *y*. If *x*→*y*, we call *y* a *successor* of *x* and *x* a *predecessor* of *y*. These concepts are illustrated in Figure 1. Here *b*1*, b*2*, b*3 are successors of *a,* and *a*1*, a*2*, a*3 are the predecessors of *b*.

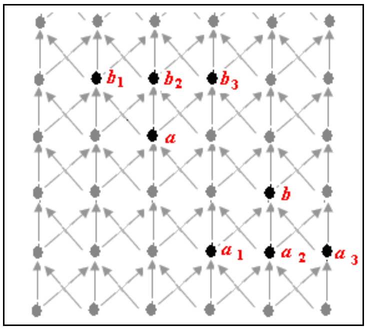


Figure 1 Adjacencies

Definition 5 *δ-path*: denote *path dilation* by $\delta(x)$, and $\delta(x)=\{y\in E: x\to y\}$ *L*-tuple $\boldsymbol{a}=(a_1, a_2\ldots a_L)$, is called a *path of length L* if $\forall k\in[1,L-1]$, $a_{k+1}\in\delta(a_k)$, Henceforth we refer to such a path as a δ-path of length *L*.

The set of δ-paths of length L contained in a subset X of E is denoted by $\prod_L(X)$.

Definition 6 *path opening*: We define the set $\alpha_L(X)$ as the union of all *δ-path*s of length *L* contained in *X*: $\alpha_L(X)=\cup\{\delta(a):a\in\prod_L(X)\}$

The operator $\alpha_L(X)$ has the algebraic properties of an *opening*, we call it *path opening*.

3 Path-Reconstruction Operator

As filters by reconstruction generally only thinks about grads or width, but often ignores orientation and shape information. Path openings can be adequate in the common situation where there exist narrow, locally oriented features, but path openings sometimes can not effectively eliminate undesirable features without affecting desirable ones. In order to more efficiently filter noise by length and orientation characters of objects, the advantages of filters by reconstruction was combined with path openings. Path-reconstruction operator was proposed just in this way.

Firstly, binary path-reconstruction was put forward, and based on the threshold decomposition method, binary path-reconstruction operator extended to gray level domain.

Definition 7 *Binary Path-reconstruction operator*: Let sets $I\in M$ and $J\in M$ be two binary images defined on the same discrete domain $M\subset Z^2$, and set *J* is a subset of *I*, $J\subseteq I$, this means $\forall x\in M$, $J(x)=1\Rightarrow I(x)=1$, the set *J* is called the marker image and the set *I* is called the mask image. Denote the set of all connected components by Ω, and Ω={ $I_1, I_2,\ldots I_n$,}, $I_1, I_2,\ldots I_n$ is connected component of set *I*, *n* is the number of connected components in set *I*. For an appointed*δ*-path length threshold *L*, now we may define path-reconstruction operator for binary image ω(*I*, *J*, *L*) as follow.

$$\omega(I,J,L)=\left\{x\in I_k \middle| I_k\in M, \alpha_j(I_k\cap J)\neq\Phi, j\geq L, k=1,2,\cdots n\right\} \quad (1)$$

In equation (1), $\alpha_j(I_k\cap J)=\cup\{\delta(a):a\in\prod_j(I_k\cap J)\}$, $\delta(a)$ is path dilation operator.

According to the definition of connected operator, it is clear that path-reconstruction operator for binary image is a kind of connected operator. And it is not difficult to establish that the operator ω(*I*, *J*, *L*) has the algebraic properties of an opening, specifically increasingness, anti-extensivity and idempotence, so path-reconstruction operator is a morphological filter.

On discrete domain, with threshold decomposition and superposition technique, binary operator having increasingness property can extend to gray level domain[5].

Definition 8 *gray Path-reconstruction operator*: Let functions *f* and *g* be two gray images defined on the same discrete domain *D*, denote the value range *V*={0,1,…,*N*-1}, functions $g \le f$, this means $\forall p \in D, g(p) \le f(p)$, the function *J* is called the marker image and the function *I* is called the mask image. For an appointed *δ*-path length threshold *L*, we may define path-reconstruction operator for gray image $\psi(f,g,L)$ as follow.

$$\psi(f,g,L)(p) = \max\{k \mid p \in \omega(T_k(f), T_k(g), L)\}, \quad p \in D \tag{2}$$

In equation (2), $T_k(f) = \{p \in D \mid f(p) \ge k\}$, $T_k(g) = \{p \in D \mid g(p) \ge k\}$.

From this definition, it is clear that binary operator is a specific case when *V*={0,1}.

4 Efficient Algorithm of Path-Reconstruction Operator and Application

In this section, we are concerned with both the binary and the gray-scale case, but the emphasis is put on gray path-reconstruction. Indeed, in the binary case, a straightforward efficient implementation of morphological path-reconstruction can be proposed as follows:

1) Construct the mask image and marker image in different restriction. Label the connected components of the mask image, i.e., each of these components is assigned a unique number. Note that this step can itself be implemented very efficiently by using algorithms based on chain and loops or queues of pixels.

2) Determine the labels of the connected components which contain at least a path whose δ-path of length not less than *L* in the corresponding regions of the marker image.

3) Remove all the connected components whose label is not one of the previous ones.

As mentioned earlier, such an algorithm could be extended to the gray-scale case by working on the different thresholds of the images. However, it would be extremely inefficient, making gray path-reconstruction a too cumbersome transformation to be used in practice. This is the reason why we are now interested in implementing this transformation as efficiently as possible.

In an attempt to reduce the number of scannings required for the computation of an image transform, we shall be concerned with an efficient algorithm. The breadth-first scannings involved are implemented by using a queue of pixels, i.e., a First-In-First-Out (FIFO) data structure: the pixels which are first put into the queue are those which can first be extracted.

Define the Adjacencies, for each pixel p four values: $\lambda^+[p]$ is the length of the longest path travelling downward from pixel p; $\lambda^-[p]$ is the length of the longest path travelling upward from pixel p; $\lambda[p]$is the length of the longest path passing through pixel *p*; flag[*p*] is the label of pixel p, flag[*p*]=0 means *p* is active.

For Adjacencies in Figure.1, the coordinate of *p* is (p^1,p^2), then we have the following equation:

$$\begin{cases} \lambda^-[p] = 1 + \max\{\lambda^-[p^1-1, p^2-1], \\ \qquad \lambda^-[p^1-1, p^2], \lambda^-[p^1-1, p^2+1]\} \\ \lambda^+[p] = 1 + \max\{\lambda^-[p^1+1, p^2-1], \\ \qquad \lambda^-[p^1+1, p^2], \lambda^-[p^1+1, p^2+1]\} \\ \lambda\ [p] = \lambda^-[p] + \lambda^+[p] - 1 \end{cases} \tag{3}$$

Algorithm 1:

1) Initialisation:

Set all pixels with $\lambda^+[p]$= $\lambda^-[p]$= $\lambda[p]$= flag[*p*]=0.

2) Define output image *O*, Scanning all the pixels in mask image *I* and marker image *J*, if *J*(*p*)>*I*(*p*), then *J*(*p*)=*I*(*p*).

3) Order the pixels in *J* based on histogram, then store in array *queue* by turns, and the array *GreyFirst_marker*[*N*] record the first pixel in gray level *N*.

4) Take out all the pixels *p*, *J*[*p*]=*m* in the *queue* by

turns.

For each pixel from *GreyFirst_marker*[*m*+1]-1 to *GreyFirst_marker*[*m*]: if flag[*p*]=0, recompute $\lambda^-[p]$ according to equation (3); if $\lambda^-[p] \geq L$, $\lambda^-[p] = L$, if $\lambda^-[p]$ did not change, set flag[p^1-1, p^2-1]=1, flag[p^1-1, p^2]=1, flag[p^1-1, p^2+1]=1.

For each pixel from *GreyFirst_marker*[*m*] to *GreyFirst_marker*[*m*+1]-1: if flag[*p*]=0, recompute $\lambda^+[p]$ according to equation (3); if $\lambda^+[p] \geq L$, $\lambda^+[p] = L$, if $\lambda^+[p]$ did not change, set flag[p^1+1, p^2-1]=1, flag[p^1+1, p^2]=1, flag[p^1+1, p^2+1]=1; if $\lambda[p] \geq L$, set flag[*p*]=1,*O*[*p*]=m., and transfer function *GetRoot*(p), details of *GetRoot*(p) was put forward in **algorithm 2**.

5) Go to step 4, set *m*-1, until all the pixels in image *J* have been scanned.

Algorithm 2:

fifo_add(p): puts the (pointer to) pixel *p* into the queue.

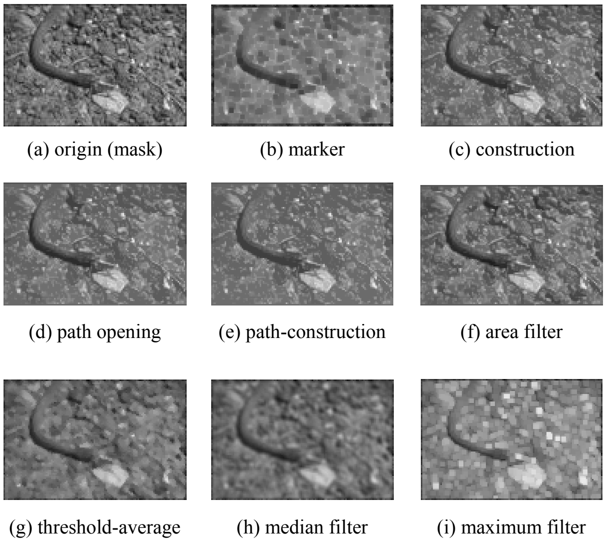
fifo_first(): returns the (pointer to) pixel which is at the beginning of the queue, and removes it.

Fifo_empty(): returns true if the queue is empty and false otherwise.

```
Int   GetRoot(int p)
{ Remove pixels in neighbourhood of p into array fifo.
    While fifo_empty()=FALSE
    {p←fifo_first()
     If I[q] ≥m and flag[q]=0
        O[q]= m and flag[q]=1
        fifo_add(q) }
}
```

The above algorithm constitutes a very clear improvement with respect to the algorithm directly using threshold decomposition and superposition. For an image 256×256, if directly use threshold decomposition and superposition, the running time of the algorithm is about 3000ms, but the running time of the proposed algorithm with a FIFO queue is less than 80 ms.

Figure 2(a) is the original image and denotes mask. Image resolution is 120×80.Figure 2(b) is the marker image by morphological opening operation. Figure 2(c) is the result of reconstruction filter. Figure 2(d) is the image enhanced by path opening with *L*=12. Figure2(e) is the result of proposed operator with *L*=12. Figure 2(f) is enhanced by area filter. Figure 2(g) is the result of median filter. Figure 2(h) is the result of threshold-average filter. . Figure 2(i) is the result of max-value filter. Figure 3 (a,b,c,d,e,f,g,h,i) are the segmentation results corresponding to Figure 2 (a,b,c,d,e,f,g,h,i). It can be seen that by the proposed method most of the noise is removed and the object is extracted correctly, however other methods can not get correct results effectively.

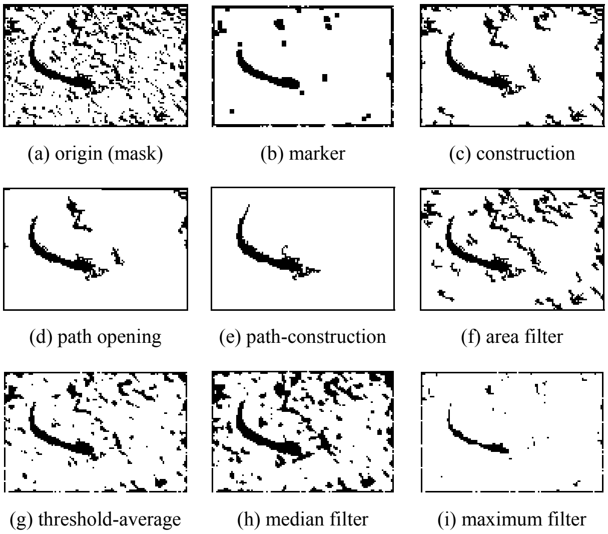


(a) origin (mask) (b) marker (c) construction

(d) path opening (e) path-construction (f) area filter

(g) threshold-average (h) median filter (i) maximum filter

Figure 2 Image enhancement results



(a) origin (mask) (b) marker (c) construction

(d) path opening (e) path-construction (f) area filter

(g) threshold-average (h) median filter (i) maximum filter

Figure 3 Segmentation results

In order to better compare the enhancing algorithms in Figure 2, here we denote a quality metric which includes flatness of background (σ_p) and signal-to-noise ratio (SNR). Background abnormality can cause problems for signal detection or incorrect measurement of the local background leve1, so the flatness of the background signal reflects these factors；The SNR quantifies how well one can resolve a true signal from the noise of the system. This metric are defined by the following equation:

$$\begin{cases} \sigma_p = \sqrt{\sum_{i=1}^{n}(p_i - B_p)^2 / n} \\ SNR = (\left|S_p - B_p\right|)/\sigma_p \end{cases} \qquad (4)$$

It is clear that when the σp is low, the intensity of background is uniform, and when the SNR is high，the intrinsic variation in the data is low and confidence in the accuracy of the data is high. Figure 4 and Figure 5 give the comparison results. It is shown that the proposed algorithm has the lowest σp and the highest SNR.

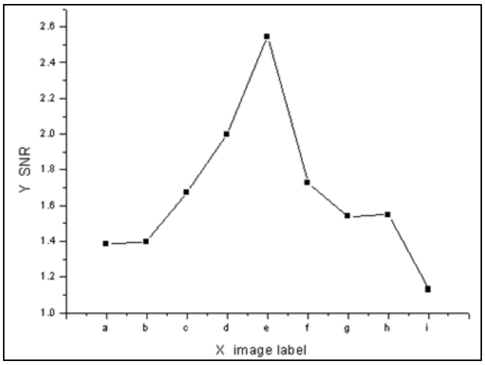


Figure 4 Evaluation by SNR

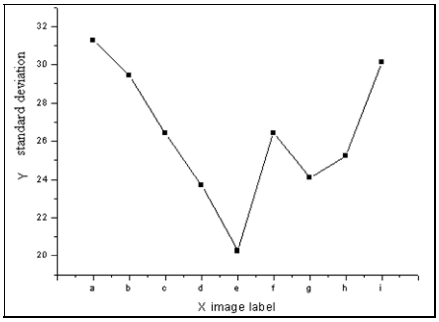


Figure 5 Evaluation by background standard deviation

5 Conclusions

In this paper, we explored the theory of path-reconstruction operator on binary and gray level image. Then to improve the running efficiency, an efficient algorithm was proposed. Results show that path-reconstruction operator effectively utilized grads information, length and orientation attributes of object, and it can be used in real-time application such as image segmentation and image filter.

References

[1] R.Ivan, Terol-Villalobos. “Openings and Closings with Reconstruction Criteria: A Study of a Class of Lower and Upper Levelings,” Journal of Electronic Imaging, Vol.14, No.1, 2005, pp.013006-1 ~013006-11

[2] k. Georgios, H.F. Wilkinson, “Mask-Based Second-Generation Connectivity and Attribute Filters,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.29, No6, 2007, pp:990~1004

[3] J. Serra, P. Salembier , “Connected Operator and Pyramids,” Proceedings of SPIE: Image Algebra and Morphological Image Processing IV, Vol.2030, 1993, pp:65~76

[4] P. Salembier, J. Serra, “Flat zones filtering, connected operators, and filters by reconstruction,” IEEE Transactions on Image Processing, IEEE Signal Processing Society, 1995, Vol.4, No.8,, pp.1153~1160

[5] J. Serra, “A Lattice Approach to Image Segmentation,” Journal of Mathematical Imaging and Vision, Vol.24, No.4, 2006, pp. 83~130

[6] U. Braga-Neto, J. Goutsias, “Grayscale Level Connectivity: Theory and Applications,” IEEE Transactions on Image Progressing, Vol.13, No.12, 2004, pp.1567~1580

[7] YANG Zhaohua,PU Zhaobang,QI Zhenqiang, “Novel Multiscale Morphological Filtering Method for Preserving the Details of Images,” Journal of Optoelectronics • Lase, Vol.14, No.8, 2003, pp:862~865

[8] P. Salembier, J. Ruiz, “Connected Operators Based on Reconstruction Process for Size and Motion Simplification,” Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol.4, 2002,pp: IV3289~IV3292

[9] H. Heijmans, M.Buckey, H.Talbol, “Path Openings and Closings,” Journal of Mathmatical Imaging and Vision, Vol.22, 2005, pp.107~119

[10]]M. Buckey, H.Talbol, “Efficient Complete and Incomplete Path Openings and Closings,” Imaging and Vision Computing, Vol.25, 2007, pp.416-425

Circular Object Recognition Based on Invariant Features

Aijun Chen

College of Mechanical and Electrical Engineering, Northeast Forestry University, Harbin, Heilongjiang 150040, China
Email: jschenaj@163.com

Abstract

A circular object recognition method utilizing a group of geometric feature parameters is presented. First, Canny operator is used to extract edges and the edge points are traced to form series of sequential sets. All points in each set are fitted to a polygon. Then, the fitted polygon is normalized to obtain invariant features and their corresponding parameters are calculated. In the end, those targets with the feature parameters satisfying certain conditions are recognized as circular objects. Experiments on synthetic images and real images and comparison with RHT method show that the proposed method has the merits of fast recognition rate and high recognition efficiency.

Keywords: Object Recognition, Circle Detection, Invariant Feature, Polygon Fitting, Edge Detection

1 Introduction

In everyday experience, many objects that we perceive are in circular form. The automatic recognition and localization to them in an image reliably and efficiently is therefore an essential task in image analysis applications.

One of the commonly used methods for circular object recognition is Hough transform (HT), which is robust to random noise, and can withstand certain degree of occlusion and boundary defects. So HT has long been considered the best unique technique for the recognition of geometrical shapes in images [1]. The technique can deal with noise-corrupted images and can be used successfully when data are partially occluded. However, HT has some disadvantages when it works on a discrete image. The large amount of storage and computing power required by HT are the major defects, which cause it difficult to be used in real-time applications. In recent yeas, development has been made in this area of circle detection based on HT [2~5], but it is not still widely used due to slower speed. In addition, the geometric symmetry is used as a feature to recognize imperfectly circular objects [6~8], but it is invalid if circular objects are some distorted, i.e. incompletely symmetric. The fuzzy recognition methods [9, 10] overcome the above-mentioned shortage; however, they need the priori knowledge about the number of circular objects to be recognized in digital images.

In this paper, an approach is proposed which uses a group of geometric parameters to recognize circular objects. There's no need to preknow the number of circular objects with our method, furthermore, both circular perfect objects and circular targets with partial distortion and little imperfection can be recognized well.

2 Algorithm Description

The general idea is described as follows. Canny operator is used to detect edges in an original image so that a binary edge image is created, in which all edges are traced to obtain several edge point sets representing complete contours of objects. Then polygon fitting is performed for each curve made up of a certain edge point set and the fitted polygons are normalized. Take the vertexes of each normalized polygon as feature points and extract geometric features with invariant characteristics of translation, rotation and scale change. In the end, the parameters are obtained which are resulted from the ratios of the geometric feature values to corresponding ones in a certain unit circle. If geometric parameters of a target are larger than the given thresholds, it is recognized as a circular object.

2.1 Edge detection

There are many edge detection methods, among which Canny operator is a better one. Canny proposed three criteria for evaluating edge detection performance. They are: 1) Good detection. There should be a low probability of failing to mark real edge points, and low probability of falsely marking non-edge points. Since both these probabilities are monotonically decreasing functions of the output signal-to-noise ratio, this criterion corresponds to maximizing signal-to-noise ratio. 2) Good localization. The points marked as edge points by the operator should be as close as possible to the center of the true edge. 3) Only one response to a single edge. This is implicitly captured in the first criterion since when there are two responses to the same edge, one of them must be considered false. However, the mathematical form of the first criterion did not capture the multiple response requirement and it had to be made explicit.

On account of the fact that edges detected by Canny operator have virtues of singe pixel wide and good continuity, it is adopted to detect edges in this paper.

After the edges are detected by Canny operator, the gray-level value of pixels located at edges is set to 1 and those of others to 0. Consequently, a binary edge image is made.

2.2 feature extraction

A series of edge points can be obtained after edges are detected and these edge points describe shape properties of object contours in the image. The computation of recognition algorithm will be time consuming if we take these edge points as feature points of recognition objects. For the sake of speeding up recognition, these edge points must be cut down. In this paper, all edge points are traced to form edge point sets and then curves resulted from the edge point sets are fitted to polygons, whose vertexes are the final feature points.

Building edge point sets: In this paper, an 8-direction tracing method with 4-neighbor precedence [12] is adopted to trace edges. A tracing start point is found by scanning pixels row by row. That is, scan in a binary edge image from up to down and left to right until a pixel with gray-level value 1 is met. The pixel just is the start point. And the algorithm of building edge point sets is as follows.

Step 1. Select a start point, erect a new coordinate chain and add the coordinates of the point to the chain.

Step 2. Set the gray-level value of the edge point from 1 to 0 (To keep from repeatedly tracing) and take the point as the current one.

Step 3. Find the next edge point.

Think about the four 4-neighbor points of the current point in a clockwise direction. If there are such points that their gray-level value is 1, set the first point with gray-level value 1 as the current point and add its coordinates to the chain, then go step2. Otherwise, continue.

Consider the four 8-neighbor diagonal points of the current point in a clockwise direction and do it as above-mentioned.

Step 4. Set the start point as the current point and do the same work as that in Step 3.

Step 5. If there still are points with gray-level value 1, go step1. Otherwise, stop.

Several edge point sets can be obtained, which represent complete contours of objects, after the binary image is processed with the aforementioned tracing algorithm.

Polygon fitting and normalization: The idea of polygon fitting is as follows:

Take the first point $S(x_s, y_s)$ and the last point $E(x_e, y_e)$ in an edge point set $\boldsymbol{P}$ as two end points of a virtual segment and calculate the distance d_k between each of the rest points in the set $\boldsymbol{P}$ and the virtual segment.

$$d_k = \frac{\left|x_k(y_s - y_e) + y_k(x_e - x_s) + y_e x_s - y_s x_e\right|}{\sqrt{(x_s - x_e)^2 + (y_s - y_e)^2}} \tag{1}$$

$$k = 2, 3, \cdots, n-1$$

with n being the number of edge points located at the virtual segment (including the two end points). Set $d_{\max}$ be the longest one among d_k and its corresponding edge point is $M(x_{\max}, y_{\max})$. $d_{\max}$ is given by

$$d_{\max} = \max(d_k) \qquad k = 2,3,\cdots,n-1 \tag{2}$$

The virtual segment is a side of a polygon if $d_{\max}$ is little than a given threshold d_{th}; otherwise, set the point M as one end point of the virtual segment, that is, $(x_e, y_e) \leftarrow (x_{\max}, y_{\max})$, keep another end point invariant, and calculate the distance d_k between the virtual segment and each point between point S and point E with Eq.(1). And then judge if $d_{\max}$ is little than d_{th}. If it does, take the new point M and the first point S in the set $\boldsymbol{P}$ as the two end points of a side of the polygon. Set point M as the start point of a new virtual segment, that is, $(x_s, y_s) \leftarrow (x_{\max}, y_{\max})$, and the last point in the set P as another end point of the virtual segment. Continue as the above-mentioned procedure. Otherwise, set the new point M as the end point of the virtual segment, keep the start point invariant and calculate with Eq.(2) until $d_{\max}$ is little than d_{th}.

A sequential set $\boldsymbol{V}$ can be obtained which consists of the vertexes of the fitted polygon after an edge point set is processed according to the aforementioned method. However, rotation, translation and scale change maybe occur due to some factors, such as imaging distance, orientation and position. For the sake of invariant to rotation, translation and scale change, a fitted polygon has to be normalized. Here, the longest distance among the centroid of an edge point set and the vertexes of the fitted polygon is selected as the normalization factor.

Let ($\bar{x}$, $\bar{y}$)be the coordinates of the centroid of an edge point set. They can be described by

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \ , \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{3}$$

with N being the number of points in an edge point set and (x_i, y_i) the coordinates of the edge point p_i. The normalization factor D can be written as

$$D = \max|v_i - C| = \max(\sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}) \tag{4}$$

$$i = 1,2,\cdots,M$$

with M being the number of vertexes in the fitted polygon.

The corresponding relationship between the coordinates (x_i', y_i') of the feature point v_i' after the polygon is fitted and those before is described by

$$x_i^{'} = (x_i - \bar{x})/D \ , \quad y_i^{'} = (y_i - \bar{y})/D \tag{5}$$

$$i = 1,2,\cdots,M$$

Feature extraction: After finishing the normalization, the rest work is to extract geometric features from the normalized polygon. These features are perimeter, average polar distance and area.

(1) Perimeter.

The perimeter of a polygon is defined as

$$L = \sum_{i=1}^{M} l_i = \sum_{i=1}^{M} \sqrt{(x_i^{'} - x_{i+1}^{'})^2 + (y_i^{'} - y_{i+1}^{'})^2} \tag{6}$$

$$i = 1,2,\cdots,M$$

with l_i being the Euclidean Distance between the two feature points $v_i^{'}$ and $v_{i+1}^{'}$; $(x_i^{'}, y_i^{'})$ and $(x_{i+1}^{'}, y_{i+1}^{'})$ respectively denote the coordinates of $v_i^{'}$ and $v_{i+1}^{'}$. While i is equal to M, set i+1 be 1.

(2) Average polar distance.

$$\bar{d} = \frac{1}{M}\sum_{i=1}^{M} d_i = \frac{1}{M}\sum_{i=1}^{M}\sqrt{x_i^{'2} + y_i^{'2}} \tag{7}$$

where d_i is the Euclidean Distance between the feature point $v_i^{'}$ and the centroid $C^{'}$. The coordinates of $C^{'}$ are (0, 0).

(3) Area.

$$A = \frac{1}{2}\sum_{i=1}^{M}(x_i^{'} y_{i+1}^{'} - x_{i+1}^{'} y_i^{'}) \qquad i = 1,2,\cdots,M \tag{8}$$

While i is equal to M, set i+1 be 1.

If the contour of an object is a standard circle, it must be a unit circle after normalized. So the feature parameters are obtained by calculate the ratios of the values of the normalized geometric feature to their corresponding ones in a unit circle. The feature parameters are as follows:

(1) Perimeter Ratio.

$$c_1 = L/(2\pi r) = L/(2\pi) \tag{9}$$

(2) Distance Ratio.

$$c_2 = \bar{d}/r = \bar{d} \tag{10}$$

(3) Area Ratio.

$$c_3 = A/(\pi r^2) = A/\pi \qquad (11)$$

where r being the radius of a unit circle, that is $r = 1$.

An object is recognized as a circular target if the three feature parameters resulted from it satisfy conditions as Ineq.(12).

$$c_1 > T_1 \text{ and } c_2 > T_2 \text{ and } c_3 > T_3 \qquad (12)$$

where T_1, T_2 and T_3 respectively denote the thresholds of c_1, c_2 and c_3 and they are set by experiments.

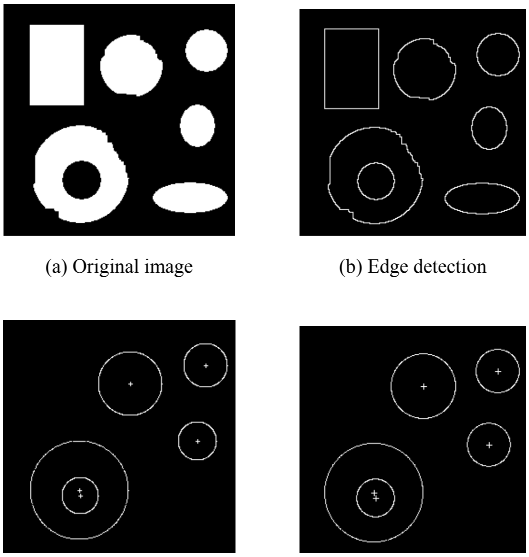
3 Experimental Results

Our experiments are performed on a Pentium(R)1.60GHz computer with Windows XP operating system using Visual C++ 6.0 program language. The images to be processed were divided into two groups: One consists of synthetic images and another are of real images. In our experiments, d_{th} is equal to 3; the values of T_1, T_2 and T_3 are 0.91, 0.90 and 0.80, respectively. For the purpose of comparison, we apply our proposed approach and the RHT method to each image individually. In this paper, the performances of the two methods are evaluated by running the programs 50 times, because running results with RHT method may be different due to randomly sampling.

The first experiment was tested on the synthetic images. The original image is shown in Figure 1a. It consists of one perfectly circular object, one circular object with some flaws, one circular object with little deform similar to an ellipse object, one ellipse object, one rectangle object and one cirque object with some flaws in the outside boundary. The binary edge image with edges detected by Canny operator is shown in Figure 1b. The recognition results to Figure 1a with the RHT method and the proposed method are shown in Figure 1c and Figure 1d, respectively. Both methods can correctly recognize the five circular objects because the edge of the synthetic images is simple. At the same time, we can see both methods are not sensitive to edge distortion and imperfect. The centers and radii of recognized circular objects in Figure 1a are shown in Table 1.

Table 1 Recognition Results on the Synthetic Image

RHT method		Our method	
Centers of objects	Radii	Centers of objects	Radii
(84, 187)	54	(84, 188)	55
(223, 50)	24	(224,51)	24
(140, 70)	35	(140, 69)	36
(85, 193)	20	(86, 194)	21
(214, 133)	21	(214, 134)	22



(a) Original image

(b) Edge detection

(c) Recognition result

(d) Recognition result with our method

Figure 1 Synthetic image and its recognition results

The second experiment was curried out on real images, shown in Figure 2a. The binary edge image with edges detected by Canny operator is shown in Figure 2b. With RHT method, the four circular objects are correctly recognized. Meanwhile, 10 false objects occur in the recognition result (as shown in Figure 2c). In Figure 2d gives the recognition result with the proposed method. All of circular objects in Figure 2a are correctly recognized with no false alarm. We can see from the results that as to real images, when their edges become more complicated, the false alarm rate of recognition with RHT method will greatly ascend, while there's no false alarm or low false alarm rate and the recognition rate still is highly kept with the proposed method. The centers and radii of recognized circular objects in Figure 2a are shown in Table 2. Figure 3 shows some other images and circle recognition results with our method.

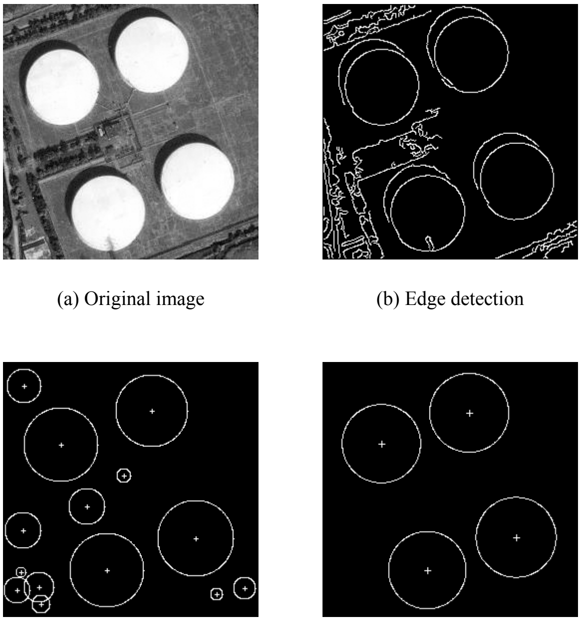
The execution time required in each method is measured in terms of seconds and it is obtained from the average in run 50 times for each image without considering the fact that the recognition result may be incorrect. The performance comparison between the RHT method and our proposed method is show in Table 3.

Table 2 Recognition Results on the Real Image in Figure 2

RHT method		Our method	
Centers of objects	Radii	Centers of objects	Radii
(149, 49)	36	(149, 49)	37
(104, 210)	37	(105, 209)	37
(58, 82)	37	(59, 82)	37
(195, 178)	37	(194, 177)	38

Table 3 Comparison of the Elapsed Time

Image No.	RHT method(s)	Our method (s)	RHT method/our method
Figure 3(a)	13.840	0.096	144.16
Figure 4(a)	40.934	0.148	276.58
Figure 5(a)1	86.995	0.155	561.26
Figure 5(a)2	32.468	0.246	131.98
Figure 5(a)3	25.097	0.124	202.40

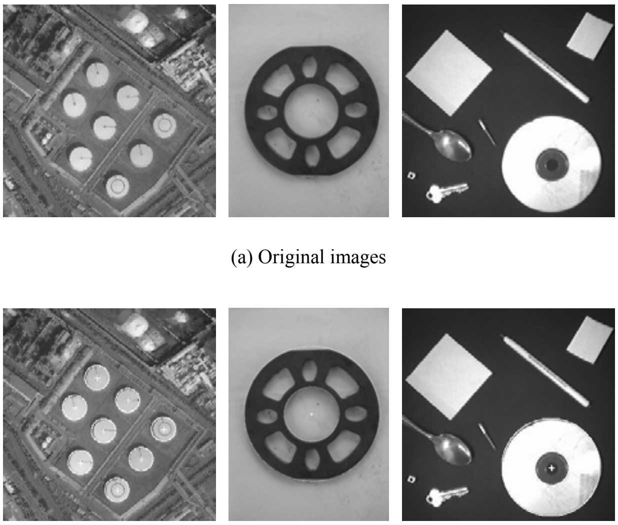

(a) Original image

(b) Edge detection

(c) Recognition result with RHT method

(d) Recognition result with our method

Figure 2 Real image and its recognition results


(a) Original images

(b) Recognition results with our method

Figure 3 Recognition results on some other real images

According to the experimental results in Table 3, Figure 3 to Figure 5, it is noticed that the proposed method is faster than the RHT. Additionally, with our method, multiple circular objects with different size can be simultaneously recognized. Moreover, the results have no reliance with one another. The proposed method performs well when it is used to recognize those circular objects with little distortion and imperfect.

4 Conclusions

We have proposed an approach for automatically recognizing circular objects which has the characteristics of little required data quantity and little memory requirements. What's more, the extracted features are invariant to translation, rotation and scale change. Experimental results on synthetic images and real images demonstrate that the proposed method is rapid and efficient in recognizing circular objects.

References

[1] J. Illingworth, J. Kittler, "A Survey of the Hough Transform," Computer Vision, Graphics, and Image Processing, 44(1), 1988, pp. 87-116

[2] L. Xu, E. Oja, and P. Kultanan, "A New Curve Detection Method: Randomized Hough Transform (RHT)," Pattern Recognition Letters, 11(5), 1990, pp. 331-338

[3] J. Qu and L. Gan, “The Application of Grads Hough Transformation in Circle Detection,” Journal of East China Jiaotong University, 24(1), 2007, pp. 101-104

[4] Q. Li and Y. Xie, “Randomised Hough Transform with Error Propagation for Line and Circle Detection,” Pattern Analysis and Application, 6(1), 2003, pp. 55-64

[5] S. H. Chiu and J. J. Liaw, “An Effective Voting Method for Circle Detection,” Pattern Recognition Letters, 26(2), 2005, pp. 121-133

[6] C. Ducottet, J. Daniere, M. Moine, et al, “Localization of Objects with Circular Symmetry in a Noisy Image Using Wavelet Transforms and Adapted Correlation,” Pattern Recognition, 27(3), 1994, pp. 351-364

[7] C. T. Ho and L. H. Chen, “A fast Ellipse /Circle Detector Using Geometric Symmetry,” Pattern Recognition, 28(1), 1995, pp. 117-124

[8] H. S. Kim and J. H. Kim, “A Two-Step Circle Detection Algorithm from the Intersecting Chords,” Pattern Recognition Letters, 22(6/7), 2001, pp. 787-798

[9] N. D. Rajesh, “Generalized Fuzzy C-shells Clustering and Detection of Circular and Elliptical Boundaries,” Pattern Recognition, 25(7), 1992, pp. 713-721

[10] Y. Man and I. Gath, “Detection and Separation of Ring-shaped Clusters Using Fuzzy Clustering,” IEEE Trans. on Pattern Analysis and Machine Intelligence, 16(8), 1994, pp. 855-861

[11] N. N. Zheng, Computer Vision and Pattern Recognition, Beijing: National Defense Industry Press, 1998

Fractal Image Compression Algorithm Improvement Study

Baosuo Wu　Wenbo Xu　Jun Sun

School of Information Technology, Jiangnan University, Wuxi, 214122, China
E-mail:wbsnow@126.com

Abstract

Fractal image compression technology has such characters as high compression rate, high Signal-to-Noise rate and better vision effect etc. , thus becomes an eye-catching topic in the field of digital image science. However, the issue of long compression time poses a hindrance to the development of this technology already becomes a well known fact, the task of how to shorten the compression time attracts more and more scholars. In this paper, we try to apply the fuzzy clustering algorithm and the QPSO algorithm to the fractal image compression in an effort to shorten the compression time. The experiment shows that this approach has significantly shortened image compression time.

Keywords: fuzzy clustering; QPSO; fractal image compression

1　Introduction

Fractal image compression is introduced by Barnsley and others in 1980s. It is famous for its high rate of compression, but the long compression time in fractal image compression has prevented its development. Many scholarships tried to improve it. This paper has carried out comparison study to the fractal encoding algorithms, and has combined fuzzy clustering algorithm with QPSO algorithm to optimize fractal image compression. Firstly, we use fuzzy clustering algorithm to class the son-block from image segmentation. Secondly, we treat the number of category created by image segmentation as the number of category in QPSO. Then we use the QPSO algorithm to clustering father-block and use search in category instead of search in all. In doing so, we can reduce the matching times and shorten coding time. Thus we enhance the fractal encoding efficiency.

2　Traditional Fractal Image Compression Algorithm

Fractal encoding technique is based on the fractal theory, utilize the self-conform between the two parts of the image to coding and decode. It uses PIFS and spell-paste theorem to search a series of affine retracting transformation to approach the original image. The key link of fractal image compression algorithm is searching the most matching range blocks in the same domain and corresponding affine transformations, and replacing the storage of original image with transform coefficient in order to get better compression ratio and lower capability demand. The related concepts and theorems are fixed point theorem, affine transformation theorem, iterative function system, attractor theorem, spell-paste theorem etc.

The main steps of fractal image compression algorithm coding:

(1) Image segmentation: Divided the original gray scale image into un-intersect range block Ri(i=1,2,⋯), (R X R dimension)

(2) Establish matching block: Using 2R×2R intercept window establish matching block Dj(j=1,□,m) in original image with step width △h in level and △v in upright. All the Dj form the search space SD={D1,⋯ ,Dm} (m=1,2,⋯).

(3) Search matching: In the matching block, we use Mean Square Error (MSE) to search Dj for each Ri. After proper affine transformation, we get wi to approach Ri. That is wi(Dj)→Ri satisfies d(Ri,wi(Dj))=||Ri-(si*(mi(G(Di))) +oi||2, G is a geometric transformation. In doing so, we accomplish the space retraction from matching to range block and mi(i=0,⋯,7) is one of eight symmetry rotation transformations, si and oi are contrast

and measurement instify factor. Note fractal encoding: To each Ri we need to find a transform set (xi,yi,mi,si,oi) (xi,yi is the coordinate of oi) which satisfies d(Ri,wi(Dj)) =||Ri-(si*(mi(G(Di)))+oi||2. Then we just quantization coding the record and can get the fractal code for each Ri.

Decode: When decoding, we just treat one gray image (the same size with the image waiting for resume) as the original decode image. Then we fetch the fractal code and the parameter of affine transformation ωi corresponding each coding block Ri and use the same order in coding process to effect on original image. After repeated iterative, we can get a fixed point set (attractor) and this is the final decode image.

3 An Improvement on the Fractal Image Compression Algorithm

3.1 Fuzzy clustering algorithm

Fuzzy clustering algorithm which first introduced by Professor Chen Fanwu is as follows.

3.1.1 The math base of fuzzy clustering algorithm

Usually, we can not certain how many classes we should partitioning a given sample set especially the multifarious nature images. How to class a gray scale vector sample set is a difficult problem. To avoid jamming, we use fuzzy cluster class method to deal with image. We confirm the final number of category by the arithmetic's objective function. And the math base of the unknown number of soft category in sample set is to find out a proper sort matrix U and sort center V to cluster criterion:

$$J_m(U,V) = \sum_{k=1}^{n} \sum_{i=1}^{m} (u_{ik})^t \| x_k - v_i \|^2 , \tag{1}$$

(Umxn={uik},m—number of category,t>1)

And when it get the least function value, we get the best U* and V* so that Jm(U*, V*)=min[Jm(U, V)]. In this paper, we cite the related conclusion and fuzzy cluster sort process directly. The detail information about consequence and analysis please refer to the references.

For a sample set X ={x1,x2,…xn}, there are many different subdivision methods. But we have:

Proposition 1, The best subdivision is which can make the objective function getting the least value.

Proposition 2, For any given category number m, the subdivision which corresponding to the least objective function may be not the best one.

Proposition 3, For the fuzzy equivalence relation matrix R*, every category number mλ(1 < mλ< n) certainly correspond a best subdivision J (mλ, Cλ);but J(m*λ, C*λ) = min{J(mλ, Cλ)} correspond the best subdivision C*λ which is under the category number m*λ. That is toward such soft sort problem, every possible category number m (correspond hart sort) has a best subdivision. And the one which correspond the best category number m* is the best subdivision to sample set.

3.1.2 How to realize the fuzzy clustering algorithm

(1) R* is a fuzzy equivalence relation matrix created by sample set X={x1,x2,…xn}. Firstly, we structure fuzzy consistent relation matrix Rmxn={aij} under some distance definition. And aij=1-μdij (parameterμmade 0<aij<1)

$$\mathrm{d}_{\mathrm{ij}} = \sum_{k=1}^{n} | x_{i,k} - v_{i,k} | \tag{2}$$

(n is sample amount)

(2) Structure fuzzy equivalence relation matrix: Firstly, using complex method to get exterior product R*:R*= R2=R⊗R={aij*} (aij*=(ai1∧a1j)∨(ai2∧a2j)∨…(ain∧anj)) between R, the operator ∧ denote choosing the little one and the operator ∨ denote choosing the bigger one. Let R* =RL = R⊗R⊗R…⊗R (aggregately L), R* is fuzzy equivalence relation matrix.

(3) Setup intercept threshold θi, we line up the nonzero elements in R* from the smallest one, 0<θ1<θ2…θn<1.

(4) Doing best cluster under given threshold: letλ1=θ1, and

$$\mathrm{a}\lambda_1, \mathrm{ij} = \begin{cases} 1, \mathrm{a}_{\mathrm{ij}}* \geq \theta_1 \\ 0, \mathrm{a}_{\mathrm{ij}}* < \theta_2 \end{cases}$$

Then we can get Rλ1;and from Rλ1 we can get the best category number mλ and the best subdivision Cλ1. So we can get cluster objective function from the following formula Jλ1:

$$J(m,C)=\sum_{i=1}^{m}\sum_{k>i>1}\|a_i^*-a_k^*\|^2\,, \qquad (3)$$

k,i∈Cj,(m is category number,a* is a vector in R*)

(5) Back to step 4, letλ2=θ2,…λn=θn,and following the method of step 3, we can get Jλ2 ,Jλ3,… Jλn;

(6) Fetch the least one of all cluster objective function to be the best subdivision.

$$J_{\lambda*}(m*,C*)=\min(J_{\lambda1}(m,c),J_{\lambda2}(m,c),)\ \ldots J_{\lambda n}(m,c))$$

,and C* is the best subdivision,m* is the best category number. Because of Rλcorresponding Jλ, so if column or row vector in Rλ*is same, we put them into the same category, that is cluster analysis.

3.2 QPSO algorithm

QPSO(Quantum—behaved Particle Swarm Optimization) is an atom group evolution algorithm which based on quanta action. Using the concept of colony and evolution and according as adapt of individual (atom), QPSO treat each individual as an atom which is no weight and no volume and it fly in stated speed in the search space. The speed is adjusted by individual and colony fly experience. Each particle stand for a position in Nd dimension space and exposed to the following two aspect adjust particle's position.

(1) The best position as yet of each particle

(2) The best position of particle group.

Each particle i has the following information,

(1) xi=(xi1,xi2,⋯,xid):the current place of particle

(2) Pi=(Pi1,Pi2,⋯,Pid):stand for the best adaptable value of particle i, viz. pbest.

(3) Pg=(Pg1,Pg2, ⋯ ,Pgd):stand for the best adaptable value of particle group, viz. gbest.

Evolution formula of particle,

mbest=$\frac{1}{M}\sum_{i=1}^{M}Pi$=($\frac{1}{M}\sum_{i=1}^{M}pi1,\ldots,\frac{1}{M}\sum_{i=1}^{M}pid$)

Pid=ψ*Pid+(1-ψ)* Pgd,ψ=rand

xid=Pid±α*|mbestd-xid|*$\ln(\frac{1}{u})$

The mbest stand for the particle group pbest's middle position and Pid is a random point between Pid and Pgd. α, a retraction extension coefficient of QPSO, is a key parameter of QPSO convergent.

Commonly, we chooseα=(1.0 - 0.5) (MAXITER - T) / MAXITER + 0.5.

The basic steps of algorithm:

1) Scavenge characteristic vector;

2) Initialization (cluster center, part excellent, global excellent)

For T =1:MAXITER

3) Clustering the sample under the Euclid geometry distance.

4) Account mbest and the adaptable function value with the evolution formula of particle.

5) Updating the part excellent pbest and global excellent gbest.

6) Account random point by the evolution formula of particle.

7) Updating the center vector of particle by the evolution formula of particle.

End

Repeat the step 2 to 6 until reach the iterative number or given condition.

3.3 Fractal image compression algorithm based on fuzzy cluster algorithm and QPSO algorithm

We choose search window from original image and toward each sub block of search window, we just choose the mean gray scale to be the sample point of cluster. We clustering the sample set according to fuzzy cluster algorithm. Then we can get the best partition of sample set and the best subdivision number of sub block sample set. Using the subdivision number of sub block sample set, we can cluster father block sample set by QPSO algorithm. After the clustering, the sub block just matching search in the same category's father block. We can use same category's search instead of global search in order to shorten search time and increase the coding efficiency. Because of the sub block partition by fuzzy cluster algorithm is the best algorithm, using QPSO cluster to father block is more reasonable under the best category number. So that the matching of sub block and father block is more nicety and the fractal code we get is excellent. Then we can get better compression effect.

The basic steps of fractal image compression algorithm based on fuzzy cluster algorithm and QPSO algorithm:

(1) Divide original image into some no superpose search window, and the size is $\omega*\omega$.

(2) To the sub block R(4*4) in each search window, we use fuzzy cluster optimize algorithm to cluster. And the cluster number is n. Just record the getting cluster centerσi (i=1,2,···,n).

(3) From the cluster number n, we cluster the father block in search window by QPSO algorithm. Because of the cluster number is from fuzzy cluster optimize algorithm, the result of father block cluster is more reasonable and effective and the cluster constringency speed is quicken.

(4) Matching search each sub block and the father block in the same category, we can get the fractal code s and o from the formula $d(Ri,wi(Dj))=||Ri-(si\times(mi(G(Di))) +oi||2$.

(5) By doing matching search in each window of original image, we can complete the coding procedure.

4 The Result of Emulation Experiment

We experiment the standard image Lena and Peppers by the algorithm refer to this paper. From the experiment result we can know under the same compression ratio, our algorithm is great enhanced in compression speed and shortened in compression time compare with traditional fractal image compression algorithm.

The result of experiment is shown in table 1.

Table 1 experiment results

Test image (256×256)	Lenna		Peppers	
Algorithm	Traditional algorithm	Our algorithm	Traditional algorithm	Our algorithm
PSNR(dB)	34.18	32.75	34.48	33.96
Coding time (m)	20.5	5.6	20.6	4.5
Compression ratio	24.16	24.28	24.28	24.30

5 Conclusion

From table 1, We can get the conclusion that our algorithm's compression ratio is almost the same with traditional algorithm, but we have a shorten compression time and an enhanced compression speed. Although the increased calculative complicacy of fuzzy cluster optimize algorithm, it has a enhanced compression speed. Take the real-time of image compression and transmission into account; we considered, for enhanced speed, a little complexity is acceptant. And considered the algorithmic optimize, how to reduce the algorithmic complexity and optimize algorithm will be the key point of our next research.

References

[1] Yuandaobo. Image information compress [M]. Beijing: Science publishing company,2004

[2] SUN J,XU WB.A Global Search Strate~ of Quantum—behaved Particle Swarm Optimization[A].Proceedings of IEEE conference on Cybemeties and Intelligent Systems [C].2004,pp.111—116

[3] Zhao D P,Zhu W Y.A Difference and Quick Image Compression Algorithm Based On Fractal Mapping[J].Journal of software,2001,12(1):pp.134—142

[4] Jacquin A. E, Image CodingBased on a Fractal Theory of Iterated Con2tractive Image Transformations, IEEE Trans. Image Processing, January1992, (1): pp.18～30

[5] Chen Fanwu.The best subdivision of fuzzy cluster soft sort problem [J]. Mini Micro Systems, 1992, (6):pp.18-24

[6] He Chuanjiang,Li Gaoping.A fractal image encoding improved algorithm [J].Computer emulation.2004,21(8):pp.62-65

[7] Gao Xinbo. Fuzzy cluster analysis and applications [M] . Xi'an, Xi'an electron science and technology university, 2004

[8] Lai C M, Lam K M, Siu W C. A fast fractal image coding based on kick-out and zero contrast conditions [J]. IEEE Transactions on Image Processing, 2003, 12 (11): pp.1398-1403

[9] He C, Yang S X, Xu X. Fast fractal image compression based on one-norm of normalised block[J]. IEE Electronics Letters. 2004, 40 (17): pp.1052-1053

[10] Ghazel, M., Freeman, G. H., and Vrscay E. R. Fractal image denoising [J], IEEE Transactions on Image Processing, 2003, 12, (12): pp.1560-157

[11] Zhou Xinhua, Huang Dao. An average value clusterbased on ant group's fuzzy C [J] . Control engineering, 2005 ,12 (2) :132

Detecting Position of Multi-objects Based on Neural Network

Junyi Hou

Xuzhou Air Force College, Xuzhou, Jiangsu, 221000, China
Email: klf030@163.com

Abstract

In this paper, by the function that BP neural network can simulate eyes, the relation between image points and special points is built .The mass center in image is input into neural network, two BP networks are applied to processing a pair of left-right image and special position of object is rapidly obtained .Satisfactory result is obtained in the experiments.

Keywords: Image processing，BP neural network

1 Introduction

BP neural network is one of the important model of manpower neural network .As a multi-layer Feed-forward neural network, it has been investigated and applied mostly ,even being approved in theory.Three-layer neural network will have the capability of simulate any complex nonlinear mapping[1],so long as has enough hinder numbers. BP learning algorithm(contradict transmit learning algorithm) has a characteristic of clear route、precise structure、stable working states and perfect manipulation. Forward neural network of BP algorithm is a manpower neural network which has been applied most extensively presently, it can resolve questions that most neural network faced. Location and track of target is important in ballistic trajectory missilery recovery、air defense、coast defense 、 section recovery and campaign scrutiny ,it has attracted the attention of scientists and engineers .Now the object detection based on machinery visualization has two methods. One is track the object via vidicon turning, It only deal with limited object and further more image is farther to sensor, the object is always in the available field ,the main aim is to track the object . Another one is locality detection of near distance and multi-object. The vidicon is fixed while the object is moving. Then analyze the movement rule of multi-objects by continuous shooting sequence images. Its characteristic is that in the image sequence numbers and pose of objects is varied, and also need displacement field and velocity field of images. So in this way we could get the movement rules of every object. For example the method of only using image manipulation, multi objects location need periodic sampling, so it will not only bring mass computation and long period, also is far to track the tracked target and the practical requirement. So we have the necessity to seek a location method of multi objects which is rapid、reliable and precise. Because of high paralleled, outstanding and accordable learning capacity and extensive applications in the fields of model recognition, so in this article the author aims at the second method puts forward a location method of multi-objects which is combination of BP network and computer vision and dynamic image manipulation.

2 BP Network Model and Algorithm

Feed-forward BP network is also contradict transmit neural network, which is most used in Feed-forward network to carry out mapping transformation, it is also the mostly research and the clearest cognize. Three-layer Feed-forward BP network reserve information by BP network, also called error contradict transmit learning method, is a typical method of correcting error. Its basic ideology is the error of deferent layer in network learning and expected results is contradictorily transmitted to input layer and then to each joint, so we can compute referenced error of each joint, after this make some correspond adjust to let the network to adapt required mapping. BP algorithm (Back Propagation algorithm) is a learning way of neural network which has been applied most widely. The processes of BP algorithm:

Input transmission forward by network

$$a^0 = p$$

$$a^{m+1} = f^{m+1}(w^{m+1}a^m + b^{m+1}), m = 0,1,\cdots m-1$$

$$a = a^m$$

Sensitivity transmit reverse by network

$$S^M = -2\dot{F}^M(n^M)(t-a)$$

$$S^m = \dot{F}^m(n^m)(W^{m+1})^T S^{m+1}, m = M-1,\cdots,2,1$$

Weight update (weight and biases by Approximate steepest descent)

$$W^m(k+1) = W^m(k) - \alpha S^m(a^{m-1})^T$$

$$b^m(k+1) = b^m(k) - \alpha S^m$$

BP network algorithm deduce clearly will make learning more precise, also make multiplayer Feed-ward network learn anything it wants. The BP network after exercised operates very fast, also can be conducted collaterally. Based on these advantages, We can use two BP networks to conduct a pair of left-right images combine in the multi-objects detection to obtain mass center position.

3 Image Processing

3.1 Remove background of image

The background in the actual applied detection is more complex, while the complex backgrounds have influence on the precision of detection, even hard to conduct. So the first task in this object detection is to remove the background by difference images. The vidicon is static when the image is obtained, but the number and pose of the sequence image is varied, the background of the image is relatively static. The scenery of the valid field in the vidicon will not be influenced by nature light only when the light was put in arc position. So difference images is adopted to remove background ,that is combined light forward image and fixed background to remove the background[3].suppose background image is $f(i,j)$,image of t_k time is $f(i,j,t_k)$, sequence image is $g(i,j,t_k)$

So: $g(i,j,t_k) = f(i,j,t_k) - f(i,j)$ (1)

3.2 Remove blur and noise

The image after processed nearly have no background, but movement blur and noise is still, so first we use image recovery and Gauss filter way to remove blur and noise, then according to the characteristic of erection chart, adopt improved DA[4] to mark and inspect objects.

3.3 Seek mass center object

The shape and pose of the object is always changing during movement, so we use space perspective invariable matrix to seek the mass center as a characteristic quantity of object detection in order to locate object.

A figure image $f(i,j)$ of N*N dimension

Its definition

$$f(i,j) = \begin{cases} 1......(i,j) \in object \\ 0.....(i,j) \in object \end{cases} \quad (2)$$

Center distance

$$\begin{cases} x_c = m(1,0)/m(0,0) \\ y_c = m(0,1)/m(0,0) \end{cases} \quad (3)$$

(x_c, y) is horizontal and vertical equality of mass center

There into $m(n,v) = \prod_{i=1}^{N}\prod_{j=1}^{N} f(i,j)i^u \cdot j^v$

Suppose a coordinate of any point in the object is (x_i, y_j), d_i is distance square of any point to mass center (x_c, y_c) , $d_i = (x_i - x_c)^2 + (y_i - y_c)^2$ (4)

From formula (4) we know that the distance of this object mass center and all the mass center below the object is smaller than any ratio, so must seek the object of the objects, choose any point in image $f(i,j)$ as the origination drop, the search process will begin from here to the mass center position of the objects, suppose (k,l) as dynamic mass center coordinate,

$$h(k,l)=\prod_{i=1}^{N}\prod_{j=1}^{N}d(i-k,j-l)f(i,j) \qquad (5)$$

Its value shows the relativity of image and model in the dynamic mass center (k,l), actually it is sum of distance from any point in image to dynamic mass center (k,l)

4 BP Network in Detecting Position of Multi-Object

4.1 Make sure the mapping relation of image points and special points

In the model recognition, neural network technology is a very effective process of sort recognition technology, substantively it is a sort of mapping. This article make use of computer vision theory, in virtue of dynamic Image processing technology, and detecting near position of multi-objects ,then make sure of relation of image points and special points. Suppose the space process from image points to special points as black box, then induct three-layer BP network, translate the problem of import and output in a set of stylebook into a high nonlinear mapping problem, regulate linked weight of neural network via iterative operation, and add connotative node of optimize parameter to approach nonlinear functions, in order to ascertain mapping relation .After many experimentation, we come to reckon input layer include two neural center is mass image coordinate (x_c, y_c), the output layer include three neural center is mass image coordinate (x_p, y_p, z_p), the midst layer include three layer BP network of 6 neural center(2-6-3).

Appropriate sample point has determinative function on detecting precision when we choose BP network to base the mapping relation[5].The sample should not only treat with fathomable extension but also reflect detectable extension, In this article we choose a model of 120×200 mm^2,the background is black, and it also includes symmetrically 5×12 white points whose diameter is 6mm, the distance between moving twice and thrice along vertical model is R, even white points in model of shooting thrice as exercised sample are put into network, then exercise network weight, educe the mapping relation of image points and special points.

4.2 Algorithm of collateral bp network in detecting multi-objects

Owing to overlap phenomenon will appear in detecting moving multi objects from a fixed angle, so we choose two vidicon to shoot synchronously, then after the images are preprocessed, we seek and compute mass center ,and input the mass center coordinate of a pair of left-right images into two 2-6-3 BP network, detect them collaterally, suppose the object mass center coordinate of left image is (x_l, y_l),compute coordinate is (x_1', y_1', z_1'), the object mass center coordinate of right image is(x_2,y_2),compute coordinate is (x_2', y_2', z_2'), if the left images is bigger than the object numbers in right images, so:

$$x_1' = x', y_1' = y', z_1' = z';$$

In opposition

$$x_2' = x', y_2' = y', z_2' = z';$$

If the object numbers in left image and right image is equal, so:

$$x' = \frac{(x_1' + x_2')}{2}, y' = \frac{(y_1' + y_2')}{2}, z' = \frac{(z_1' + z_2')}{2}$$

5 Application and Analysis

Put this method in detecting multi-objects of near distance, suppose as a row of peanut seed in free fall, the distance about 200 mm, process as follows:

1) Make sure of the mapping relation of image points and special points to get network weight matrix;

2) Shoot continuously 30 frames of peanut seed of sequenced fall;

3) Remove the background blur and noise, identify the object, compute mass center;

4) Choose the computed images mass as sample, then input network, educe the space position of

objects;

5) According to the collection velocity in 25 frame/second of Metro Ⅱ image collection and the formula s=1/2g t^2 to reckon the relative position between the upper frame and lower frame, compare with every object position P_i(I=1,2…n)in the lower frame, n is the numbers of object in the lower frame, to find the same objects in the two frames then joint them;

6) After being jointed, we can get the moving rule of the object.

Table 1,2,3 are detection results of three sequenced frames , process time of each frame about 2 seconds, max error of computed position less than 1mm.

Table 1 Compare the actual coordinate of the first frame with exercised results

No.	actual coordinate(mm)	Exercised coordinate(mm)	Absolute value of error(mm)
1	(36.5,110.2,3.5)	(36.2,109.9,3.1)	(0.3,0.3,0.4)
2	(98.3,196.1,4.0)	(96.5,195.9,3.7)	(0.2,0.2,0.3)
3	(61.0,306.7,4.1)	(60.6,307.0,3.7)	(0.4,0.3,0.4)

Table 2 Compare the actual coordinate of the second frame with exercised results:

No.	Actual coordinate (mm)	Exercised coordinate(mm)	Absolute value of error(mm)
1	(24.6,85.1,3.1)	(26.1,85.0,3.4)	(0.5,0.1,0.3)
2	(39.3,165.0,3.2)	(39.1,164.8,2.)	(0.2,0.2,0.4)
3	(48.0,274.2,2.8)	(48.1,274.6,3.)	(0.1,0.4,0.3)

6 Discussion

In this article the author aims at the characteristic of detecting position of multi objects, introduce the neural network to ascertain the mapping relation of image points and special points, avoid of the complex nonlinear algorithm locked in vidicon, and seek the object position quickly by image processing and BP collateral algorithm, then after jointed the frames to get the rule of object movement. This algorithm is rude, high precision in location and easy to compute and design.

Table 3 Compare the actual coordinate of the third frame with exercised results

No.	actual coordinate (mm)	Exercised coordinate(mm)	absolute value of error(mm)
1	(25.，129.6，3.2)	(25.5,129.3，3.2)	(0.4，0.3，0.0)
2	(35.7，227.4，3.1)	(35.9，226.7，2.8)	(0.2，0.3，0.3)
3	(46.0，350.0，3.5)	(46.2，349.1，3.1)	(0.2，0.1，0.4)

References

[1] Engozingers, Tomsene, "An accelerated learning algorithm for multiplayer perceptions Optimization layer by layer", IEEE Trans on Neural Networks, Vol.6, No.1, 1995, pp.31~42

[2] Zhou Hongyi, Jing Zhongliang etc, Mobile Object Track, Beijing: National defence industrial publishing company, 1991

[3] Martin T, Hagan, Howard B, Demuth, Mark H, Beale Dai kui etc translated，Machine industrial publishing company，2002

[4] Hu Shaoxing, Ma Chenglin, Zhang Aiwu, High Technology Communication，2002

[5] Bar-Shalom Y, Fortman TE, Tracking and Data Association, New work: Acdemic press, Inc，1988

Infrared Target Image Compression based on Region of Interest

Ying Wang[1,2] Shengzhi Yuan[2] Ao Sun[1, 2] Yufeng Wei[1, 2]

1 Unit 91550 of PLA, Dalian, Liaoning, China
Email: alpsblue@163.com

2 Department of Science and Technology of Weapons, Naval Aeronautical Engineering Institute
Yantai, Shandong, China
Email: yuanshengzhi_hy@sina.com

Abstract

In order to meet the needs of automatic target recognition (ATR), Infrared target image has been used more frequently than ever. But the transmission question for Infrared target image through Data Link is hard to solve. For the good real-time performance, Image compression based on region of interest (ROI) has been one of the hot issues in the field of image compression and coding. However, there is not a fixed model for region of interest automatically detected. A new stepwise approaching and recurring threshold search algorithm based on 2D(two-dimension) maximum entropy principle was proposed for ROI automatically detected while a compressed scheme based on ROI was studied. An experimental study was also conducted after the compressed scheme realized in the frame of JPEG2000. It was proved that the method of ROI automatic detected not only can meet real-time requirements, but also is reliable, effective and significant in applications.

Keywords: Infrared Target Image, Image Compression, Region of Interest (ROI), JPEG2000

1 Introduction

Along with the development of UAV and LAM (Loitering Attack Missile), the image of the battlefield (such as Infrared target image and SAR image) should be compressed and transmitted through Data Link for automatic target recognition (ATR). Because of the bandwidth limitation in the data link channel, it is desired to achieve high compression rates while the quality of the reconstructed images should been enhanced. But the applicability of the reconstructed images depends on whether some significant characteristics of the original images are preserved after the compression process has be finished, it is necessary to do research on the technology of image compression and coding based on Region of Interest (ROI), which can compress different parts of a image in different compression rates without important information losing and has become one of the hot issues in the image coding domain. At present, the research on the image compression and coding of ROI is mostly based on the discrete wavelet transformation (DWT), SPIHT algorithm and EBCOT algorithm. A novel image compression scheme based on DWT, Bitplane shift algorithm and SPIHT was suggested in document [1][2].While most research on image compression of ROI based on EBCOT, is the application of JPEG2000, such as document [3][4].

Obviously the region of interest (ROI) in an infrared target image about the battlefield is the region in which the targets maybe exist. The other region of the image is the background (BG), which we don’t care about. Through ROI and BG coding separately, the information of the targets in the ROI can be preserved while the information of BG can be curtailed, or even neglected. Thus, high compression rates can be achieved, the complexity of code operation also reduced. Because of uncertainty of image content and real time requirement for compression and transmission, a novel image compression scheme based on ROI automatically detected was suggested in this paper. The paper is organized in the following way: In Section 2, a new stepwise approaching and recurring threshold search algorithm based on 2D maximum entropy principle was proposed for ROI automatically detected. In Section 3, the infrared image compression based on JPEG2000 was designed and realized. An experimental study of the image compression scheme based on ROI automatically detected was shown in Section 4. In Section 5 the issues mentioned in this paper were summarized, and further work also described.

2 Region of Interest Automatically Detected

The technology of ROI automatically detected belongs to image segmentation, which is a very difficult problem in practice. In general, there are three kinds of approaches for image segmentation: the method based on image threshold, the method based on image edge and the method on image boundary. In the complex environment, the infrared target image may be contaminated by a variety of noise sources because of the sensor effect of the infrared equipment which is inherent and atmospheric radiation. Thus it is very difficult to analyze the boundary and edge of the infrared image. The grey range of the infrared target and background is different while the temperature of them is significantly different. In comparison with the method based on image edge and image boundary, it is very common and effective to adopt the method based on image threshold for image segmentation. However, the major algorithms existed on image threshold cannot fulfill the requirement of practice because of bad adaptability and real-time performance. It is hard to achieve satisfactory effect for different targets with different sizes in the infrared target image by the algorithms. If the area of target in an image is more than 30%, the satisfactory effect can be achieved by the traditional methods on image threshold, such as Prewitt method and OTSU method [5]. Along with the relative area of target reducing, the performance of the traditional methods decreased rapidly. In comparison with 1D maximum entropy approach, 2D maximum entropy, which was purposed by Kapu, Abutelab in 1989, not only has greater adaptability for ifferent targets with different sizes in the infrared image, but also greater resistance capability to noise, greater robustness. But the operation capacity of 2D maximum entropy approach increases according to the index growth. It is hard to satisfy the real time requirement [6][7]. Thus it is necessary to suggest a new stepwise approaching and recurring threshold search algorithm based on 2D maximum entropy principle for good real time performance.

2.1 A New Stepwise Approaching and Recurring Threshold Search Algorithm

Based on 2D maximum entropy approach, 2D histogram should be first formed by gray value of each pixel and its neighboring region in the image. A typical 2D gradation histogram is shown as Figure 1. Coordinate y is gray average of region while Coordinate x is gray level of the pixel. In the histogram, r_{ij} is the number of pixel in the image, whose gray level is i and average gray of region is j. If the threshold vector is (s,t), the histogram is divided into four areas. In an infrared target image, the gray and the average gray of the target or the background have little difference due to the slowly change about the gray value of pixel in the target or the background region. So the diagonal areas in two-dimension histogram are chiefly occupied by the target and the background region. The A-region represents the background, and B-region is the target, and C-region and D-region are chiefly noise pixels and edge pixels.

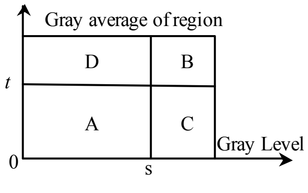


Figure1 Two-dimentional aray histogram divided by vector(s,t)

The definition of the global two-dimension entropy about an infrared image was shown as Eq.(2-1).

$$H(s,t)=\ln\left[P_A(1-P_A)\right]+\frac{H_A}{P_A}+\frac{H_L-H_A}{1-P_A} \qquad (2\text{-}1)$$

In the Eq.2-1,

$$H(A)=-\sum_{i=1}^{s}\sum_{j=1}^{t}(\frac{p_{ij}}{P_A})\ln(\frac{p_{ij}}{P_A}),$$

$$H_L=-\sum_{i=1}^{L}\sum_{j=1}^{L}p_{ij}\ln p_{ij},\ P_A=\sum_{i=1}^{S}\sum_{j=1}^{T}p_{ij}.$$

Through seeking the maximum value of $H(s,t)$, the optimal vector (s,t) is found. The recurring formulas are shown as Eq.(2-2), Eq. (2-3), Eq. (2-4), Eq. (2-5).

$$P_A(s+1,t)=P_A(s,t)+\sum_{j=1}^{t}p_{s+1,j} \qquad (2\text{-}2)$$

$$\begin{aligned}P_A(s,t+1)&=\sum_{i=1}^{s}\sum_{j=1}^{t+1}p_{ij}\\&=P_A(s,t)+P_A(s-1,t+1)-P_A(s-1,t)+p_{s,t+1}\end{aligned} \qquad (2\text{-}3)$$

$$H_A(s+1,t)=H_A(s,t)-\sum_{j=1}^{t}p_{s+1,j}\lg p_{s+1,j} \qquad (2\text{-}4)$$

$$\begin{aligned}H_A(s,t+1)&=-\sum_{i=1}^{s}\sum_{j=1}^{t+1}p_{i,j}\lg p_{i,j}\\&=H_A(s,t)+H_A(s-1,t+1)\\&\quad-H_A(s-1,t)-p_{s-1,t}\lg p_{s-1,t}\end{aligned} \qquad (2\text{-}5)$$

Through recursive optimization, the complexity of the image segmentation algorithm based on 2D maximum entropy has been decreased from $O(n^4)$ to $O(n^2)$ .If we want to further reduce the operating time, the number of logarithm operation and cycle operation, which is the key to improve operation efficiency of the algorithm, should be reduced. The basic principle of the progressive approach and recurring search algorithm is shown as following:

1) Searching the rough threshold value on the two-dimension histogram at rough scale;

2) Searching the accurate threshold value in the region nearby the rough threshold value.

This method avoids the unnecessary logarithm operation and cycle operation in the region of maximum entropy impossible existed, and raises the operating efficiency. After the elements at some scale in the original histogram unifying, 2D histogram at rough scale is formed.

Suppose $f(x,y)$ is a 2D gray image, the size is $M\times N$, the whole gray level is L; $G(s,t)$ is the two-dimension gray histogram. The domain of $G(s,t)$ is ：$D=\{(s,t)\big|1\le s\le L,1\le t\le L\}$ .The element of $G(s,t)$, $g_{st}=r_{ij}/(M\times N)$. Suppose 2D histogram at rough scale is $G^{'}(s^{'},t^{'})$; the formula about the element of $G^{'}(s^{'},t^{'})$ was shown as Eq.(2-6).

$$g'_{s',t'}=\sum_{s=s'*2^m-(2^m-1)}^{s'*2^m}\ \sum_{t=t'*2^m-(2^m-1)}^{t'*2^m}g_{s,t} \qquad (2\text{-}6)$$

While the domain of $G^{'}(s^{'},t^{'})$ is $D^{'}=\{(s^{'},t^{'})\big|1\le s^{'}\le L/2,1\le t^{'}\le L/2\}$ .The size of $G(s,t)$ is $L\times L$, the size of $G^{'}(s^{'},t^{'})$ is $L/2\times L/2$, while 2^m is the gray level span for the histogram at rough scale and the original histogram. As a matter of convenience, it can be discussed in the series field, $s,t,s^{'}$ and $t^{'}$ are continuous variables.

Because $\iint\limits_D G\left(s,t\right)dsdt=\iint\limits_D G^{'}\left(s^{'},t^{'}\right)ds^{'}dt^{'}$, while s= $s^{'}*2^m$, $t=t^{'}*2^m$ we can get $G'(s',t')=2^m*2^m*$ $G(2^m s',2^m t')$. Thus it is clear that $G^{'}$ is the scale transformation of G, the histograms of $G^{'}$ and G are similar. In theory of the histogram, only if the gray of some special point is used as the threshold for image segmentation, two-dimension entropy of the image is maximum.The larger deviation of the special point, the smaller two-dimension entropy of image. Therefore from the histogram of $G^{'}$, rough threshold about the image we can get.

The realization process of the progressive approach and recurring search algorithm proposed above is as follows: Suppose $G^{'}$ and its element g_{st} known, three Matrixes $PS^{'}$, $HS^{'}$ and $H^{'}$ are defined. The formulas are shown as Eq.(2-7), (2-8), (2-9).

$$PS'(s,t)=\sum_{i=1}^{s}\sum_{j=1}^{t}g'_{i,j} \qquad (2\text{-}7)$$

$$HS'(s,t)=-\sum_{i=1}^{s}\sum_{j=1}^{t}g'_{i,j}\log g'_{i,j} \qquad (2\text{-}8)$$

$$HS'(s,t)=-\sum_{i=1}^{s}\sum_{j=1}^{t}g'_{i,j}\log g'_{i,j} \qquad (2\text{-}9)$$

Obviously, the recursive operations based on the recurring formulas Eq.(2-2), (2-3), (2-4), (2-5) of $G^{'}$ are changed to the recursive operations of Matrices $PS^{'}$, $HS^{'}$ and $H^{'}$. In the Matrix $H^{'}$, the largest element of

$H^{'}$ is the rough threshold. Equally, Matrices PS , HS and H of G can be calculated. The domain D is defined as below.

$$D=\{(s,t)\,|\,\frac{L}{2^m}S-\frac{L}{2^m}\le s\le\frac{L}{2^m}S+\frac{L}{2^m},\frac{L}{2^m}T-\frac{L}{2^m}\le t\le\frac{L}{2^m}T+\frac{L}{2^m}\}.$$

In the progressive approach algorithm, two-step searching or multilevel searching method can be used. In comparison with the two-step searching method, the number of logarithm operation and cycle operation based on the multilevel searching method is much less in theory. But through repeated experiment on different images, it is found that the adaptability of the two-step searching method is much better while the execution time and efficiency of both methods are no great different. Therefore, the two-step searching method is chosen in practice. The algorithmic flow based on the two-step searching method is shown as Figure 2.

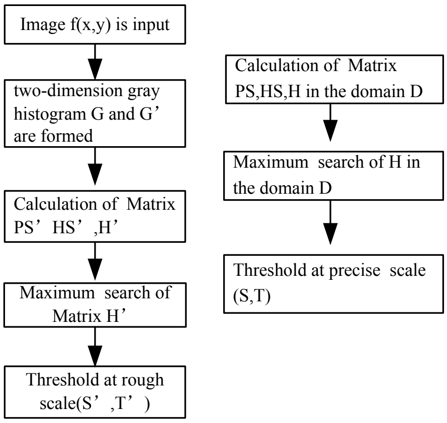


Image f(x,y) is input
two-dimension gray histogram G and G’ are formed
Calculation of Matrix PS’ HS’ ,H’
Maximum search of Matrix H’
Threshold at rough scale(S’ ,T’)
Calculation of Matrix PS,HS,H in the domain D
Maximum search of H in the domain D
Threshold at precise scale (S,T)

Figure 2 Algorithmic flow based on the two-step searching method

2.2 Process of ROI automatically detected

The region, in which the targets maybe exist, is obviously the region of interest (ROI) of the infrared image. In comparison with the background (BG), the connectivity and aggregation in the Region of Interest is much better. After image segmentation through 2D maximum entropy approach, a binary image including targets and noise is formed. In order to realize Region of Interest automatically detected, the connectivity of target region should be recovered and the visual effect of the binary image enhanced while false alarm points eliminated. In accordance with the characteristic of the binary image, there are three steps for region of interest automatically detected: 1) order statistic filtering; 2) mathematical morphological filtering; 3) rectangular extension about the region of interest (ROI).

Before the image segmentation, median filter, which is the special case of order statistic filter, can be used for pretreatment about the infrared image .the effect of noise is, to a certain extent, decreased. Because of different aggregation about the pixels of the target and noise in the binary image which is formed by image segmentation, order statistic filter can be used to remove discrete points about noise partly. Based on the order statistic filter, the connectivity of target region has been enhanced without information loss about targets. After order statistic filtering, the region about targets and polymerization noise is included in the binary image. The cascade operation about open and close in mathematical morphological filtering is used to eliminate the polymerization noise.

Owing to Region of Interest automatically detected for ATR (automatic target recognition), it is unsuitable to consider the target shape directly as Region of Interest. The serious disturbance about the recognition effect of targets will be found if the segmentation errors occur. In practice, the target is defined as a connective region whose pixel number is more than N or equal to N, region of interest (ROI) as an extended circumscribed rectangular in which the targets contained (shown as Figure 3).

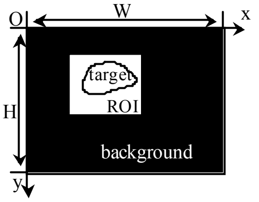


Figure 3 Rectangular extension about region of interest

3 Infrared Target Image Compression Based on JPEG2000

As the new standard of image compression, JPEG2000 not only provides good compressed performance at the high bit rate or the low bit rate, but also support the new characteristics that JPEG does not support. ROI coding is one of them. A user-defined ROI of the image is allowed in JPEG2000.With applying JPEG2000 in the field of infrared target image compression based on ROI, whether the good effect can be achieved is a question worth studying.

3.1 Selection of wavelet bases

There is not any final conclusion of the research on wavelet base selection. In the most chances, wavelet bases should be chosen by experience according to the special application. The different effects of image compression can be achieved while the different wavelet base chosen. There are some related studies on the evaluation of different effects [8-10]. Wavelet bases commonly used are: Haar, Daubechies, Biorthogonal, Coiflets, and Symlets. Whether the wavelet bases above are suitable to the infrared target image compression should be discussed.

The peak signal to noise ratio (PSNR) and energy compactness [11] are chosen for Evaluation standards of wavelet base. The PSNR, which is the common evaluation index for quality of image compression, is the reflection of the quality of the reconstruction image. Energy compactness, which has a very important effect on the compression efficiency, is the reflection of specific property of energy concentration about DWT.

Wavelet bases commonly used are evaluated by Lena image and other 4 infrared images. After 3-level wavelet decomposition, the same quantization operation is executed without entropy coding. The coefficients after quantization are used for image reconstruction. Based on the data of PSNR, it is found that the compressed effect of Biorthogonal wavelet bases is the best; especially bior3.5 and bior3.7.Based on the analysis about the energy compactness curves, it is found that bior2.2 wavelet is best while the compactness of bior3.7 is worst. Obviously, In comparison with bior3.5, the wavelet filter length of bior3.7 is longer and the operation complexity is higher, the operating time is longer. Thus the bior3.5 is more suitable than Daubechies (9, 7) or Daubechies (5, 3).

3.2 Coding method of ROI

Recently, in the frame of EBCOT, there are two ways to realize the coding of ROI: the Bitplane shift algorithm and the Rate Distortion Slope algorithm. Many Bitplane shift algorithms such as General Scaling based algorithm; Max Shift algorithm and BbB shift algorithm、PSB shift algorithm，HB shift algorithm were discussed in many papers. The Bitplane shift algorithms were based on the wavelet coefficients of BG scale-down. The Rate Distortion Slope algorithm was based on the optimization of Rate Distortion Slope in the frame of EBCOT. The wavelet coefficients of ROI were encoded preferentially through Rate Distortion Slope of the ROI code area shifting in the code organization process.

At excessively low bit-rate without account of information of BG, the algorithm complexity of Max Shift algorithm is least in all Bitplane shift algorithms. The execution efficiency is also best in theory because the quality of the reconstructed image about ROI and BG is no use for control. But the enlarge phenomena will appear if the Rate Distortion Slope algorithm used, especially the ROI is very little. Thus the Max Shift algorithm was chosen for the coding method of ROI in practice in view of uncertainty of ROI in the Infrared target image.

3.3 Compressed scheme based on region of interest automatically detected

In the frame of JPEG2000, a compressed scheme based on ROI automatically detected for the infrared target image is proposed. In the compressed scheme, the bior3.5 as the basic wavelet basis of JPEG2000.The Max Shift algorithm was chosen for ROI coding at low bit-rate. During the process of realization (in Figure 4), there are five steps: 1) the wavelet coefficients of an infrared target image is formed through DWT after preprocessing; 2) the wavelet coefficients are quantified;

3) Based on the fast recurring algorithm based on 2D entropy threshold, ROI of the infrared image was detected; 4) the wavelet coefficients are shifted by the Max Shift Algorithm after the mask of ROI forming; 5) after Tier-1and Tier-2 coding, the combined bit flow is formed for transmission.

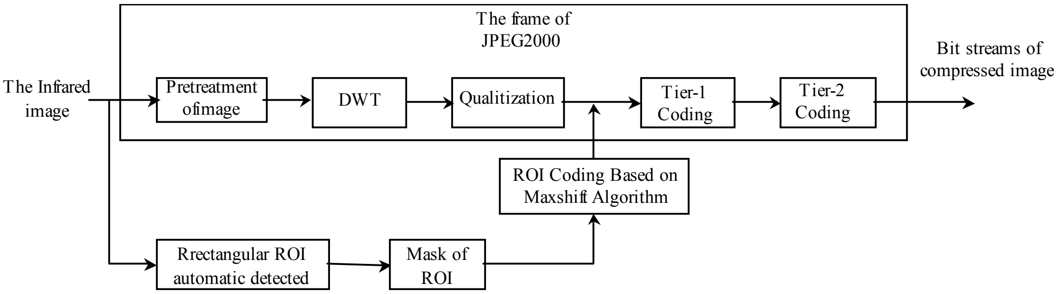
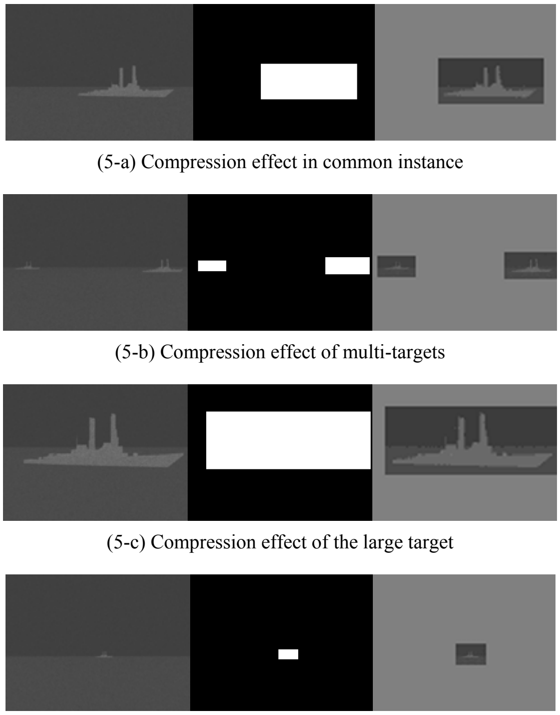


Figure 4 Compressed scheme based on ROI automatic detected

4 An Experimental Study for Compressed Scheme Based on ROI

In the computer of Pentium IV 2.4GHz CPU，1GB RAM, 4 infrared target images whose size are all 768×576 were chosen for experimental study on compressed scheme based on ROI automatically detected. The compressed rate chosen in this experiment is the minimum compressed rate in which reconstructed images can be detected easily. The compression effect was shown in Figure 5.



(5-a) Compression effect in common instance

(5-b) Compression effect of multi-targets

(5-c) Compression effect of the large target

(5-d) Compression effect of the small target

Figure 5 Compression effect of infrared images in various conditions

The experimental data we had gotten are listed in Tab.1.

Table 1 Experimental Data of Image Compression

Parameters	Image 1#	Image 2#	Image 3#	Image 4#
Compressed Rate (bpp)	0.007	0.008	0.008	0.006
File Size after Compression (b)	349	436	410	320
PSNR of ROI(db)	31.567	34.178	32.030	34.242

Suppose the bandwidth of Data Link is 9.6kbps，time of image compression, time of transmission and time of image reconstruction in theory are listed in Tab.2.

Table 2 Time Parameters of infrared image compression

Time Parameters	Image 1#	Image 2#	Image 3#	Image 4#
Time of image compression(s)	0. 2657	0. 2555	0. 2590	0. 2544
Time of transmission(s)	0. 2908	0. 3633	0. 3417	0. 2667
time of image reconstruction(s)	0. 108 3	0. 106 2	0. 092	0. 1065

From the experimental data above, some conclusion we have get is shown:

1) The compressed scheme based on region of interest automatically detected, which is proposed above, can be used for high strength compression. File size after compression is about several hundred bits. Thus the compressed scheme in this paper can satisfy the low bandwidth requirements of Data Link.

2) Based on the result of experiment, it is found that the good recognition effect can be achieved if the compressed rate is kept above 0.006bpp. At this position, the file size after compression is above 320 bits. In general, if the compressed rate is 0.008bpp, the file size after compression is about 420 bits. If the bandwidth of Data link is 9600 bps, 2.86 frames (9600/420=2.86) will be transmitted per second in theory. The time of image compression should be less than 0.35s (1/2.86=0.35).Based on the data in the table 6-5, it is found that the time of image compression has met the requirement of time.

5 Summary and Future Work

In order to solve the image transmission based on low bandwidth data link for advanced equipment such as UAV and LAM, the technology of automatic extracting about ROI of the infrared image is studied in this paper while a compressed scheme based on region of interest automatically detected was proposed. After realization in the frame JPEG2000, An experimental study is conducted to qualitatively assess the compressed scheme .From the compressed effect and data of experiment; it is found that the contradiction between the low bandwidth and image information can be solved by the compressed scheme proposed in this paper to a large extent.

Although the good compressed effect has achieved, the real-time character of the compressed scheme should be improved. The application of the compressed scheme in a special project should be further studied.

References

[1] Zhang Ye, “Research on the Compression of Region of Interest in still image”, Mater degree thesis, Suzhou University, Suzhou, pp. 8-15

[2] Li Xiaofei and Ma Dawei, “A New Regions of Interest Image Coding Method Based on SPIHT Algorithm”, Application Research of Computers, Editorial office of Application Research of Computers, Chengdu, 2007:2, pp. 189-191

[3] Sun Wu and Wang Youzhao, “An Introduction to Region of Interest Coding Techniques in JPEG2000”, Computer Engineering and Applications, Editorial office of Computer Engineering and Applications, Beijing, 2003:24, pp.67-69

[4] Fang Ruijun, “Research on Region of Interest Coding Algorithm Based on JPEG2000”, Mater degree thesis, Northwestern Polytechnical University, Xi’an, pp. 12-18

[5] S.U.LEE and S. Y CHUANG, “A comparative performance study of several global thresholding techniques for segmentation”, Computer Vision Graphics and Image Processing，Academic Press, Inc, New York, 1990: 52(3), pp.171-190

[6] Gong Jian, Li Liyuan and Chen Weinan, “A Fast Threshold Segment Algorithm based Two-dimensional Entropy”, Journal of Southeast University, Editorial office of Journal of Southeast University, Nanjing, 1996: 25(4),pp.31-36

[7] W.T.Chen and C.H.Wen, “A Fast Two-dimensional Entropic Thresholding Algorithm”, Pattern Recognition, Elservier.Academic. Press, New York 1994: 27(7), pp. 885-893

[8] Ke Li and Huang Lianqin, “Choice of Wavelet Base in Real-time Compression for Remote Sensing Image”, Optical Technique, Editorial office of Optical Technique, Beijing, 2005: 31(1), pp. 77-80

[9] Zhang Ye and Wang Yimin, “Wavelets selection and evaluation for image compression”, Journal of Suzhou University Natural Science, Editorial office of Journal of Suzhou University, Suzhou, 2003: 19(1),pp. 54-58

[10] Yu Xiaohong and Yao Ming, “Wavelet Transform and the Choice of Wavelet Base in Image Compression”, Journal of Computer Applications, Editorial office of Journal of Computer Applications, Chengdu, 2001: 21(7), pp.: 20-22

[11] Yao min, Digital Image Processing, China Machine Press, Beijing, 2007, pp.210-214

Block Segmentation in Cross Stitch Application

Noraziah Ahmad   Suryanti Awang   Nurshabah Yalah   Norazaliza Mohd. Jamil

University Malaysia Pahang, Faculty of Computer Systems & Software Engineering,
Locked Bag 12, 25000 Kuantan, Pahang, Malaysia
Email: noraziah@ump.edu.my, suryanti@ump.edu.my

Abstract

Image segmentation is process of partitioning a digital image into multiple sets of pixels called regions. In this paper, the technique of image segmentation used focused on color image since the output of pattern must be in color. We present the implementation of block segmentation as a technique to convert a color image into cross stitch pattern. Block segmentation is used by considering the characteristic of cross stitch pattern itself.

Keywords: Block segmentation, color image, RGB, cross stitch, pixel

1 Introduction

Cross-stitch is entertaining needlework where a pattern sewed using needle and threads in canvases fabric to form a picture. A cross stitch pattern is simply a rectangular grid where some square in the grid are filled with colors. Figure 1 shows the pattern and its actual stitching.

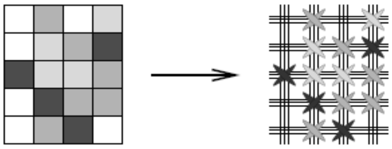


Figure 1   The pattern and its actual stitching

Most image processing techniques involving the image treated as two-dimensional signal and applying standard signal processing technique to it. In color image processing, the pixels of a color image can be thought of as vector quantities. A typical encoding of a color image consists or the red, green and blue (RGB) color components, therefore each pixel can be identified as a dimensional vector. The use of color in image processing is motivated by two principal factors; (i) Color is a powerful descriptor that often simplifies object identification and extraction from a scene. (ii) Humans can discern thousands of color shades and intensities, compared to about only two dozen shades of gray.

In this paper, the major of problems focused is color segmentation which is used to obtain on color values. The problems may occur in segmenting an image such as the difficulty to segment a color image which is a three dimension matrix that contains three layers that is red, green and blue. Besides, performing an automatic segmentation method does not perform well when used to process an input containing noise. Other problems may occur is in the color classification. Color classification is used to characterize the color tone from each color within the image. For some images, color classification is hard to recognize because of the color tone. This color tone is in the same class through our eyes but actually different in real tone.

2 Literature Review

RGB Image

RGB image represents an image with three matrices of size matching image format [1]. In other word, RGB color image is an 3 dimensional array of color pixels, where each pixel is corresponds to one of the colors red, green or blue and gives an instruction of how much of each of these colors a certain pixel should use. Figure 2 shows three layers of a color images that is stores in *M x N x3* arrays. An RGB image actually formed from this three color components and stacked in third dimensions [1].

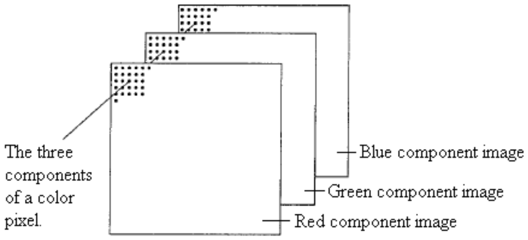


Figure 2 The three components of the RGB image

Block Segmentation

The idea of applying block segmentation in cross stitch application is derived from the mosaic method. This is one of the manual methods used to turn an image into cross stitch pattern [2]. This method is basically performed using Adobe Photoshop application. A desired image is edited manually by pixelate the image into mosaic in Adobe Photoshop to perform the. The original image is divided into cell (the small square part) that can be turn into X-stitch in cross stitch. One cell equals to one stitch. This can be done by using block based segmentation.

Block segmentation is the process of segmenting image into square small size sub-image called block and process each block to obtain the output. The image will be partitioned into blocks with the desired size, and the color value of each block will be obtained. This process is shown in Figure 3 shows block segmentation process.

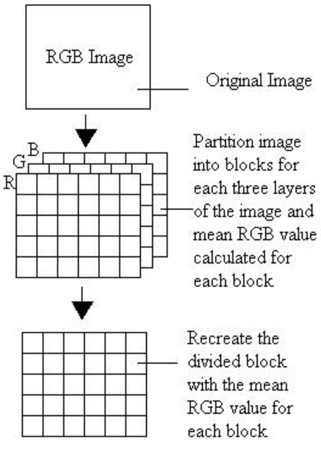

RGB Image
Original Image
B
G
R
Partition image into blocks for each three layers of the image and mean RGB value calculated for each block
Recreate the divided block with the mean RGB value for each block

Figure 3 Block segmentation process

To segment an image, the system partitions the image into small blocks. The mean pixels value is then extracted for each block. This is based on the block processing technique. Block processing is often used for motion analysis. However, it can be implemented in this project by considering the process of dividing the image into small block and process each block to gain the output. In block processing, an image is subdivided into square image blocks [3]. The formula used to get the mean color value is

$$B_k = \frac{\sum_{i=1}^{N} RGB_k(i)}{N} \tag{1}$$

k = each block in image

B = the average value of block *k*

RGB = color value for each pixel in the block *k*

N = number of pixel in a block.

For example, RGB image having a red channel mean value of 87, a green channel mean value of 110 and a blue channel mean value of 77 are seen as a dark green [3]. For color image, this formula can be adopted by process image based on each color R, G and B channels separately and calculates the mean RGB value for each block. As mentioned earlier, RGB images contained three color channels which is when combined, it will produce a color image [1].

3 Methodology

In this paper, the color image size used is a square size and a small size image not more than 300 x 300 dimensions. It’s easier to divide a square image into blocks rather than rectangle since the output block to produce is square blocks.

The process of defining the block size is based on the original image’s dimension size. For image size below than 200 x 200 the block size defined is 2 and for the image size more than 200 x 200 the block size defined is 3. This is to assure that the color tone of the output image still can be characterized and the pattern produced is clear.

After defining the block size, the original image will be divided into the blocks and the number of block for X axis and Y axis of the color image can be acquired.

Let say for 100 x 100 dimensions size of image and the block size is 2 (block of 2 by 2), then the number of block for each axis is 50. The mean RGB values of each block will be calculated by separating three components of red, green and blue layers. This process will be done by looping the calculating process for each layer and each block in each layer.

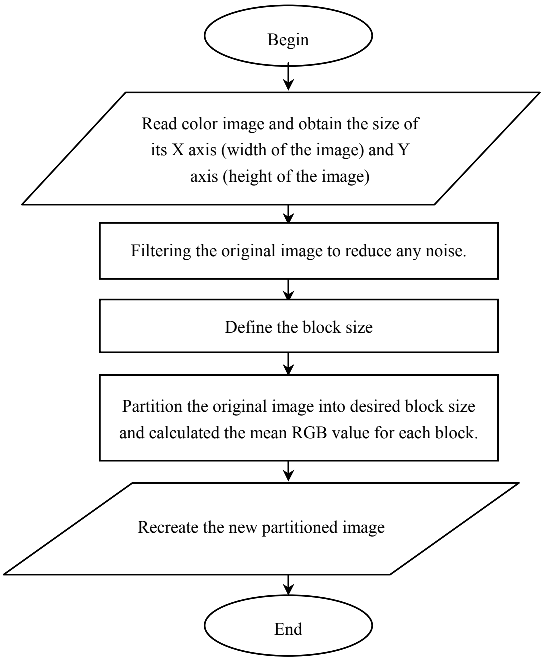
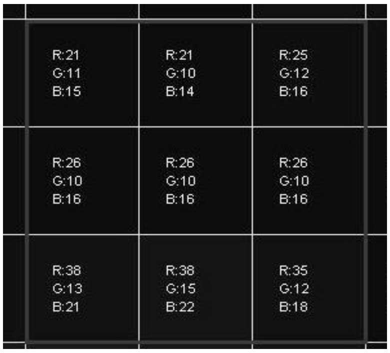

Begin
Read color image and obtain the size of its X axis (width of the image) and Y axis (height of the image)
Filtering the original image to reduce any noise.
Define the block size
Partition the original image into desired block size and calculated the mean RGB value for each block.
Recreate the new partitioned image
End

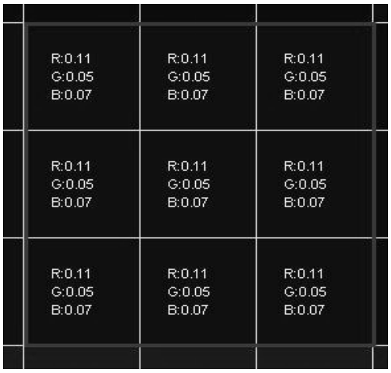
Figure 4 Block Segmentation Flow Chart

After the mean RGB values of each block calculated, the new image with the new RGB values will be recreate to perform a pattern as in Figure 5. For the better view, refer to Figure 5. In the figure we can the larger view block which is 9. Let see the block which has been highlighted in red line. The RGB value of each pixel is displayed. We can conclude that the color of each pixel is same that is brown.



R:21 G:11 B:15	R:21 G:10 B:14	R:25 G:12 B:16
R:26 G:10 B:16	R:26 G:10 B:16	R:26 G:10 B:16
R:38 G:13 B:21	R:38 G:15 B:22	R:35 G:12 B:18

Figure 5 Larger Views of RGB Values in Each Pixel of Original Image



R:0.11 G:0.05 B:0.07	R:0.11 G:0.05 B:0.07	R:0.11 G:0.05 B:0.07
R:0.11 G:0.05 B:0.07	R:0.11 G:0.05 B:0.07	R:0.11 G:0.05 B:0.07
R:0.11 G:0.05 B:0.07	R:0.11 G:0.05 B:0.07	R:0.11 G:0.05 B:0.07

Figure 6 Larger Views of RGB Values in Each Pixel of Segmented Image

However, the RGB value of each pixel is different. In order to make all the pixels in each block are same color, the mean RGB values are calculated. Then the image recreated with the averaged RGB values, which means all the 9 pixels share the same RGB values as shown in Figure 6.

4 Result

Figure 8 shows the 175 x 175 dimension size of image that has been divided block of 3 by 3. It means the total of pixel by each block is 9. An original image is shown in Figure7.

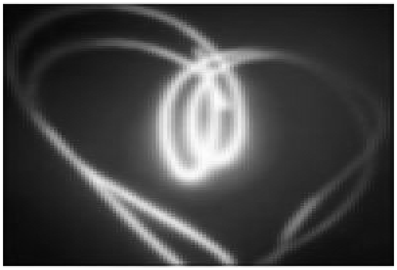


Figure 7 Orignal Image (Input Image)

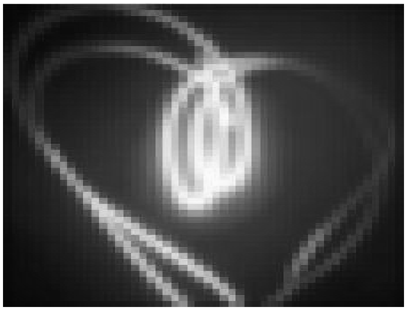


Figure 8 Segmented Image (Result)

From the Figure 8, we can see that the segmented image result can be use as cross stitch pattern, for the image has been divided into small sub-square and on square can be consider as one stitch. The same image process by using block segmentation but the block size defined is 5 as shown in Figure 9. The image produced was not clear.

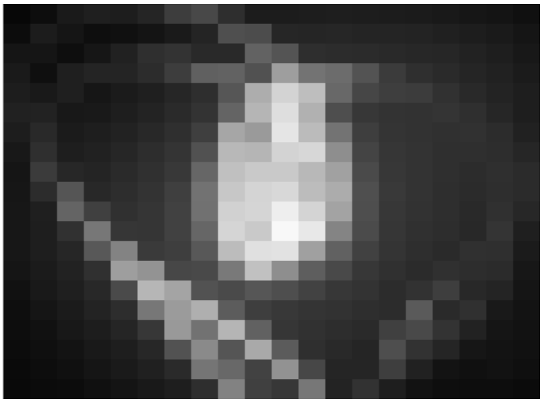


Figure 9 Segmented Image with Block size = 5

5 Conclusion

This paper presented the usage of block segmentation in cross stitch application. The experimental result shows that block segmentation technique is suitable for generating cross stitch pattern this is because the color tone still can be characterized. It is also easier to make the cross stitch pattern instead of using manual methods that is by editing the image using application such as Adobe Photoshop.

References

[1] Rafael C. Gonzalez, Richard E. Woods and Steve Eddins, “Digital Image Processing Using Matlab”, pp. 196 – 205, 2004

[2] Werner D. Streidt, “Digital Image Processing with Feilter Meister”, pp. 10-13, 2000

[3] “How to Make Your Own Cross Stitch Pattern Using Adobe Photoshop”, http://www.geocities.com/Heartland/Prairie/5588/cs4.html, Accessed on 26 March 2008

A Novel Color Subcarrier Recovery Algorithm Based on Phase Estimation and Frequency Offset Estimation

Qinghong Shen[1] Bo Xiao[1] Tao Han[2]

1 Department of Electronic Science and Engineering, Nanjing University, Nanjing, 210093, China
Email: qhshen@nju.edu.cn

2 Genesis Microchip Canada Co. Ottawa, 61350, Canada
Email: hantao@gmail.com

Abstract

This paper presents implementations of recovering the color subcarrier to demodulate the chroma signal used in digital video decoder system. The signal orthogonal decomposition method is used for obtaining the phase estimation of the color burst. On the basis of this, the frequency offset can be estimated by the gradient of the phase estimations of two lines. The obtained phase and frequency offset of color burst are sent to the NCO (Numeric Controlled Oscillator) for recovering the local color subcarrier for chroma demodulation. The proposed method is compared with the conventional close loop and open loop methods. The numeric simulations and the decoding picture prove that the proposed method is effective and practical.

Keywords: Digital Video Decoder, Color Subcarrier Recovery, Phase Estimation, Frequency Offset Estimation, Chroma Demodulation

1 Introduction

In the digital video decode system, there are two main methods to recover color subcarrier: the close loop PLL method [1, 2, 3, 4] based on automatic control theory; and the open loop phase measurement method [5, 6, 7, 8] based on signal estimation theory. The PLL method can resume the local color subcarrier which has the same frequency and phase with the color burst, but its structure is relatively complicate to realize. When the color burst has a big frequency offset, the PLL need a time of several fields to be locked up. Also, the sampling frequency should be close to the integral multiple of the color subcarrier frequency, otherwise the phase detector will output a wrong result [9]; the open loop method uses correlation algorithms to detect the initial phase of color burst of every line in order to recover the local color subcarrier only with the same initial phase. Because the broadcast TV signal's subcarrier frequency is stable and close to the standard value, the local color subcarrier frequency is set to the standard value for simplifying the design. For broadcast TV signal, the method has a good result; for non standard TV signal, such as VCR signal, whose color subcarrier frequency drifts slightly, this method will affect picture quality, because the high accurate local color subcarrier is important for guaranteeing the precision of hue of color [10]. The proposed scheme in this paper uses the signal orthogonal decomposition method to do high accuracy phase approximation of color burst. Then, the gradient of phase of two lines is used to calculate the frequency offset. This scheme has a simple structure and a wide range of application, and is easy to be realized by either hardware or software. The practical decoded color bar picture proves the methods are effective. The block diagram of our video decoder is depicted in Figure 1.

[*] Supported by Scientific and Technological Project of JiangSu Province under Grant No.BE2005006

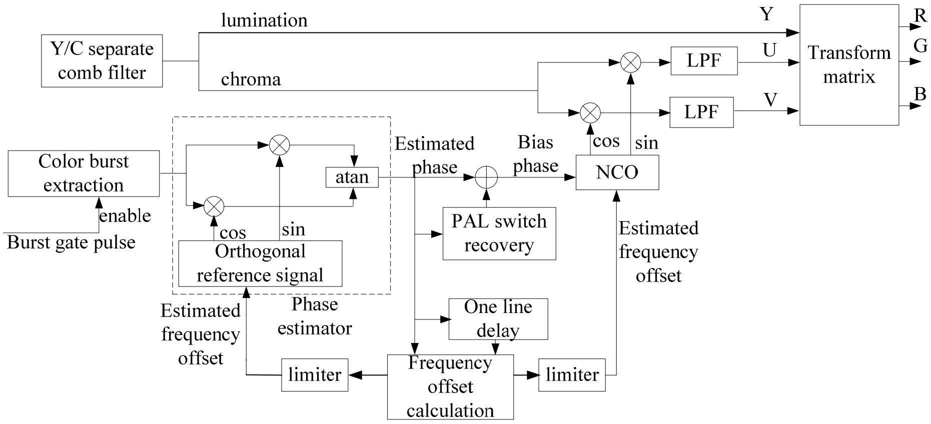


Figure 1　Block diagram of digital video color decoding

2　Phase Estimation

The least-squares method is the most used approach to do phase estimation. But owing to the need for matrix calculation, this approach is not suitable for hardware implementation. In this scheme, the signal orthogonal decomposition approach is used for phase estimation. The color burst signal is a sine wave containing 10~11 cycles, of which expression can be defined as:

$$s(t) = a\sin(2\pi(f+\Delta f)t+\theta_0) \tag{1}$$

Where a is the amplitude, f is the standard frequency (4.43361875MHz), Δf is the frequency offset, θ_0 is the initial phase. Because the color burst has the rising and falling edge, only the intermediate 6~8 cycles are extracted for phase estimation to avoid estimation error caused by waveform distortion. The waveform of the color burst is shown in Figure 2.

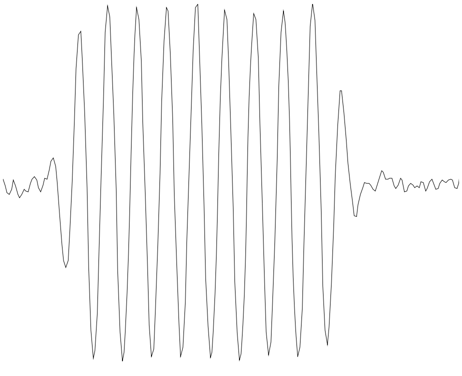


Figure 2　Waveform of the color burst

The orthogonal reference signals are defined as:

$$\begin{aligned} r_1(t) &= \sin(2\pi(f+\hat{\Delta f})t) \\ r_2(t) &= \cos(2\pi(f+\hat{\Delta f})t) \end{aligned} \tag{2}$$

Where $\hat{\Delta f}$ is the estimation of frequency offset. Then, the concept of the decomposition signals can be described by the following equations:

$$\begin{aligned} R_1(t) &= \int_0^{nT} s(t)r_1(t)dt \\ &= \int_0^{nT} \frac{1}{2}(\cos((\Delta f-\hat{\Delta f})t+\theta_0) \\ &\quad -\cos((2f+\Delta f+\hat{\Delta f})t+\theta_0))dt \\ &\approx \int_0^{nT} \frac{1}{2}\cos(\theta_0)dt = \frac{nT}{2}\cos(\theta_0) \end{aligned} \tag{3}$$

$$\begin{aligned} R_2(t) &= \int_0^{nT} s(t)r_2(t)dt \\ &= \int_0^{nT} \frac{1}{2}(\sin((\Delta f-\hat{\Delta f})t+\theta_0) \\ &\quad +\sin((2f+\Delta f+\hat{\Delta f})t+\theta_0))dt \\ &\approx \int_0^{nT} \frac{1}{2}\sin(\theta_0)dt = \frac{nT}{2}\sin(\theta_0) \end{aligned} \tag{4}$$

So, the expression of the estimated phase is:

$$\hat{\theta}_0 = \operatorname{atan}\left(\frac{R_2(t)}{R_1(t)}\right) \tag{5}$$

When this estimation approach is implemented by hardware, only several MAC-structures (Multiply and Accumulate) are needed. Table 1 shows the results of the numerical simulation experiment of phase estimation.

Table 1 The results of phase estimation

Actual θ_0	Estimation $\hat{\theta}_0$
-90^{o}	-89.89^{o}
-80^{o}	-79.12^{o}
-70^{o}	-70.09^{o}
-60^{o}	-61.09^{o}
-55^{o}	-54.92^{o}
-50^{o}	-49.29^{o}
-45^{o}	-45.02^{o}
-40^{o}	-40.28^{o}
-30^{o}	-29.80^{o}
-20^{o}	-20.29^{o}
-10^{o}	-10.19^{o}
-5^{o}	-4.90^{o}
0^{o}	-0.10^{o}
5^{o}	5.70^{o}
10^{o}	9.78^{o}
20^{o}	20.42^{o}
30^{o}	30.19^{o}
40^{o}	40.37^{o}
45^{o}	45.22^{o}
50^{o}	48.65^{o}
55^{o}	54.12^{o}
60^{o}	61.02^{o}
70^{o}	70.32^{o}
80^{o}	81.08^{o}
90^{o}	90.44^{o}
115^{o}	114.86^{o}
140^{o}	142.72^{o}

The recovered phase has three roles: Adjust the oscillation phase of NCO after the bias phase （225^{o} for NTSC line, 135^{o} for PAL line）is added; Estimate the frequency offset; Recover the PAL line-to-line phase inverting signal *palsw* [11, 12]. The method is: If the nth line is NTSC line and the initial phase of color burst is $\hat{\theta}_n$, then the (n+1)th line is PAL line and its initial phase is $\hat{\theta}_{n+1}$, according to the principle of PAL system, $\hat{\theta}_n = \hat{\theta}_{n+1}$, $palsw = \cos(\hat{\theta}_n - \hat{\theta}_{n+1}) = 1$; Similarly, if the nth line is PAL line, then $\hat{\theta}_n = \hat{\theta}_{n+1} + \pi$, $palsw = \cos(\hat{\theta}_n - \hat{\theta}_{n+1}) = -1$.

3 Frequency Offset Estimation

The subcarrier frequency of non standard TV signal may drifts slightly within the range of ±500Hz [1]. When the frequency offset exists, the value of frequency offset can be considered as stable in a few contiguous lines. Because of the principle that frequency is the gradient of phase changing, the frequency offset can be estimated from the linear difference between the estimated phases of 2 lines.

According to the principle of PAL system, the phases of the interlaced lines are continuous. The two interlaced estimated phases are used for frequency offset estimation. Assuming the estimated phase of the nth h line is $\hat{\theta}_n$, the estimated frequency offset of the nth line is $\hat{\Delta f}_n$, the actual frequency offset is Δf ,then the expressions of $\hat{\theta}_{n+2}"$ which is the NCO output result after two lines of oscillation from the initial phase $\hat{\theta}_n$ and the estimated phase of the (n+2)th line $\hat{\theta}_{n+2}$ are:

$$\begin{aligned} \hat{\theta}_{n+2}" &= \hat{\theta}_n + 2\pi(f + \hat{\Delta f}_n) * (6400 / f_s) \\ \hat{\theta}_{n+2} &\approx \hat{\theta}_n + 2\pi(f + \Delta f) * (6400 / f_s) \end{aligned} \tag{6}$$

Where f_s is the sampling frequency (50MHz), 6400 is the number of sampling points of two lines, then:

$$\Delta f - \hat{\Delta f}_n \approx \frac{(\hat{\theta}_{n+2} - \hat{\theta}_{n+2}") * f}{2\pi * 6400} \tag{7}$$

So, the recursive expression of frequency offset is:

$$\hat{\Delta f}_{n+1} = \frac{\hat{\Delta f}_n + (\hat{\theta}_{n+2} - \hat{\theta}_{n+2}")}{2\pi * 6400} \tag{8}$$

Table 2 shows the results of the numerical simulation experiment of frequency offset estimation.

Table 2 The results of frequency offset estimation

Actual Δf （Hz）	Estimation $\hat{\Delta f}$ （Hz）
-500	-503.36
-440	-443.00
-380	-384.72
-300	-301.10
-200	-200.82
-120	-120.88
-100	-101.02
-80	-82.20
-10	-11.70
0	0.08
10	10.30
50	52.12
90	90.14
100	101.07
120	122.00
150	150.40
200	201.51
270	268.87
300	303.32
350	351.55
400	398.08
500	595.68

The estimated frequency offset of every iteration can be sent to the low-pass filter, in order to obtain the average value to increase estimation accuracy. Before adjusting the oscillation frequency of NCO, the value of the estimated frequency offset should be limited first to prevent the situation that the estimated frequency offset has a big deviation from the actual value because of the factors such as noise.

4 Experimental Result

With the estimated phase and frequency offset, the local color subcarrier can be resumed to demodulate the chroma signal. The cutoff frequency of the LPF which filters out the high frequency component in the chroma signal can be 0.6MHz or 1.3MHz. The demodulated YUV component signals are sent into the matrix converter to get the RGB signals. According to the principle of television, if the phase error of the local subcarrier is within ±5^{o} range, the variations in the hue of the picture is not visible. Through the phase and frequency offset estimation, the phase error of both standard and nonstandard TV signal is controlled within the scope of the permit.

Figure 3 and Figure 4 are the pictures decoded in the matlab platform, of which the sources are the PAL standard TV signal generator and the DVD player. The pictures have little difference with the TV show, and achieve a good display effect. The estimated frequency offset is about 38Hz, which proves that the signal source output signal is standard.

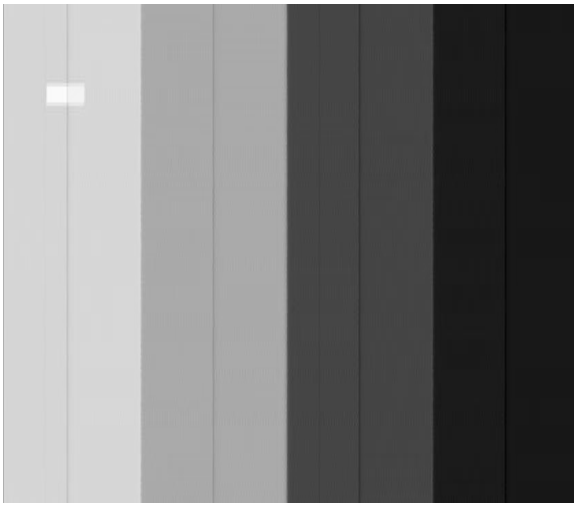


Figure 3 The decoded color bar picture

Figure 4 The decoded portrait picture

5 Conclusions

This paper studies a novel color subcarrier recovery algorithm used in the digital video decoder. The method that combines the use of the phase and frequency offset estimation to recovery a high-precision local color subcarrier is proposed. Also, the other related technology is studied. Through the experiment verification, the method is proved effective.

References

[1] Keith Jack, Video Demystified 4rd Edition, New York: Elsevier, 2005

[2] Y. Suzuki, T. Gai, M. Ymakawa, “NTSC / PAL / SECAM Digital Video Decoder with High-Precision Resamplers”, IEEE Transactions on Consumer Electronics, 51(1), 2005, pp.287-294

[3] R. Donald, Phase-Locked Loops for Wireless Communications Digital, Analog and Optical Implementations 2rd Edition, New York: Kluwer Academic, 2002

[4] E.B. Roland, Phase-locked loops:Design, Simulation, and Applications, Beijng: :Tsinghua University Press, 2004

[5] J. Dennis, “High quality digital PAL decoding for single frames”, IEEE Electronics Letters, 34(23), 1998, pp.2221-2222

[6] Kong Zhuang, Fang Jia-Yin, “Subcarrier Recovery on the Basis of the Spectrum of PAL Signal”, Journal of the Academy of Equipment Command & Technology, 15(3), 2004, pp.94-96

[7] R. Lares, A. Rothermel, “Sync Signal Processing for Asynchronously Sampled Video Signals”, IEEE Transactions on Circuits and Systems, 3(3), 2000, pp.575-578

[8] A. Rothermel, R. Lares, “Synchronization of Analog Video Signals with Improved Image Stability”, IEEE Transactions on Consumer Electronics, 49(4), 2003, pp.1292-1300

[9] Wang Chua-Chin, Lee Ching-Li, Chang Ming-Kai, “Low-cost Video Decoder with 2D2L Comb Filter for NTSC Digital TVs”, IEEE Transactions on Consumer Electronics, 51(22), 2005, pp.684-697

[10] H. Beintken, M. Hahn, “System-on-silicon for 100Hz TV:New Concepts Qualified for Highest Integration”, IEEE Transactions on Consumer Electronics, 46(3), 2000, pp.812-818

[11] J. Yuan, Q.H. Shen, “Digital video based on LSI OSDC”, Journal of Nanjing University(Natural Sciences), 39(4), 2003, pp.510-516

[12] F.L. Chao, Q.H. Shen, “Realization of keystone correction based on a new programmable gate array chip”, Journal of Nanjing University(Natural Sciences), 42(4), 2006, pp.362-367

An Unsupervised Fuzzy Clustering Method for Shot Clustering

Zhihao Zhou Xiaonan Chen

Department of Information Engineering, Wuxi Professional College of Science and Technology
Wuxi, Jiangsu, 214028

Abstract

With the development of the internet, the needs of video retrieval become more and more high. So we should investigate new methods to do the retrieval to instead of the old methods which are mainly by manual marking and so on. Using the fuzzy clustering methods, we can extract the key frames and key shots by similarity measures such as SCD and CLD for it. With the key frames and key shots, the retrieval result will be more efficient. In this paper, we proposed a content-based retrieval method which mainly with the algorithm MRLC to get superior performance and representation capability with less manual operation.

Keyword: Fuzzy Clustering, MRLC Algorithm, Shot Retrieval, Key Shot, Extraction, Similarity Measure

1 Introduction

With the constitution of the coding audio-visual information Standard MPEG (Moving Picture Experts Group) -1/2/4 and Multimedia Content Description Interface Standard Interface Standard MPEG-7 [1, 2], digital video is used in a great of in the field of the education, entertainments and so on. Because of the limit of the description capability, the strength of the subjectivity and the mark by hand, traditional keyword-based video retrieval could not be satisfied with the great capacity for liquor need. It is always to segment the shots first, so it could use the shots to be basic cells of the construction and retrieval for the video serial. And then to extract the key frames, which are representing the shots. Having done these, the problem of the content-based video retrieval is translated into the content-based image retrieval.

It is supported that by using on-line unsupervised clustering [3, 4], shots can be retrievable. We can use this method to extract the key frames and key shots by similarity measures. The proposed method lessens the problems what there are quite many parameters settled artificially in a certain extent.

2 Expression of Video Content

2.1 Shot Segmentation

Shot conversion contains two methods. They are abrupt changing and gradual changing [5]. The joint between two shots is always at the transformation point of the scene. To check the difference between two adjacent frames, we can do the shot segmentation. The SDM (Space Difference Measure) doesn’t take account of the movement of vidicon. Using the SDM may bring on the error check results. So when we use it, we must take movement compensation into account in the compare process. By shifting up and down, left and right in a distance respectively, we can get the maximum SDM from these data. There are 25 frames per second in the sample video, so the difference between two adjacent frames is quite little. Considering it, the bound of the movement which is in the comparing process is taken ±1/25 each border in our experiment which is suitably (fixed by the video swatch).

For a pixel (i, j) in the frame t, let $I_t(i, j)$ denote its RGB value. The *SDM* D_s is defined for the adjacent frame pair (i, j) respectively as

$$D_s(I_i, I_j) = \frac{1}{M \times N} \cdot \sum_{x=1}^{M}\sum_{y=1}^{N} | I_i(x, y) - I_j(x, y) | \qquad (1)$$

The modified SDM D_s ‘ is defined as

$$D_s^{'}(I_i,I_j)=\max\left[\frac{1}{(M-|p_M|)\times(N-|p_N|)}\times\right.$$
$$\left.\sum_{x=1}^{M}\sum_{y=1}^{N}|I_i(x-p_M,y-p_N)-I_j(x,y)|\right] \quad (2)$$

Where p_M and p_N are stand movement for each border, who is a serial value calculate from

$$p_M=\lfloor -M/25\rfloor,\lfloor -M/25\rfloor+1,\cdots,\lfloor M/25\rfloor-1,\lfloor M/25\rfloor$$
$$p_N=\lfloor -N/25\rfloor,\lfloor -N/25\rfloor+1,\cdots,\lfloor N/25\rfloor-1,\lfloor N/25\rfloor$$

Having got all D_s ‘ of whole video, we could take every frame into two kinds. One is distinct changing, the other is indistinct changing. If a frame belongs to the distinct kind, we use 1 to denote it, and else is 0. The video can be expressed as a binary array in such way, such as 011100010…. By the array, the shot abrupt changing and gradual changing can be detected. The module 010 denotes shot abrupt changing, and the module 011 or 110 denotes shot gradual changing. With D_s‘$(I_i, I_j)_{\max}$<0.4, it is considered that shot changing doesn’t happen. With D_s‘$(I_i, I_j)_{\max}$>0.6, the shot abrupt changing happens. If there are continuous (or single) D_s‘$(I_i, I_j)_{\max}$ whose values are between 0.4 and 0.6, it is indicated that shot gradual changing may appears. In this time, we should use nonlinear characters space which is composed of RGB Color feature difference and HSV Color feature difference with histogram, to enlarge the difference. Thereby, the precision is can be enhanced.

2.2 Feature Extraction

The content of the video is various, so using the only one feature is insufficient. We import two detect features to make the similarity measure more robust in this paper. In the experiment, color descriptor SCD and CLD are using as the vision characters. SCD (scalable color descriptor) is a histogram of HSV color space. It is used to describe the color distribution status across-the-board, and it isn’t sensitive to the image rolling and flexing. CLD (color layout descriptor) is a descriptor of YCbCr color space. It is used to describe the color distribution status of an image or a certain shape area. These two descriptors are the complements each other.

SCD is used to calculate the QHDM (Quadratic Histogram Distance Measure). To divide the chroma, saturation and luminance these three degree into 16, 4 and 4 different grades for each. At last, we get a eigenvector of 256 grades. It depends on the size of sample frame. The bigger the picture is, more the grades it has, so as the color digits. For a frame i, let $\mathbf{H}_i$ denote its HSV histogram. The size of a frame is M × N. The $\mathbf{H}_i$ is defined as

$$\mathbf{H}_i=\{\,h_{1,1,1}\,,\,\ldots,\,h_{i,j,k}\,,\,\ldots,\,h_{16,4,4}\}$$
$$i=1,\ldots,16;\,j=1,\ldots,4;\,k=1,\ldots,4 \quad (3)$$

The Similarity Measure of *SCD* S_{SCD} is defined as

$$S_{SCD}(i,j)=\frac{1}{M\times N}\cdot\sum_{h=1}^{16}\sum_{s=1}^{4}\sum_{v=1}^{4}\min\left\{H_i\left(h,s,v\right),H_j(h,s,v)\right\} \quad (4)$$

The modified histogram intersection algorithm proposed above could reduce redundancy and expense greatly.

CLD should change the original color space to YCbCr color space. Then disjoin the *Y*, *Cb* and *Cr* to do the DCT changing. After that, by doing the heterogeneous quantization for the DCT coefficient of each degree, we do Zigzag scanning. At last, CLD feature should be by Gaussian normalization.

The D_{DCT} is defined for the frame pair (i, j) respectively as

$$D_{DCT}(i,j)=\sqrt{\sum_{k=1}^{6}\omega_Y\left(DY_{i_k}-DY_{j_k}\right)^2}+$$
$$\sqrt{\sum_{k=1}^{3}\omega_Y\left(DCb_{i_k}-DCb_{j_k}\right)^2}+$$
$$\sqrt{\sum_{k=1}^{3}\omega_Y\left(DCr_{i_k}-DCr_{j_k}\right)^2} \quad (5)$$

Where $\sum\omega_k$ is equa to 1, $\omega_k\in[0,1]$, $k\in\{Y, Cr, Cb\}$

The Similarity Measure of *CLD* S_{CLD} is defined as

$$S_{CLD}(i,j)=\frac{\max\left(\sum_{y=1}^{6}\sum_{cb=1}^{3}\sum_{cr=1}^{3}\left\{Y_i\left(y,cb,cr\right),Y_j\left(y,cb,cr\right)\right\}\right)-D_{DCT}(i,j)}{\max\left(\sum_{y=1}^{6}\sum_{cb=1}^{3}\sum_{cr=1}^{3}\left\{Y_i\left(y,cb,cr\right),Y_j\left(y,cb,cr\right)\right\}\right)} \qquad （6）$$

The Similarity Measure of the frame pair (i, j) is defined as

$$Sim(i,j)=\omega_1\cdot S_{SCD}\,(i,j)+\omega_2\cdot S_{CLD}\,(i,j) \qquad （7）$$

Where $\omega_1+\omega_2=1$, $\omega_1\in[\,0,1\,]$

3 The Method of Shot Retrieval

3.1 Key Frame Extraction

In the paper, key frames can be chosen by using on-line clustering algorithm. The fundamental is that, after clustering inside of the shot, take the frame which are closest to the center of the class as the key frame. And the number of the key frame is lying on the number of the class of the shot mainly. It means that the shot will be divided into several semishots, including subshots[6]. An algorithm called MRLC (Modified Radius-center Leader-follower Clustering) is detailed in the paper. It is based on the on-line unsupervised clustering algorithm Leader-Follower. The algorithm settles the thresholds automatically with the number of frame, the variance of the adjacent frame and so on which are known quantity. With the adaptive thresholds, the key frames are automatically [7] extracted by the unsupervised clustering method.

Input: An Sample *Shot* ***Sh***, which include *NF* frames, where $\mathbf{f}_i$ (i = 1, 2,…, NF) denotes the frame i.

Output: Key frame *Array* **KF** = { $\mathbf{f}_1{}'$, $\mathbf{f}_2{}'$,…, $\mathbf{f}_{nf}{}'$ }, $nf\geqslant 1$.

Using the unsupervised clustering, *NF* frames are clustered into *MF* classes. They are $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$,…, $\boldsymbol{\theta}_{MF}$. $\mathbf{C}_1$, $\mathbf{C}_2$,…, $\mathbf{C}_{MF}$ denote the centers of each class, and R_1, R_2,…, R_{MF} denote the radius of each class, and L_1, L_2,…, L_{MF} denote the frame number of each class. The similar measure of *frame* $\mathbf{f}_i$ and *frame* $\mathbf{f}_j$ is calculated by the formula *Sim* (i, j) above. Control the number of the class is by the *threshold* α_F. The process as below:

Step 1. Calculate the similar measure of the shot. Get the *Array* **QS** = {QS_1, QS_2,…, QS_{NF-1}};

Step 2. Get the sequence of the *Array* **QS** in sort ascending as $QS_1{}'\geqslant QS_2{}'\geqslant\ldots\geqslant QS_{NF-1}{}'$, the requested *value T* is defined as

$$\mathrm{T}=\arg\min\sigma \qquad （8）$$

Where $\boldsymbol{\sigma}=M_u\cdot\boldsymbol{\sigma_u}^2+M_d\cdot\boldsymbol{\sigma_u}^2$, $M_u=T$, $M_d=N-T-1$, $m_u=\frac{1}{M_u}\sum_{i=1}^{T}QS_i^{'}$, $m_d=\frac{1}{M_d}\sum_{i=T+1}^{N-1}QS_i^{'}$, $\sigma_u{}^2=\frac{1}{M_u}\sum_{i=1}^{T}\left(QS_i^{'}-m_u\right)^2$, $\sigma_d{}^2=\frac{1}{M_d}\sum_{i=T+1}^{N-1}\left(QS_i^{'}-m_d\right)^2$ QS_T is the threshold we want. Now let α_F denote it. If *threshold* α_F bigger than 0.9, we take it as 0.9;

Step 3. Initialize the clustering: Put the *frame* $\mathbf{f}_1$ into the *class* $\boldsymbol{\theta}_1$, and $\mathbf{f}_1$ also is *center* $\mathbf{C}_1$ of the *class* $\boldsymbol{\theta}_1$. L_1 is 1, and the *number M* of the classes is 1;

Step 4. As if the aggregation is not empty, take the next *frame* $\mathbf{f}_i$:

① Using the formula proposed in Section 2 above, let calculate the similarity measure of frame pair ($\mathbf{f}_i$, $\mathbf{C}_k$), where $\mathbf{C}_k$ is center of the known *class* $\boldsymbol{\theta}_k$ (k = 1, 2,…, MF). Then choose the biggest one of the *Sim* ($\mathbf{f}_i$, **Ck**) as S_{max}, and register the *value k*. The S_{max} is defined as

$$S_{max}=\max\ \{Sim_{ik}\,(\mathbf{f}_i,\mathbf{C}_k)\}\ k=1,2,\ldots,M \qquad （9）$$

② If $S_{\max}<\alpha_F$, it tells that *frame* $\mathbf{f}_i$ is far away from known classes. At this time, a new class comes into being. As Step 3, just M plus 1 become new M. Else if $S_{max}\geqslant\alpha_F$, put the *frame* $\mathbf{f}_i$ into the *class* $\boldsymbol{\theta}_k$. Before the joining of new element, we regard the original radius as $R_k{}'$. The new *Radius* R_k is defined as

$$R_k=[(1\text{-}S_{\max})+R_k{}'\cdot L_k\,]\,/\,(L_k+1) \qquad （10）$$

The new *Center* R_k is $\mathbf{C}_k$ equals to $\mathbf{f}_{ki}$, and the *coefficient ki* is defined as

$$ki=\arg\min|\,R_{ki}-R_k\,| \qquad （11）$$

Where the *coefficient ki* is the *swatch i* of the *class k*, R_{ki} is the average of the distance of the *swatch i* and other swatch for each. The *Radium* R_{ki} is defined as

$$R_{ki} = \frac{1}{L_k + 1} \cdot \sum_{j=1}^{L_j} \left\| \mathbf{f}_{ki} - \mathbf{f}_{kj} \right\| \qquad （12）$$

The new number L_k of the *class* $\boldsymbol{\theta}_k$ adds one.

Step 5. Otherwise, the clustering process is over;

Step 6. When the clustering is finished, if the frame *number* L_i of the *class i* is smaller than 10 (get the 2/5 number of frames per second, and the number is 25 in the paper), or smaller than N / M, we will fail it. In other situation, preserve its center $\mathbf{f}_{ki}$ in the key frame *Array* **KF**;

Step 7. Save and show on the result, the key frame extraction.

3.2 Key Show Extraction

After the key frame extraction, we can use the result to do the shot retrieval. But not every shot is in need, what we need are the representative ones. Some corresponding key frames may show the same scenes, or similar content. For example, in the long-distance education video, the teacher scenes repeat again and again. Feeding back all the key frames is redundant, so one is enough. Having done the key frame extraction, we do the clustering to the shots. Accordingly, the reconstruction for the hierarchy of the video data has been done. The shot retrieval is divided into two levels, one is common retrieval, the other is key retrieval, and it can satisfy different demands.

The clustering method starts with the original algorithm MRCL, and then according to the criterion of breaking and merging [8], we should manage the initial clustering result to amend the error that may come into being in initial clustering. There is something be different from the key frame extraction process. It reduces the limit of the samples' number. Using the fuzzy clustering, we can do the adjusting for the *m* key frames.

Input: An Sample *Video* ***Vi***, which include *NS* shots, where $\mathbf{s}_i$ (i = 1, 2,…, *NS*) denotes the *semishot i*.

Output: Key Shot *Array* **KS** = { $\mathbf{s}_1^{'}$, $\mathbf{s}_2^{'}$,…, $\mathbf{s}_{ns}^{'}$ }, *ns* ≥1.

Using the unsupervised clustering, *NS* shots are clustered into *MS* classes. They are $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$,…, $\boldsymbol{\theta}_{MS}$. $\mathbf{C}_1$,$\mathbf{C}_2$,…,$\mathbf{C}_{MS}$ denote the centers of each class, and R_1, R_2,…, R_{MS} denote the radius of each class, and L_1,L_2,…,L_{MS} denote the frame number of each class. The similar measure of *shot* $\mathbf{s}_i$ and s*hot* $\mathbf{s}_j$ is calculated by formula *Sim*(*i*, *j*) above. Control the number of the class is by the *threshold* α_S. The process as below:

Step 1. Initialize the clustering: choose the value. The value is the average of the thresholds (those are less than or equal to the center value) got from the *Array* a_F. It should be bigger than the threshold of key frame extraction relatively. With the threshold we can do the clustering.

Step 2. Calculate the maximum spaces $Spa_{\max}$ of every class inside and the minimum spaces $Gap_{\min}$ of each pair of classes for all. The *space Spa*(*i*, *j*) between *shot* $\mathbf{s}_i$ and *shot* $\mathbf{s}_j$ is defined as

$$Spa(i, j) = 1 - Sim(\mathbf{s}_i, \mathbf{s}_j) \qquad （13）$$

The *space Gap* (*i*, *j*) between *class* $\boldsymbol{\theta}_i$ and *class* $\boldsymbol{\theta}_j$ is defined as

$$Gap(i, j) = 1 - Sim(\mathbf{C}_i, \mathbf{C}_j) \qquad （14）$$

When $Spa_{\max} > (1 - a_S)$ and $Spa_{\max} > Gap_{\min}$, it tells us that we can do the breaking the largest maximum space which class has. We also use the clustering algorithm MRCL to divide it into two classes, and get their centers. The number of classes adds one. If there is no class of this kind, the adjusting is over. Else continue.

Step 3. Calculate the spaces of each pair of classes for all and the radius of each class. If the space of a pair of classes is less than the sum of their radius, and the average of their radius is less than the average of the whole shot, we should affirm whether these two classes are the two broken recently. And if not, we use the clustering algorithm MRCL to unite them into one class, and then get its center. The number of classes reduces one.

4 Experiment Results

The video contains 1710 frames in the test, which size is 352 × 288, such as meeting scenes, the speakers, the wooden figures and so on. It can embody the quality of the method intuitively. We gained 57 key frames and 32 key shots.

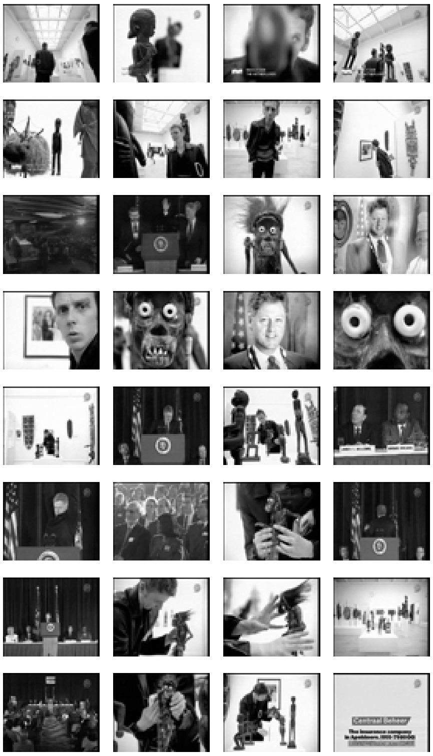


Figure 1 The key shots group got by the proposed method.

Compared with the clustering ways in the past, based on the algorithm MRCL, our method increases the autoadaptation function. The paper takes ANMRR and AR as the objective appraising method for the search results.

Figure 2 Shots' sample frames for query: sample（1）, sample（2）and sample（3）from left to right.

Table 1 Value NMRR and R of 3 kinds of methods for each.

Kind	1st		2nd		3rd	
Shot Query	NMRR	R	NMRR	R	NMRR	R
（1）	0.133	1.00	0.466	0.50	0.061	1.00
（2）	0.450	0.60	0.325	0.75	0.340	0.75
（3）	0.175	1.00	0.232	1.00	0.171	1.00
Average	0.253	0.80	0.341	0.75	0.191	0.92

The inner change of the chosen shot query is great, the similar scenes and the relative scenes are much. The 1st method takes the middle frame of the shot as the key frame. After the retrieval, there are 57 semishots in the key shots group. The method 2 is based on, the key shot group contains 81 semishots. The 3rd is the method proposed in this paper. From the Table 1, we can see that to the Query Shot（2）, the result of method 1 is the worst, because only one key frame is in lack and it isn't representative. Using the method 2, the result group is short of one shot. The result with the method proposed in this paper is more perfect. We can conclude that our method owns its feasibility and validity

5 Conclusion and Discussion

With the on-line unsupervised clustering, we do the key frames extraction and key shots extraction step by step, and using some adjusting, the result can be used in shot retrieval. It shows that this method is efficient subjectively and objectively, removes the influence from the experience on the thresholds basically, and indicates superior performance and representation capability of the proposed method here for the video. However, the problem of retrieval of shots and images based on semantic content can not be satisfied only with color-based visual descriptors, regardless of their efficiency. Techniques from information science, artificial intelligence, and combination of content analysis techniques of multiple information sources (audio, text) are required to achieve satisfactory retrieval results.

References

[1] Sikora T, Member S. The MPEG-7 Visual Standard for Content Description - An Overview [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2001, 11（6）: 696 – 702

[2] José M M, Rob K, Fernando P. Mpeg-7: The Generic Multimedia Content Description Standard [J]. IEEE Multimedia, 2002, 4: 78 – 87

[3] Zhaoqi Bian, Xuegong Zhang. Pattern Recognition [M]. Beijing: Tsinghua University Press, 2000

[4] Hong Jin, Guirong Shi, Yuanhua Zhou. Hierarchical Video Content Representation using Unsupervised Fuzzy Clustering Methods [J]. Computer Engineering and Applications, 2002, 2: 163-165

[5] Xinbo G, Xiaoou T. Unsupervised Video-shot Segmentation

and Model-free Anchorperson Detection for News Video Story Parsing [J]. IEEE Trans. on Circuits and Systems for Video Technology, 2002, 12（9）: 765 – 776

[6] Tong Lin, Hongjiang Zhang, Jufu Feng. Shot Content Analysis for Video Retrieval Applications [J]. Journal of Software, 2002, 2: 163 – 165

[7] Fangshi Lin, De Xu, Weixin Wu. A Cluster Algorithm of Automatic Key Frame Extraction Based on Adaptive Threshold [J]. Journal of Computer Research and Development, 2005, 42（10）: 1752 – 1757

[8] Hua Xiong, Xiaofeng Hu. A Self-Adjusting Shot-Clustering Technique without Experiential Parameters [J]. Journal of Image and Graphics, 2001, 6（3）: 243 – 249

[9] Qinjie Dong, Yuxin Peng, Zongming Guo. A New Approach for Content-based Shot Retrieval by Fuzzy Classification [J], 2004.1: 56 – 57, 102

Image Segmentation Based on Improved Genetic Algorithm

Xu Ling[1,2]

1 School of Information technology, Jiangnan University, Wuxi, Jiangsu, 214122

2 Lake Tai College, Jiangnan University, Wuxi, Jiangsu, 214064

Abstract

An evolutionary image segmentation method by genetic algorithm (GA) is presented. GA has global search abilities but it may suffer from premature convergence on multi-modal problems. This is due to the decrease of population's diversity in search space that leads to fitness stagnation. An improved GA (IGA) is proposed and the mutation rate in IGA is controlled by the population's diversity which can balance the population's diversity and have more chances located in the region of global solutions. The effectiveness of IGA is shown by the experiment of image segmentation, the results show that IGA has great advantage of convergence speed and success ratio over GA.

Keywords：Genetic Algorithm; global optimization; image segmentation

1 Introduction

Segmentation refers to the grouping of image elements that exhibit similar characteristics. Image segmentation is the first essential and important step of low level vision of image processing techniques. Results of the image segmentation can reveal pictorial information of regions with similar properties and is a crucial step in image analysis. Image segmentation is a complex visual computation problem and is the key step from the image processing to the image analysis. The other tasks of image segmentation are object detection, feature extraction, object recognition and classification. Many approaches have been used to image segmentation, such as [1-5]. Some of these methods are more or less heuristic and specific to a particular application. These methods can be boiled down to two types which are boundary detection-based approaches and region clustering-based approaches. Though much attention has been attached to it for many years, it develops very slowly. Bhanu and Lee pointed out that an efficient and heuristic method is needed to solve the problem of image segmentation [2]. They said the method can efficiently search the complex space of plausible parameter combinations and locate the values which yield optimal results. The approach should not be dependent on the particular application domain nor should it have to rely on detailed knowledge pertinent to the selected segmentation algorithm.

Genetic Algorithms (GAs) [7,8] are computational models inspired from biological evolution, which is based on natural selection and genetics, involving a structured yet randomized information exchange resulting in the survival of the fittest amongst a population of string structures. GAs can be used to search for an optimal solution to the evolution function of an optimization problem. The main characteristics of GAs are the information exchange between the individuals, excellent searching ability and the ability that prevents the premature convergence. So it is very suitable for the nonlinear problems that the traditional algorithm can not solve [7]. The backbone of GA is the reproduction of an original population, the performance of crossover and mutation and the selection of the best. The search area for the GAs is very wide and it usually converges to a point near the global optimum. But a major problem with GA in multi-modal optimization is premature convergence, which results in great performance loss and sub-optimal solutions. The main reason for premature convergence is a too high selection pressure or a too high gene flow between population individuals [6]. Then it is important to improve GA and make it with less probability trap into the local minima. In this paper, the population's diversity is calculated during the search procedure. The main purpose is to alter the mutation rate adaptively according to the population's diversity and make the algorithm search efficiently. By using the improved GA (IGA) in solving the problem of image segmentation, it is shown that IGA can get better results than GA.

The remaining of this paper is organized as follows. In section 2, the problem formulation of image segmentation is introduced. GA and IGA are described in section 3. In section 4, the steps of using IGA in solving the problem of image segmentation and the experimental results are shown. Finally, conclusions are giving in section 5.

2 Problem Formulation of Image Segmation

In histogram entropy method, the concept of entropy in information theory is used for image segmentation, which was proposed by Pun [9]. Based on the method in [9], Kapur and Sahoo proposed the KSW entropy method based on two distribution hypothesis. In this paper, we will use KSW entropy method for image segmentation. The details of KSW entropy method is as follows.

According to the concept of entropy, for an image which gray level range is $[0, L-1]$, its entropy is

$$H = -\sum_{i=0}^{L-1} p_i \ln p_i \qquad (1)$$

where p_i is the appearing probability of grey level *i*.

Suppose an image is partitioned into target W and background B by threshold *t*. Therefore, the distribution of $[0,t]$ is

$$\frac{p_0}{P_t}, \frac{p_1}{P_t}, \cdots, \frac{p_t}{P_t}, \qquad (2)$$

and the distribution of $[t+1, L-1]$ is

$$\frac{p_{t+1}}{1-P_t}, \frac{p_{t+2}}{1-P_t}, \cdots, \frac{p_{L-1}}{1-P_t}, \qquad (3)$$

where

$$P_t = \sum_{i=0}^{t} P_i \qquad (4)$$

Let

$$H_t = -\sum_{i=0}^{t} p_i \ln p_i . \qquad (5)$$

Then for the above two distributions, the entropy of them are $H_W(t)$ and $H_B(t)$ respectively and their expressions are as following,

$$H_B(t) = \ln P_t + \frac{H_t}{P_t}, \qquad (6)$$

$$H_W(t) = \ln(1-P_t) + \frac{H - H_t}{1-P_t}, \qquad (7)$$

where $H(t)$ is the summation of $H_B(t)$ and $H_W(t)$,

$$H(t) = \ln P_t(1-P_t) + \frac{H_t}{P_t} + \frac{H-H_t}{1-P_t} . \qquad (8)$$

From the equations above, one can see that the image will get the best segment results which are the target and background if *t* makes the value of entropy $H(t)$ maximize.

3 GA and IGA

3.1 Overview of GA

The GAs is a population-based, robust search and optimization technique, which finds applications in numerous practical problems, especially used to tackle high-dimensional, multi-modal search space problems. GAs has been shown to outperform conventional non-linear optimization and local search techniques on difficult search spaces. The robustness of the GA is due to its capacity to locate the global optimum in a multi-modal landscape. The flowchart of GA is shown in figure 1.

In GA, a chromosome is considered as an individual and each individual represents one candidate solution of the given problem. Then the population is consisting of a constant size of individuals. Each individual is encoded by a fixed length string, which is usually binary string. The individual components of the string are known as genes and each may take one of a small range of values. From figure 1, during the search procedure of GA, after the old population of solutions is evaluated according to the fitness function, the rule of survival of the fittest is used to select parents from the old population. And then the crossover operator is applied to pairs of parents to create offsprings and they will be mutated through the mutation operator according to a certain probability. The mutated individuals will insert into a new population. The details of GA used in this paper are described below.

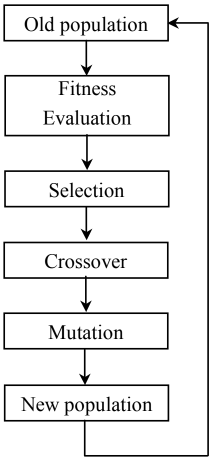


Figure 1 Flowchart of GA

a) Encoding and fitness evaluation

The most critical problem in applying a genetic algorithm is in finding a suitable encoding of the examples in the problem domain to a chromosome. A good choice of representation will make the search easy by limiting the search space; a poor choice will result in a large search space. The individuals usually represented by fixed size of strings. In this paper, the problem is to segment the grey image and the object is the pixels. As the value of a pixel ranges from 0 to 255, we will take the binary encoding and the length of the string is fixed 8. The form of the string is shown in Figure 2

a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7

Figure 2 Binary encoding

In figure 2, each bit can take values 0 or 1. For three individuals X and Y and Z:

X= 00110111, Y=11001010, Z= 10100101.

These three binary strings can decoded to decimal values which can be treated as fitness values, which are $f(X) = 55$, $f(Y) = 202$, $f(Z)=165$. For a minimum optimization, X is the fittest.

b) Selection

After all the individuals evaluated, according to the fitness value the population will apply the selection operator. Many schemes have been proposed for the selection process of choosing parents for subsequent recombination, such as roulette wheel selection, stochastic universal sampling and tournament selection. One of the most popular methods is roulette wheel selection (also called fitness proportionate selection). In this selection method, the fitness functions must be nonnegative. An individual's slice of a Monte Carlo-based roulette wheel is an area proportional to its fitness. The "wheel" is spun in a simulated times and the parents are chosen based on where the pointer stops. Another popular approach is called tournament selection. In this method, chromosomes are compared in a "tournament," with the better chromosome being more likely to win. The tournament process is continued by sampling (with replacement) from the original population until a full complement of parents has been chosen. The most common tournament method is the binary approach, where one selects two pairs of chromosomes and chooses as the two parents the chromosome in each pair having the higher fitness value. Empirical evidence suggests that the tournament selection method often performs better than roulette selection. In this paper, we will choose the tournament selection.

c) Crossover operator and crossover rate

The crossover operator will be applied after the selection operator. The crossover operation creates offspring of the pairs of parents from the selection step. As binary encoding is considered, there are three mechanisms for crossover, which are single-point crossover, multiple-point crossover and uniform crossover. In this paper, we will choose the single-point crossover mechanism. Suppose X and Y are the parents selected, the crossover occurs between the three and the fourth binary position, swapping the fragments of the two strings produces two offsprings *X’* and *Y’*, where *X’*= 00101010, *Y’*=11010111.

A crossover probability Pc is used to determine if the offspring will represent a blend of the chromosomes of the parents. If no crossover takes place, then the two offspring are clones of the two parents. If crossover does take place, then the two offspring are procedure according to an interchange of parts of the chromosome structure of the parents. The crossover rate decides the frequency of crossover operator. The algorithm can obtain a fast convergence speed if the crossover rate is high. But too high crossover rate will cause premature convergence. Generally, Pc= 0.4~0.9.

d) Mutation operator and mutation rate

Mutation operator is the final steps of GA, which can provide the new individuals. Mutation occurs randomly and very rarely both in natural and artificial genetic systems. Because the initial population may not contain enough variability to find the solution via crossover operations alone, the GA also uses a mutation operator where the chromosomes are randomly changed. When the mutation does happen to a individual, one bit of the chromosome is chosen and set to its complementary value.

Chosen of mutation rate (Pm) is influenced by the population size, the length of chromosomes and etc. Mutation of a given bit occurs with small probability. Generally, Pm=0.001~0.1.

The above operations are repeated until a certain conditions are satisfied.

3.2 Improvement of GA (IGA)

As mentioned in section 1, the premature convergence is a major problem with GA in multi-modal optimization. Premature will result in great performance loss and sub-optimal solutions. And the main reason for premature convergence is a too high selection pressure or a too high gene flow between population individuals. The selection operator will make the population’s diversity decline and then the individuals may stuck at local optima and difficult to escape from them. It is known that the mutation operator can make the diversity increase. While the mutation rate of GA is usually a small value and during the iterations it is fixed. Therefore, in order to improve the performance of GA, we can modify the mutation rate according to the population’s diversity in order to find out the best region containing the global optimum and escape from the local minima. Based on the motivation, we firstly give the diversity measurement of the population. In the paper, the diversity of the population is set according to the following equation:

$$Div(population) = \frac{1}{N}\sum_{i=1}^{N}\frac{|f_i - \overline{f}|}{f_{\max} - \overline{f}}, \qquad (9)$$

where N is the number of individuals, f is the fitness value of the *i*th individual and $\overline{f}$ is the average fitness value of the population, $f_{\max}$ is the maximum fitness value of the population. From the equation, one can see that if the population is collected, the value of *Div* will small and the value of *Div* will be large if the individuals distribute in a wide space.

In the improved GA, the mutation rate is determined by the value of *Div*, that is

$$pm = \begin{cases} pm_1 & Div > div \\ pm_2 & Div > div \end{cases}. \qquad (10)$$

In the above equation, Pm$_1$ and Pm$_2$ are the mutation rates which are settled by the users and Pm$_1$> Pm$_2$, div is the lowest limit of the diversity which is settled according to the problems.

Through this method, the population can perform the mutation operator according to a big probability in order to generate more new individual and then the population can keep a certain diversity.

4 Experimental Settings and Results

The experiment was carried out with matlab 7 on the computer with P4 2.0G Hz and 512M memory. The standard lenna image is taken for segmentation. GA and IGA were used to complete the image segmentation. For

GA and IGA, the population size is 20, the maximum iteration is 500, crossover rate is 0.7. For GA, the mutation rate is 0.01. For IGA, Pm_1 is 0.1 and Pm_2 is 0.01, *div* is 1. The experiment executes 20 times independently.

The segmentation time of image Lenna that used by GA and IGA is shown in table 1. The averaged iteration that reach converge and success times of segmentation are shown in table 2.

From table 1, IGA cost much less time than GA which means IGA can perform much faster than GA. From table 2, the convergence speed of IGA is more rapid than GA. In the total 20 runs, IGA success more times than GA. From thes results, IGA exhibited better performance than GA in the image segmentation.

Table 1 Segmentation Time

Method	Averaged time(s)	Accelerate ratio
GA	0.42	3.231
IGA	0.13	

Table 2 Results of Segmentation

	GA	IGA
Averaged evolution iterations	8.56	5.33
Success times	18	15

5 Conclusions

In this paper, an improved GA is proposed to enhance the ability of GA escaping from the local minima. In IGA, the mutation rate is controlled by the population's diversity which can balance the population's diversity and have more chances located in the region of global solutions. In the experiment of image segmentation, the results show that IGA can get better performance than GA.

References

[1] R.M. Haralick, L.G. Shapiro, Image Segmentation Techniques, Computer Vision, Graphics and image Processing, Vol.29, pp.100-132, 1985

[2] B.Bhanu, S.Lee, and S.Das, Adaptive image segmentation using multi-objective evaluation and hybrid search methods, AAAI Fall Symp. Machine Learning in Computer Vision, Oct, 1993

[3] B.Bhanu, S.Lee and J.C.Ming Self-optimizing image segmentation system using a genetic algorithm, in Proc. Fourth Int. Conf. on Genetic Algorithms, San Diego, GA, July 1991, pp.362-369

[4] M.Cheriet, J.N.Said, and C.Y.Suen, A recursive threshold ing technique for image segmentation, IEEE Trans. Image Processing, vol. 7, pp.918-921, 1998

[5] N.Pal, and S.Pal, A review on image segmentation techniques, Pattern Recognition, vol.26, no.9, pp.1277-1294, 1993

[6] Rasmus K. Ursem, Diversity-Guided Evolutionary Algorithms, Proceedings of Parallel Problem Solving from Nature VII (PPSN-2002), pp. 462-471,2002

[7] J.H. Holland. Adaptation in Natural and Artificial systems. The MIT Press, Cambridge, Massuchusetts, 1992

[8] D.H.Goldberg. Genetic Algorithms- In Search, Optimization and Machine Learning. Addison-wesley Publishing Company, 1989

[9] T. Pun. A new method for grey-level picture thresholding using the entropy of the histogram. Signal Processing, 1980, 2（3）: 223～237

[10] J. N. Kapur, P. K. Sahoo, A. K. C. Wong. A new method for gray-level picture thresholding using the entropy of the histogram. Computer Vision, Graphics, and Image Processing, 1985, 29（3）: 273～285

Acquire 3D Information of Vehicle Body Contour from Two Picture

Lifang Chen

The School of Information Technology, Jiangnan University 1800 Lihu Road, Wuxi, Jiangsu , 214122, China

Email: may7366@163.com

Abstract

In this paper one mathematical model for extracting 3D information from two pictures is built on vehicle body contour design. A new method based on six known 3D coordinates points or consulting objects using characteristic point stereo matching— "characteristic point marking, multiple angle picturing and corresponding point match" for extracting vehicle body contour from two pictures are presented. The 3D coordinates are calculated by Direct Linear Transformation Techniques(DLTT).In the last part of this paper, several application research are made taking feihu vehicle hubcap as example.

Keywords： Vehicle body contour design, Image processing, 3D information extracting, Direct Linear Transformation Techniques

1 Introduction

A new tape of vehicle takeing place by conventional sculpture model facture must pass 3-5 years, it is not seasoned with the information age of fast change, so extracting 3D information from 2D pictures by digital image processing will be adopted in the future. aim at the insufficient actuality of vehicle trade of china and use for reference from the other CAD trade, extracting 3D information from 2D pictures by digital image processing or acquiring 3D geometric data of vehicle for a CAD model from images by photogrammetry technique are presented. these technique will cut procreative period of produce and can be used to accelerate the development of economy.

2 The Acquisition of the 2D Data at the normal Case of Photogrammetry

Each point, which is required for the complete restitution of the object, has to be displayed on at least two images from different points of view. Fig 1 is a geometric vehicle model at the normal case of photogrammetry. By a series of algorithmic commutation, the relation of the coordinate of picture and the dimensional coordinate of object can be expressed as Eq.（1）

$$\begin{cases} x+f\dfrac{a_1(X-X_s)+b_1(Y-Y_s)+c_1(Z-Z_s)}{a_3(X-X_s)+b_3(Y-Y_s)+c_3(Z-Z_s)}=0 \\ y+f\dfrac{a_2(X-X_s)+b_2(Y-Y_s)+c_2(Z-Z_s)}{a_3(X-X_s)+b_3(Y-Y_s)+c_3(Z-Z_s)}=0 \end{cases} \quad (1)$$

where X,Y,Z is the dimensional coordinate of the point M of the object;X_S,Y_S,Z_S is the dimensional coordinate of the center photogrammetry; x,y is the coordinate of image; a_i,b_i,c_i (i=1,2,3)is the direction cosine of is the dimensional coordinate of the object and is the dimensional coordinate of the image; f is the focus of video camera.

From Eq.（1）,we can gain a result what every point of one picture can list two equation ,but the 3D coordinate of the point must have three parameter, so we must require at least two picture to acquire the 3D coordinate. for it is desired to view the object in a stereoscopic manner, the images have to be taken according to the so called normal case of photogrammetry

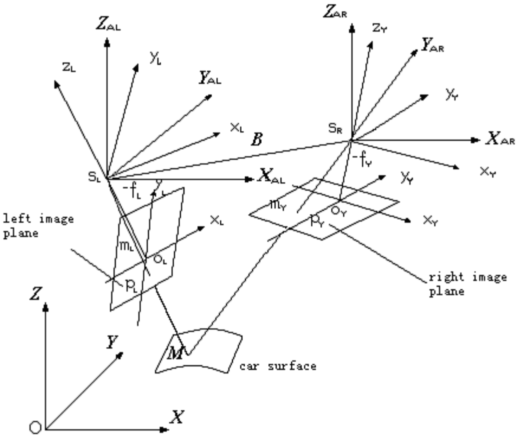


Figure 1　Geometric model by the normal case of photogrammetry

with nearly parallel directions of view from two points on a horizontal base, perpendicular to the viewing directions. This arrangement is similar to the arrangement of the human eyes. For this purpose at least three control points and 5-10 tie points per image are required. Tie points have to be identified and measured in at least two images. For each image at least six unknowns (three coordinates for the position, three rotations and if required parameters of the interior orientation) and for each object point three coordinates have to be estimated. The more control and tie points are available, the better the results of the orientation process in terms of accuracy and reliability can be obtained. A new method based on six known three dimension points or consulting objects using characteristic point stereo matching— “characteristic point marking, multiple angle picturing and corresponding point match” for extracting vehicle body contour from two pictures are presented. where the introduce of characteristic point marking such as:

（1）select characteristic point, there are always some especially points, these point have distinct and assured place, so we call these points as characteristic point what include spinous point、inflexion and border point. we set then as initial point .

（2）partition area, we partition the object by different tinctorial color line, so we can identify in two image.

（3）insert poin, insert different chromatic point to the area which is partitioned.

Then we can acquire the pictures by the normal case of photogrammetry. In the courts of taking picture it is necessary to identify the pictures of one object point in the different images. so we can match the 3D object from two picture.

The matching process is express as Figure 2 .by semi-automatic match, we can acquire the coordinate value of image(xl,yl)and (xr,yr) from two picture.

3 The Calculation of the 3D Coordinate of VEHICLE Body Contour

The calculational methods of extracting the vehicle body contour has many kinds, In many cases the photoparameter measure is adopted, but the method lay some disadvantages such as the error is not amended. In the paper Direct Linear Transformation Techniques(DLTT) is adopted. DLTT is a measure which can adjust pickup camera’s error and the aberration of photo and calculational error, also can calculate the 3D coordinate of the surface of the object and import non-line error. In theory, the host point of quadrant must be superposition to the center point of photo, but in commonly instances camera axis is not lay the center of photo. so the coordinate of host point of quadrant is unknowns, we must adjust the value of x and y of the Eq.（1）. assume that the correctional modulus of x and y as U and v_y , base on the Eq.（1）, after we import a series of line error and non-line error and all kinds of aberration of revised modulus of photo, we can list error’s equation which can extract unknowns modulus in the relation modulus of the Direct Linear Transformation, it can be expressed as Eq.（2）

$$\begin{cases} v_x = -\frac{1}{A}\left[l_1X + l_2Y + l_3Z + l_4 + xXl_9 + xYl_{10} + xZl_{11} + A(x - x_0)r^2k_1 + x\right] \\ v_y = -\frac{1}{A}\left[l_5X + l_6Y + l_7Z + l_8 + yXl_9 + yYl_{10} + yZl_{11} + A(y - y_0)r^2k_1 + y\right] \end{cases} \tag{2}$$

where $l_i(i = 1 \sim 11)$ is unknowns modulus ; $A = l_9X + l_{10}Y + l_{11}Z + 1$; k_1 is non-linear aberration modulus of objective；r is vector radius

$$r = \left[\left(x - x_0\right)^2 + \left(y - y_0\right)^2 \right]^{1/2} ;$$

the other signal is the same as Eq.（1）。

the error equation and the relevant vertical equation of matrix form is expressed as Eq.（3）

$$\begin{cases} V = ML + W \\ M^T ML + M^T W = 0 \end{cases} \qquad (3)$$

where

$$V = \begin{bmatrix} v_x & v_y \end{bmatrix}^T$$

$$M = -\frac{1}{A}\begin{bmatrix} X & Y & Z & 1 & 0 & 0 & 0 & 0 & xX & xY & xZ & A\left(x - x_0\right)r^2 \\ 0 & 0 & 0 & 0 & X & Y & Z & 1 & yX & yY & yZ & A\left(y - y_0\right)r^2 \end{bmatrix}$$

$$L = \begin{bmatrix} l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 & l_8 & l_9 & l_{10} & l_{11} & k_1 \end{bmatrix}^T$$

$$W = -\frac{1}{A}\begin{bmatrix} x \\ y \end{bmatrix}$$

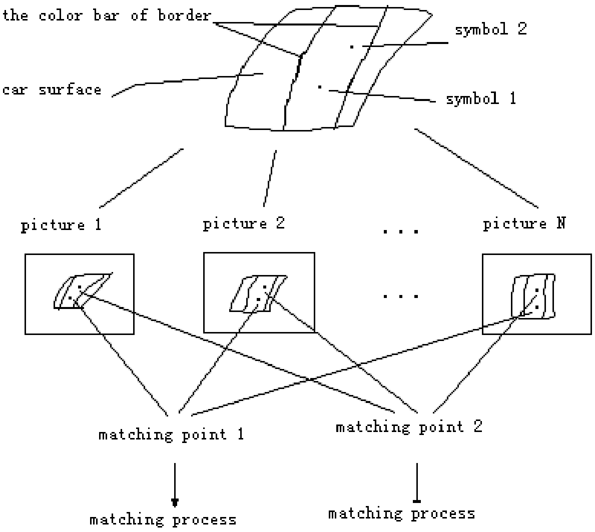


Figure 2 Process of corresponding point match

if the vehicle body surface or Reference object have at least six known points, then we can acquire 12 unknowns modulus of the left photo and the right photo by express(3).so we can acquire the 3D coordinate of the vehicle body surface for it has the 12 known modulus.

For extracting the 3D coordinate of the vehicle body surface, we construct the error equation of left photo as Eq. (4)

$$V\,\mathrm{n}^l = N^l S + Q^l \qquad (4)$$

where

$$V\,\mathrm{n}^l = \begin{bmatrix} v\mathrm{n}_x^l & v\mathrm{n}_y^l \end{bmatrix}^T$$

$$N^l = -\frac{1}{A^l}\begin{bmatrix} l_1^l + l_9^l x^l & l_2^l + l_{10}^l x^l & l_3^l + l_{11}^l x^l \\ l_5^l + l_9^l y^l & l_6^l + l_{10}^l y^l & l_7^l + l_{11}^l y^l \end{bmatrix}$$

$$S = \begin{bmatrix} X & Y & Z \end{bmatrix}^T$$

$$Q^l = -\frac{1}{A^l}\begin{bmatrix} l_4^l + x^l \\ l_8^l + x^l \end{bmatrix}$$

$$A^l = l_9^l X + l_{10}^l Y + l_{11}^l Z + 1$$

and construct the error equation of left photo as Eq. (5)

$$V\,\mathrm{n}^r = N^r S + Q^r \qquad (5)$$

where

$$V\,\mathrm{n}^r = \begin{bmatrix} v\mathrm{n}_x^r & v\mathrm{n}_y^r \end{bmatrix}^T$$

$$N^r = -\frac{1}{A^r}\begin{bmatrix} l_1^r + l_9^r x^r & l_2^r + l_{10}^r x^r & l_3^r + l_{11}^r x^r \\ l_5^r + l_9^r y^r & l_6^r + l_{10}^r y^r & l_7^r + l_{11}^r y^r \end{bmatrix}$$

$$S = \begin{bmatrix} X & Y & Z \end{bmatrix}^T$$

$$Q^r = -\frac{1}{A^r}\begin{bmatrix} l_4^r + x^r \\ l_8^r + x^r \end{bmatrix}$$

$$A^r = l_9^r X + l_{10}^r Y + l_{11}^r Z + 1$$

combining the left photo with the right photo, we can acquire the error equation and the vertical equation of the 3D coordinates of object as Eq. (6)

$$\begin{cases} \begin{bmatrix} V\,\mathrm{n}^l \\ V\,\mathrm{n}^r \end{bmatrix} = \begin{bmatrix} N^l \\ N^r \end{bmatrix} S + \begin{bmatrix} Q^l \\ Q^r \end{bmatrix} \\ \begin{bmatrix} N^l \\ N^r \end{bmatrix}^T \begin{bmatrix} N^l \\ N^r \end{bmatrix} S + \begin{bmatrix} N^l \\ N^r \end{bmatrix}^T \begin{bmatrix} Q^l \\ Q^r \end{bmatrix} = 0 \end{cases} \qquad (6)$$

in the express（6）,there have four equations and 3 unknown data(X，Y，Z), which is super determinate equation, so we can adopt smallest two multiplication sign to extract the 3D coordinates.

4 Applications

Image Acquisition: To verify the result of extracting 3D coordinates from two photo and the models of validity, first we mark the signal in the feihu vehicle hubcap by sticking chromatic circinal plane on the feihu vehicle hubcap and insert the different altitudinal bar, then we take pictures for feihu vehicle hubcap flat from two different points of view, acquiring the left picture and the right picture as Figure 3 and Figure 4 .

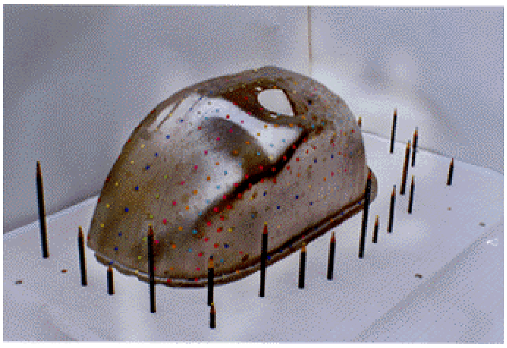


Figure 3 Left picture of feihu vehicle hubcap

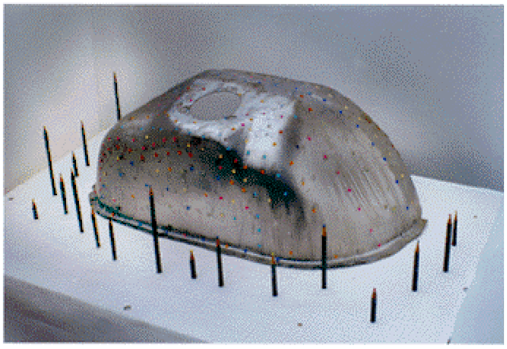
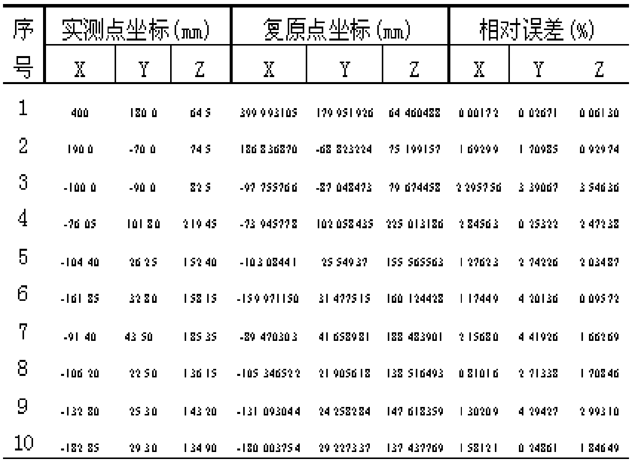


Figure 4 Right picture of feihu vehicle hubcap

Corresponding Points: in the two picture, by semi-automatic match we can acquire the corresponding point coordinate of the different chromatic point. semi-automatic match is necessary to support the matching process by a human operator through depend on software such as visual c etc.

Analyze error: by calculation, we can acquire the 3D coordinates of feihu vehicle hubcap. we list Part data of recovered result and on-the-spot survey as table1. and Opposite error for recovered 3D coordinates of feihu vehicle hubcap by Direct Linear Transformation Techniques of the normal case of photogrammetry as table 2.

Table 1 Part data of recovered result and on-the-spot survey of feihu vehicle hubcap



序号	实测点坐标(mm)			复原点坐标(mm)			相对误差(%)		
	X	Y	Z	X	Y	Z	X	Y	Z
1	400	180 0	64 5	399 993105	179 951926	64 460488	0 00172	0 02671	0 06130
2	190 0	-70 0	74 5	186 836870	-68 823224	75 199157	1 69299	1 70985	0 92974
3	-100 0	-90 0	82 5	-97 755766	-87 048473	79 674458	2 295756	3 39067	3 54636
4	-76 05	101 80	219 45	-73 945778	102 058435	225 013186	2 84563	0 25322	2 47238
5	-104 40	26 25	152 40	-103 08441	25 54937	155 565563	1 27623	2 74226	2 03487
6	-161 85	32 80	158 15	-159 971150	31 477515	160 124428	1 17449	4 20136	0 09572
7	-91 40	43 50	185 35	-89 470303	41 658981	188 483901	2 15680	4 41926	1 66269
8	-106 20	22 50	136 15	-105 346522	21 905618	138 516493	0 81016	2 71338	1 70846
9	-132 80	25 30	143 20	-131 093044	24 258284	147 618359	1 30209	4 29427	2 99310
10	-182 85	29 30	134 90	-180 003754	29 227337	137 437769	1 58121	0 24861	1 84649

Table 2 the Opposite error of the 3D reductive coordinates of feihu vehicle hubcap

fractional error (%)	Direct Linear Transformation Techniques of the normal case of photogrammetry
X	0.00172～2.84563
Y	0.02671～4.41926
Z	0.06130～3.54636

From table 2 we can acquire the maximal Opposite error is 5.62032,the error is mainly from two factors: one is camera ,the other is calculation.

Image Restitution: By AutoCAD software, A result of a restitution on feihu vehicle hubcap is displayed in Figure 5 .

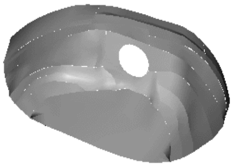


Figure 5 the result of a restitution on feihu vehicle hubcap

5 Conclusions

The paper triumphantly realize extracting 3D coordinates of the vehicle body outline from two pictures, and for the maximal fractiona error is 5.62032,so the measure content in substance in the design of the vehicle body outline. with further developments in digital photogrammetry, it will also become a fast and economic in the other production of formative design.

References

[1] D.Marr and E.Hildreth,Theory of edge detection,Proc.Royal Soc.,London,Vol.B207,1980,P187-217

[2] M.Ito and A.Ishii, Computer Vision ,IEEE trans.Pattern Anal.and Machine Intell.,Vol PAMI-8.No.4,July 1986, P362-369

[3] Marsha,Jo.Hannah,Photogra.Eng.and Remote Sensing, Vol. 55, No.12,Dec.1989,P1765-1770

[4] ALBERTZ, JQRG & ALBERT WIEDEMANN 1995: Acquisition of CAD data from existing buildings by photgrammetry.In Eds: P. J. Pahl & H. Werner, Computing in Civil and Building Engineering, A.A.Balkema, Rotterdam,Brookfield 1995, pp. 859-866

[5] S.Peleg,IEEE Trans.Pattern Anal.Machine Intell., Vol. PAMI-2,1980,P362-369

[6] Rafael C. Gonzalez and Richard E. Woods, Digital mage Processing, Prentice Hall, 2nd edition, January 2002

[7] James G. Nagy and Dianne P. O'Leary, Restoring images degraded by spatially variant blur, SIAM Journal on Scientific Computing, 19（4）,1998,P1063–1082

[8] H. J. Trussel and S. Fogel, Identification and restoration of spatially variant motion blurs in sequential images, IEEE Transactions on Image Processing,1（1）,1992,P123–126

[9] T. Haist and H.J. Tiziani, Iterative nonlinear joint transform correlation for the detection of objects in cluttered scenes, Preprint of Haist ET AL. Optics Communications 161, 1999,P 310-317

[10] D. I. Barnea and H. F. Silverman,A class of algorithms for fast digital image registration, IEEE Transactions on Computers C-21, 1972,P179-186

[11] D.Marr and T.Poggio.Tech.Rep.AI Memo 451,Artificial Intel.Lab.,Mass.Inst.Technol.,CamBridge,Nov.1977,P213-215

A Simple Algorithm of B-Spline Wavelets Fairing Based-on Geometric Meanings on Reverse Engineering

Xiaogang Ji

School of Mechanical Engineering, Jiangnan University, Wuxi, Jiangsu, 214122, China

Email: bhearts@jiangnan.edu.cn, brookstonewx@126.com

Abstract

Being an excellent filter tool, wavelet analysis is applied to curve and surface fairing in reverse engineering more and more universally. In wavelet fairing, B-Spline basis function is used as wavelet basis popularly. Unfortunately, B-Spline basis functions and corresponding wavelets are lack of translation orthogonality, widely used Mallat rapid algorithm does not work for this kind of wavelet decomposition and reconstruction. On the basis of analysis of decomposition and reconstruction theory of B-Spline wavelets, inherent relationships between different dilating and translating serieses on different scale of B-Spline basis functions are researched thoroughly. The solution process of wavelet decomposition and reconstruction is described by clear geometry meanings and the corresponding solution of reconstruction matrix P_j is elaborated, too. This algorithm is clear, efficient and robust and avoids the Abstract and complexity of wavelet analysis. At last, an example of decomposition and reconstruction of complicated curve is provided by means of this algorithm. It is proved to be feasible to fairing of complicated curve. By tensor product operation, this algorithm can be applied to surface fairing easily.

Keywords：Reverse engineering; Wavelets; B-Spline Curve; Multiresolution Analysis; Decomposition and reconstruction; Computer Graphics

1 Introduction

Wavelet analysis technology is a new time-frequency analysis mathematical method developed in 80’s. It is a useful tool for depicting internal relations of data. Different from traditional Fourier analysis, wavelet analysis has locality of time domain and frequency domain at the same time. In recent years, with the development of wavelet analysis technology, it begins to be used in the field of computer graphics widely. In 1994, Quak E and Weyrich N put forward wavelet decomposition and reconstruction algorithm based on B-Spline wavelets on closed interval[1]. Based on this algorithm, Finkelstein A and Salesin D H studied the multiresolution presentation and editing of B-Spline curves based on semi-orthogonal B-Spline wavelets[2,8]. Stollnitz E J、DeRose T D et al. described the wavelet application to curves and surfaces easily and clearly, too[3]. At the same time, Zhu Xinxiong, Sun Yankui and Zhao Gang et al. also studied the applications of wavelet technology on reverse engineering[4,6,7,9,10]. Unfortunately, detail calculation process of reconstruction matrix is not introduced in any existing papers. In this paper, on the basis of analysis on B-Spline basis function geometric meanings and wavelet analysis essential, an easy algorithm to reconstruction matrix is derived. Although this algorithm does not follow by rigorous mathematics derivation, its thinking is reasonable and it is proved to be correct and acceptable.

2 Basic of B-Spline Wavelets Analysis

Cubic B-Spline wavelet is studied in this paper.

V_j is a finite linear space on closed interval [0,1] with scalar production. It is made up of endpoint interpolating cubic B-Spline basis functions in 2^j uniform intervals. Its dilation and translation series constitute a nested set of linear vector spaces

$V_0 \subset V_1 \subset V_2 \cdots$

At integer scale *j*, a B-Spline curve is controlled by column matrix $C_j = [c_1^{(j)}, c_2^{(j)}, \ldots, c_{2^j+3}^{(j)}]^T$, which is constituted by 2^j+3 control points $c_1^{(j)}, c_2^{(j)}, \ldots, c_{2^j+3}^{(j)}$. The substance of wavelet analysis is to create a low-resolution version C_{j-1} of C_j with a smaller number of coefficients 2^{j-1}+3 approximately and seize lost detail with column matrix D_{j-1}, which is used in recovering C_j accurately. Usually, P_j, Q_j is called reconstruction matrixes and A_j, B_j is called decomposition matrixes. The above process can be expressed as matrix equations:

$$C_j = P_j C_{j-1} + Q_j D_{j-1} \tag{1}$$

$$C_{j-1} = A_j C_j, \qquad D_{j-1} = B_j C_j \tag{2}$$

The process of wavelet decomposition and reconstruction about above two equations is show in Figure 1.

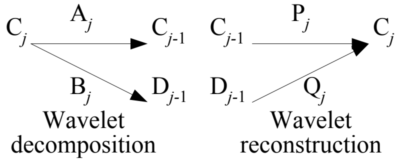


Figure 1 Wavelet decomposition and reconstruction

Because V_j is a nested set of vector space, decomposition process can be used in new control points recursively. So C_j can be expressed as tower structure with low-resolution version $C_{2^{j-1}+3}$, $C_{2^{j-1}+2}$, ···, C_0 and detail version $D_{2^{j-1}-1}$, $D_{2^{j-1}-2}$, ···, D_0, as shown in Figure 2.

There exists inner production in space V_j and for *f*, *g*∈V_j, <*f*, *g*> means inner production. At scale *j*-1, there is only one orthogonal complement of $V_{j\text{-}1}$ in V_j, denoted by $W_{j\text{-}1}$, i.e. $W_{j\text{-}1}$={*f*∈V_j|<*f*, *g*>=0, for any *g*∈$V_{j\text{-}1}$}. $W_{j\text{-}1}$ is called wavelet space and $V_j = V_{j\text{-}1} \oplus W_{j\text{-}1}$.

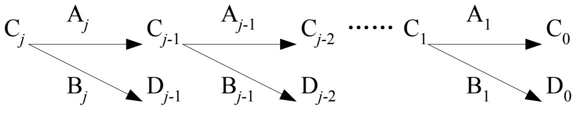


Figure 2 Recursion process of wavelet analysis

For any function f_j∈V_j, there exits only function $f_{j\text{-}1}$ ∈$V_{j\text{-}1}$ and $g_{j\text{-}1}$∈$W_{j\text{-}1}$, which meet following function

$$f_j = f_{j\text{-}1} + g_{j\text{-}1} \tag{3}$$

Then scaling functions in $V_{j\text{-}1}$ and wavelets in $W_{j\text{-}1}$ are arranged in row vector and denoted ϕ_{j-1} and ψ_{j-1} respectively, which compose a group of bases of $V_{j\text{-}1}$ and $W_{j\text{-}1}$. As a result

$$f_j = \phi_j C_j \tag{4}$$

$$f_{j-1} = \phi_{j-1} C_{j-1} \tag{5}$$

$$g_{j-1} = \psi_{j-1} D_{j-1} \tag{6}$$

Combining Eq.（3）gives

$\phi_j C_j = \phi_{j-1} A_j C_j + \psi_{j-1} B_j C_j = (\phi_{j-1} A_j + \psi_{j-1} B_j) C_j$, so

$$[\phi_{j-1} \mid \psi_{j-1}] \left[\frac{A_j}{B_j} \right] = \phi_j \tag{7}$$

At the same time, because $V_{j\text{-}1} \subset V_j$, $W_{j\text{-}1} \subset V_j$, there exist constant matrixes P_j and Q_j, which satisfy the following equations

$$\phi_{j-1} = \phi_j P_j \tag{8}$$

$$\psi_{j-1} = \phi_j Q_j \tag{9}$$

Combining Eq.（8）and Eq.（9）gives

$$[\phi_{j-1} \mid \psi_{j-1}] = \phi_j [P_j \mid Q_j] \tag{10}$$

Combining Eq.（7）and Eq.（10）gives

$$\left[\frac{A_j}{B_j} \right] = \left[P_j \mid Q_j \right]^{-1} \tag{11}$$

3 Process of Decomposition and Reconstruction of B-Spline Wavelets

According to above discussion, the general process of decomposition and reconstruction of B-Spline wavelets can be summed up as follows.

3.1 Solution of Reconstruction Matrix P_j

Bases ϕ_j and ϕ_{j-1} in vector space V_j and $V_{j\text{-}1}$ are expressed as $\phi_j = [\varphi_1^j, \varphi_2^j, \cdots, \varphi_{2^j+3}^j]$, $\phi_{j-1} = [\varphi_1^{j-1}, \varphi_2^{j-1}, \cdots, \varphi_{2^{j-1}+3}^{j-1}]$ respectively. Solution of P_j with Eq.（8）amount to find a certain coefficient series, so any scaling function φ_i^{j-1} of ϕ_{j-1} can be made up of ϕ_j linearly, i.e. $\varphi_i^{j-1} = \sum_{l=1}^{2^j+3} \varphi_l^j P_{l,i}^j$. In a general way, P_j can be solved according to relationship between ϕ_j and ϕ_{j-1}. This is the first step to wavelet decomposition and reconstruction.

3.2 Solution of Reconstruction Matrix Q_j

Cubic B-Spline $\varphi(t)$ and corresponding wavelet $\psi(t)$ are lack of translation orthogonality[5], so only orthogonality between $\varphi(t)$ and $\psi(t)$ can be assured when wavelets are deduced. This kind of wavelets are called semi-orthogonal B-Spline wavelets. According to orthogonality between $\varphi(t)$ and $\psi(t)$, we have

$$(\phi_{j-1})^T(\psi_{j-1}) = 0 \qquad （12）$$

Substituting Eq.（8）and Eq.（9）into Eq.（12）gives

$$(P_j)^T[(\phi_j)^T \cdot (\phi_j)]Q_j = 0 \qquad （13）$$

[·] means that any element of matrix is equal to inner production of corresponding scaling function of ϕ_j. Because P_j and ϕ_j are known a prior, Q_j can be solved with Eq.（13）. In a general way, Q_j is not unique. Q_j with more naught elements is more suitable.

3.3 Solution of Decomposition Matrixes A_j and B_j

Now, P_j and Q_j have been obtained, so A_j and B_j can be solved with Eq.（11）. Unfortunately, obtained matrixes A_j and B_j are popularly dense, for convenience, not A_j and B_j but direct decomposi- tion results C_{j-1} and D_{j-1} are solved.

According to Eq.（2）, we have

$$\begin{bmatrix} C_{j-1} \\ \hline D_{j-1} \end{bmatrix} = \begin{bmatrix} A_j \\ \hline B_j \end{bmatrix} C_j \qquad （14）$$

Substituting Eq.（14）into Eq.（11）gives

$$\left[P_j \mid Q_j\right] \begin{bmatrix} C_{j-1} \\ \hline D_{j-1} \end{bmatrix} = C_j \qquad （15）$$

Here, P_j、Q_j、C_j are known a prior, so the decomposition of B-Spline curve can be sum up as the solution of linear equation system, Eq.（15）.

Now, decomposition and reconstruction of wavelets are finished. It is easy to see that the most important step of above discussiong is the solution of reconstruction matrix P_j, the following solutions are all on the basis of the fact that P_j is known a prior.

4 Easy Solution of Reconstruction Matrix P_j

The graph of B-Spline $\varphi(t)$ is shown in Figure 3, its dyadic dilation and translation basis functions are $\varphi^{j,k} = \varphi(2^j t - k)_{k\in Z}$.

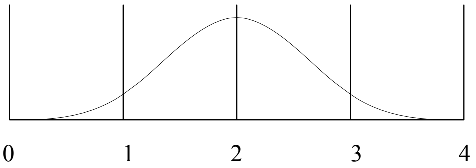


Figure 3 B-Spline wavelet basis function

According to section 2.1, the basis functions at low-resolution j-1 can be composed of basis functions at high-resolution j, i.e. the ith component $\varphi_i^{j-1} = \sum_{l=1}^{2^j+3} \varphi_l^j P_{l,i}^j$. The process of construction is shown in Figure 4.

Although P_j is unknown, P_j is subsistent and determined. In theory, equation is always tenable to different t. From the contrary point of view, if specific t are determined, a linear equation system can be constructed to solve $P_{l,i}^j$. This is the concrete thinking of algorithm. Generally, t is equal to the middle value of demain of every B-Spline sector, as shown in Figure 4 by dashed lines. The obtained linear equation system is as follows:

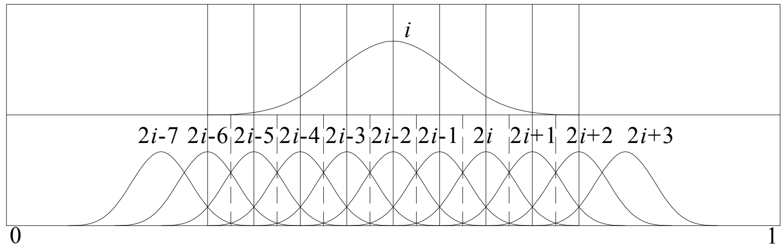


Figure 4 Geometry meanings of constitution of low-resolution basis functions

$$\begin{aligned}
&\vdots\\
0 &= \varphi_{2i-8}^{j}(\tfrac{2i-7}{2^{j+1}})P_{2i-8,i}^{j} + \varphi_{2i-7}^{j}(\tfrac{2i-7}{2^{j+1}})P_{2i-7,i}^{j} + \varphi_{2i-6}^{j}(\tfrac{2i-7}{2^{j+1}})P_{2i-6,i}^{j} + \varphi_{2i-5}^{j}(\tfrac{2i-7}{2^{j+1}})P_{2i-5,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i-5}{2^{j+1}}) &= \varphi_{2i-7}^{j}(\tfrac{2i-5}{2^{j+1}})P_{2i-7,i}^{j} + \varphi_{2i-6}^{j}(\tfrac{2i-5}{2^{j+1}})P_{2i-6,i}^{j} + \varphi_{2i-5}^{j}(\tfrac{2i-5}{2^{j+1}})P_{2i-5,i}^{j} + \varphi_{2i-4}^{j}(\tfrac{2i-5}{2^{j+1}})P_{2i-4,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i-3}{2^{j+1}}) &= \varphi_{2i-6}^{j}(\tfrac{2i-3}{2^{j+1}})P_{2i-6,i}^{j} + \varphi_{2i-5}^{j}(\tfrac{2i-3}{2^{j+1}})P_{2i-5,i}^{j} + \varphi_{2i-4}^{j}(\tfrac{2i-3}{2^{j+1}})P_{2i-4,i}^{j} + \varphi_{2i-3}^{j}(\tfrac{2i-3}{2^{j+1}})P_{2i-3,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i-1}{2^{j+1}}) &= \varphi_{2i-5}^{j}(\tfrac{2i-1}{2^{j+1}})P_{2i-5,i}^{j} + \varphi_{2i-4}^{j}(\tfrac{2i-1}{2^{j+1}})P_{2i-4,i}^{j} + \varphi_{2i-3}^{j}(\tfrac{2i-1}{2^{j+1}})P_{2i-3,i}^{j} + \varphi_{2i-2}^{j}(\tfrac{2i-1}{2^{j+1}})P_{2i-2,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i+1}{2^{j+1}}) &= \varphi_{2i-4}^{j}(\tfrac{2i+1}{2^{j+1}})P_{2i-4,i}^{j} + \varphi_{2i-3}^{j}(\tfrac{2i+1}{2^{j+1}})P_{2i-3,i}^{j} + \varphi_{2i-2}^{j}(\tfrac{2i+1}{2^{j+1}})P_{2i-2,i}^{j} + \varphi_{2i-1}^{j}(\tfrac{2i+1}{2^{j+1}})P_{2i-1,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i+3}{2^{j+1}}) &= \varphi_{2i-3}^{j}(\tfrac{2i+3}{2^{j+1}})P_{2i-3,i}^{j} + \varphi_{2i-2}^{j}(\tfrac{2i+3}{2^{j+1}})P_{2i-2,i}^{j} + \varphi_{2i-1}^{j}(\tfrac{2i+3}{2^{j+1}})P_{2i-1,i}^{j} + \varphi_{2i}^{j}(\tfrac{2i+3}{2^{j+1}})P_{2i,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i+5}{2^{j+1}}) &= \varphi_{2i-2}^{j}(\tfrac{2i+5}{2^{j+1}})P_{2i-2,i}^{j} + \varphi_{2i-1}^{j}(\tfrac{2i+5}{2^{j+1}})P_{2i-1,i}^{j} + \varphi_{2i}^{j}(\tfrac{2i+5}{2^{j+1}})P_{2i,i}^{j} + \varphi_{2i+1}^{j}(\tfrac{2i+5}{2^{j+1}})P_{2i+1,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i+7}{2^{j+1}}) &= \varphi_{2i-1}^{j}(\tfrac{2i+7}{2^{j+1}})P_{2i-1,i}^{j} + \varphi_{2i}^{j}(\tfrac{2i+7}{2^{j+1}})P_{2i,i}^{j} + \varphi_{2i+1}^{j}(\tfrac{2i+7}{2^{j+1}})P_{2i+1,i}^{j} + \varphi_{2i+2}^{j}(\tfrac{2i+7}{2^{j+1}})P_{2i+2,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i+9}{2^{j+1}}) &= \varphi_{2i}^{j}(\tfrac{2i+9}{2^{j+1}})P_{2i,i}^{j} + \varphi_{2i+1}^{j}(\tfrac{2i+9}{2^{j+1}})P_{2i+1,i}^{j} + \varphi_{2i+2}^{j}(\tfrac{2i+9}{2^{j+1}})P_{2i+2,i}^{j} + \varphi_{2i+3}^{j}(\tfrac{2i+9}{2^{j+1}})P_{2i+3,i}^{j}\\
0 &= \varphi_{2i+1}^{j}(\tfrac{2i+11}{2^{j+1}})P_{2i+1,i}^{j} + \varphi_{2i+2}^{j}(\tfrac{2i+11}{2^{j+1}})P_{2i+2,i}^{j} + \varphi_{2i+3}^{j}(\tfrac{2i+11}{2^{j+1}})P_{2i+3,i}^{j} + \varphi_{2i+4}^{j}(\tfrac{2i+11}{2^{j+1}})P_{2i+4,i}^{j}\\
&\vdots
\end{aligned}$$

Because t is equal to the middle value of domain and every corresponding $\varphi_{l}^{j}(t)$ is not equal to 0, according to the beginning and end equations, $P_{1,i}^{j}$, $P_{2,i}^{j}$, $\cdots$, $P_{2i-5,i}^{j}$, $P_{2i+1,i}^{j}$, $P_{2i+2,i}^{j}$, $\cdots$, $P_{2^{j}+3,i}^{j}$ are all equal to 0, so the equation system can be simplified as follows:

$$\begin{aligned}
\varphi_{i}^{j-1}(\tfrac{2i-5}{2^{j+1}}) &= \varphi_{2i-4}^{j}(\tfrac{2i-5}{2^{j+1}})P_{2i-4,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i-3}{2^{j+1}}) &= \varphi_{2i-4}^{j}(\tfrac{2i-3}{2^{j+1}})P_{2i-4,i}^{j} + \varphi_{2i-3}^{j}(\tfrac{2i-3}{2^{j+1}})P_{2i-3,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i-1}{2^{j+1}}) &= \varphi_{2i-4}^{j}(\tfrac{2i-1}{2^{j+1}})P_{2i-4,i}^{j} + \varphi_{2i-3}^{j}(\tfrac{2i-1}{2^{j+1}})P_{2i-3,i}^{j}\\
&\quad + \varphi_{2i-2}^{j}(\tfrac{2i-1}{2^{j+1}})P_{2i-2,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i+1}{2^{j+1}}) &= \varphi_{2i-4}^{j}(\tfrac{2i+1}{2^{j+1}})P_{2i-4,i}^{j} + \varphi_{2i-3}^{j}(\tfrac{2i+1}{2^{j+1}})P_{2i-3,i}^{j}\\
&\quad + \varphi_{2i-2}^{j}(\tfrac{2i+1}{2^{j+1}})P_{2i-2,i}^{j} + \varphi_{2i-1}^{j}(\tfrac{2i+1}{2^{j+1}})P_{2i-1,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i+3}{2^{j+1}}) &= \varphi_{2i-3}^{j}(\tfrac{2i+3}{2^{j+1}})P_{2i-3,i}^{j} + \varphi_{2i-2}^{j}(\tfrac{2i+3}{2^{j+1}})P_{2i-2,i}^{j}\\
&\quad + \varphi_{2i-1}^{j}(\tfrac{2i+3}{2^{j+1}})P_{2i-1,i}^{j} + \varphi_{2i}^{j}(\tfrac{2i+3}{2^{j+1}})P_{2i,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i+5}{2^{j+1}}) &= \varphi_{2i-2}^{j}(\tfrac{2i+5}{2^{j+1}})P_{2i-2,i}^{j} + \varphi_{2i-1}^{j}(\tfrac{2i+5}{2^{j+1}})P_{2i-1,i}^{j}\\
&\quad + \varphi_{2i}^{j}(\tfrac{2i+5}{2^{j+1}})P_{2i,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i+7}{2^{j+1}}) &= \varphi_{2i-1}^{j}(\tfrac{2i+7}{2^{j+1}})P_{2i-1,i}^{j} + \varphi_{2i}^{j}(\tfrac{2i+7}{2^{j+1}})P_{2i,i}^{j}\\
\varphi_{i}^{j-1}(\tfrac{2i+9}{2^{j+1}}) &= \varphi_{2i}^{j}(\tfrac{2i+9}{2^{j+1}})P_{2i,i}^{j}
\end{aligned}$$

This equation system are overdetermined and only five values $P_{2i-4,i}^{j}$, $\cdots$, $P_{2i,i}^{j}$ are unknown, so equation system can be solved by arbitrary five equations. For convenience, the first three and last two equations are selected.

$$\begin{aligned}
P_{2i-4,i}^{j} &= \varphi_{i}^{j-1}(\tfrac{2i-5}{2^{j+1}}) / \varphi_{2i-4}^{j}(\tfrac{2i-5}{2^{j+1}}) = \varphi(\tfrac{1}{4}) / \varphi(\tfrac{1}{2}) = \tfrac{1}{8}\\
P_{2i-3,i}^{j} &= [\varphi_{i}^{j-1}(\tfrac{2i-3}{2^{j+1}}) - \varphi_{2i-4}^{j}(\tfrac{2i-3}{2^{j+1}})P_{2i-4,i}^{j}] / \varphi_{2i-3}^{j}(\tfrac{2i-3}{2^{j+1}})\\
&= [\varphi(\tfrac{3}{4}) - \tfrac{1}{8}\varphi(\tfrac{3}{2})] / \varphi(\tfrac{1}{2}) = \tfrac{1}{2}\\
P_{2i-2,i}^{j} &= [\varphi_{i}^{j-1}(\tfrac{2i-1}{2^{j+1}}) - \varphi_{2i-4}^{j}(\tfrac{2i-1}{2^{j+1}})P_{2i-4,i}^{j} -\\
&\quad \varphi_{2i-3}^{j}(\tfrac{2i-1}{2^{j+1}})P_{2i-3,i}^{j}] / \varphi_{2i-2}^{j}(\tfrac{2i-1}{2^{j+1}})\\
&= [\varphi(\tfrac{5}{4}) - \tfrac{1}{8}\varphi(\tfrac{5}{2})P_{2i-4,i}^{j} - \tfrac{1}{2}\varphi(\tfrac{3}{2})] / \varphi(\tfrac{1}{2}) = \tfrac{3}{4}
\end{aligned}$$

According to symmetry, $P_{2i-1,i}^{j} = \frac{1}{2}$, $P_{2i,i}^{j} = \frac{1}{8}$ 。

Reconstruction matrix P_j can be obtained by above algorithm to different B-Spline basis functions $\varphi_{l}^{j-1}(t)$ (i ∈ [1,2^{j-1}+3]). Since knot vector is quasi-uniform, $P_{l,i}^{j}$ at end points must be recalculated with the same principle. The answer of P_j is as follows:

$$P_j = \frac{1}{16}\begin{bmatrix}
16 & & & & & & & & \\
8 & 8 & & & & & & & \\
 & 12 & 4 & & & & & & \\
 & 3 & 11 & 2 & & & & & \\
 & & 8 & 8 & & & & & \\
 & & 2 & 12 & & & & & \\
 & & & 8 & \vdots & & & & \\
 & & & 2 & \vdots & 2 & & & \\
 & & & & \vdots & 8 & & & \\
 & & & & & 12 & 2 & & \\
 & & & & & 8 & 8 & & \\
 & & & & & 2 & 11 & 3 & \\
 & & & & & & 4 & 12 & \\
 & & & & & & & 8 & 8 \\
 & & & & & & & & 16
\end{bmatrix}$$

Compared with literature [3] and [4], the answer is absolutely correct.

5 Examples

According to above described algorithm, an example of wavelet decomposition and reconstruction is provided. As shown in Figure 5, curve (a) is a complicated curve with 131 control points at scale j =7, after two times wavelet analysis, curve with 67 control points at scale j =6 and curve with 35 control points at scale j =5 are obtained, which are shown in Figure 5(b) and Figure 5(c). Curve (d) is the lost detail when curve (a) is decomposed to curve (b) and Curve (e) is the lost detail when curve (b) is decomposed to curve (c), too.

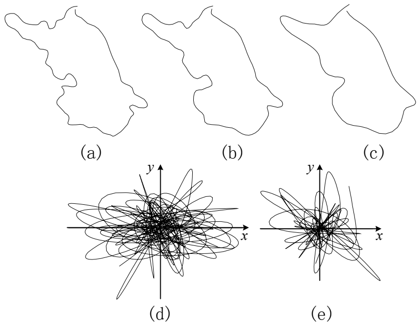


Figure 5 Wavelet decomposition and reconstruction of complicated curve

In order to simplify the drawing of detail curves, which are mapped to basis ϕ_j. Substituting Eq.（9）into Eq.（6）gives

$$g_{j-1} = \psi_{j-1}D_{j-1} = \phi_j Q_j D_{j-1}\text{。}$$

$g_{j\text{-}1}$ can be regarded as curves with control points $Q_jD_{j\text{-}1}$ at basis ϕ_j.

In order to describe detail curves clearly, curve (d) is magnified 80 times and curve (e) is magnified 20 times.

Analyzing curve (d) and curve (e) gives the conclusion that detail curves are controlled by the points around the coordinate original point and are independent on the location of curves. Detail curves are invariable if curves themselves are kept in shape, whether or not curves are translated or rotated. That is to say, detail curves really describe the differencs alone curves at different resolution and extract the lost informations in the process of curve decomposition.

At the same time, curve (b) can be absolutely reconstructed by curve (c) and corresponding curve (e) and curve (a) can be absolutely reconstructed by curve (b) and corresponding curve (d).

6 Conclusion

Application of wavelet analysis technology on reverse engineering is studied in this paper. Because B-Spline basis function and corresponding wavelets are lake of translation orthogonality, rapid Mallat algorithm does not work for this kind of wavelets translation. On the basis of analysis of decomposition and reconstruction theory of B-Spline wavelets, inherent relations at different scale of B-Spline basis functions are researched thoroughly and the solution process of reconstruction matrix P_j is described in details. This algorithm is clear, simple and robust. Although this algorithm stems from intuitionistic geometric meanings and is deficient in rigorous mathematics derivation, the thinking is reasonable and the results are absolutely right. At last, an example of decomposition and reconstruction of complicated curve with this algorithm is provided. By tensor product operation, this algorithm can be applied to surface fairing easily.

References

[1] Quak E, Weyrich N, “Decomposition and reconstruction algorithms for spline wavelets on a bounded interval”, Applied and Computational Harmonic Analysis, Vol.1, No.3, 1994, pp.217~231

[2] Adam Finkelstein, David H Salesin, “Multiresolution curves [A]”, In: Computer Graphics Proceedings, Annual Conference Series, ACM SIGGRAPH, Orlando, Florida, 1994, pp.261~268

[3] Eric J Stollnitz, Tony D DeRose, David H Salesin, “Wavelets for computer graphics: A primer, Part2”, IEEE Computer Graphics and Applications, Vol. 15, No. 4, 1995, pp.75~84

[4] Zhu Xinxiong, Modeling of freeform curves and surfaces, Beijing: Science Press, 2000

[5] Chui C K, Cheng Zhengxing, An introduction to wavelets, Xi’an: Xi’an Jiaotong University Press, 2000

[6] Ji Xiaogang, Gong Guangrong, “Semi-orthogonal B-spline wavelets and its application to curve and surface fairing”, Chinese Journal of Mechanical Engineering, Vol. 42, No. 9, 2006, pp.54~59

[7] Ji Xiaogang, Gong Guangrong, “Curve Fairing With Arbitrary Number of Control Vertices By Semi- orthogonal B-spline Wavelets”, Journal of Engineering Graphics, Vol. 27, No. 2, 2006, pp.90~95

[8] Elber G, Gotsman C. “Multiresolution control for nonuniform B-spline curve editing”, In: Proceedings of the 3rd Pacific Conference on Computer Graphics and Applications 1995, IEEE Computer Society Los Alamitos,1995, pp.267~278

[9] Sun Y K, Zhu X X, “Wavelet-Based fairing of B-spline surfaces”, Chinese Journal of Aeronautics, Vol 12, No. 3, 1999, pp.176~182

[10] Zhao G, Xu S H, Li W S, Teo O E, “Fast variational design of multiresolution curves and surfaces with B-spline wavelets”, Computer-Aided Design, Vol. 37, 2005, pp.73～82

Medical Image Registration Using Particle Swarm Optimization Algorithm with Mutation Operation

Zhijun Zhu Wenbo Xu Jun Sun

School of Information Technology , Jiangnan University Wuxi , Jiangsu 214122, China
Email:zhuzhijun123@163.com

Abstract

Image registration based on mutual information is of high accuracy and robustness. Hence, it has received much attention these years. Unfortunately, the mutual information function is generally not a smooth function but one containing many local maxima, which has a large influence on optimization. PSO has fewer parameters to control and can find the best solution quickly and guarantee to be global convergent. This paper proposes a registration method based on wavelet transform. In this method the mutual information is used as the similarity measure and a hybrid algorithm combined by PSO algorithm and a mutation the search technique. This method is applied to the 2D registration of MRI. Experiment results show that this registration method could efficiently restrain local maxima of mutual information function and not only it can improve accuracy and speed but also the subvoxel accuracy can be achieved.

Keywords: Image registration; Mutual information; Particle Swarm Optimization Algorithm; Mutation Operation

1 Introduction

Medical imaging provides insights into the size, shape, and spatial relationships among anatomical structures. For instance, computer-assisted tomography (CT) is very useful for imaging bony structures and dense tissue, whereas magnetic resonance imaging (MRI) and ultrasound (US) provide views of soft tissues. Additionally, functional imaging is becoming increasingly important both clinically and in medical research. For example, positron emission tomography (PET) and single-photon computed tomography (SPECT) imaging provide information on blood flow and metabolic processes. Very often, areas of the body are imaged with different modalities. These images are used in a complimentary manner to gain additional insights into a phenomenon.

For these different modalities to be useful, they must be appropriately combined, or *fused*. Before images can be fused, they must first be geometrically and/or temporally aligned. This alignment process is known as *registration*. One of the most important applications of registration is image-guided therapy, and registration in neurosurgery and orthopaedic surgery is now common. Registration is also used in treatment planning, in functional brain imaging and in brain atlases and mapping. Developing applications include multimedia patient records, postgenomic registration to characterize gene function, registration of intra and preoperative images in surgical interventions and treatment monitoring.

Multimodal image registration plays an increasingly important role in medical imaging. Its objective is to find a transformation that maps two or more images, acquired using different imaging modalities, by optimizing a certain similarity measure. Among the different similarity measures that have been proposed, mutual information (MI) and normalized mutual information (NMI) are the most commonly used since they produce satisfactory results in terms of accuracy, robustness and reliability. However, MI-based methods are very sensitive to implementation decisions, such as interpolation and optimization methods and multiresolution strategies [4][5][6].

This paper proposes a multiresolution registration method based on wavelet transform. In this method the mutual information is used as the similarity measure and PSO algorithm with mutation operation as the search technique, it can improve accuracy and speed. Experiments shows that this registration method could efficiently restrain local maxima of mutual information function and it can improve accuracy.

2 Mutual Information

Similarity metrics for image registration must be robust [7][8]; that is, they should attain a global maximum (or a very distinct local maximum) at the correct registration. The “best registration” is very often a local (not global) optimum. Thus, in addition to exercising care when selecting an initial orientation, other features, such as intensity gradient information, should also be utilized. However, in this paper, it is assumed that the global optimum is attained at the correct registration transformation.

Much of the current work on biomedical image registration utilizes information theoretic voxel similarity measures, in particular,mutual information based on the Shannon definition of entropy. Mutual information has been shown to be robust for multimodal registration [1][2], and does not depend on the specific dynamic range or intensity scaling of the images. It is a measure of the relative independence of two images. High values indicate high dependence. The mutual information of two images A and B is given as:

$$I(A,B) = H(A) + H(B) - H(A,B) \tag{1}$$

where H denotes Shannon entropy. For an image A with N pixels and with each intensity value $i = 1, \ldots N$, occurring with frequency p_i, the Shannon entropy is computed as:

$$H(A) = -\frac{1}{N}\sum_{i=1}^{N} p_i \log p_i \tag{2}$$

Where, the p_i is the density of pixel $H(A,B)$ is computed with an estimate of the joint density.

$$I_N(A,B) = \frac{H(A) + H(B)}{2H(A,B)} \tag{3}$$

3 Wipe Off The Background

For the sake of prevention the disturb of no use information of the image, wo must wipe off the background of the image. The method of wiping off the background is the method was proposed by rui wan [8],step as follow:

(1) Find out the maximum and minimum density of the image Z_1 and Z_k, Let original threshold value is:

$$T_0 = \frac{Z_1 + Z_k}{2} \tag{4}$$

(2) Use threshold value T_k, divide image into two part R_1 and R_2, and find out the average density of the two part separately, Z_0 and Z_B:

$$Z_0 = \frac{\sum_{z(i,j)<T_k} z(i,j) \times N(i,j)}{\sum_{z(i,j)<T_k} N(i,j)} \tag{5}$$

$$Z_B = \frac{\sum_{z(i,j)>T_k} z(i,j) \times N(i,j)}{\sum_{z(i,j)>T_k} N(i,j)} \tag{6}$$

Where, $z(i,j)$ is the density of the image point (i,j), $N(i,j)$ is the coefficient of the point (i,j) ,where $N(i,j) = 1.0$.

（3）Find out the new threshold value:

$$T_{k+1} = \frac{Z_0 + Z_B}{2} \tag{7}$$

if $T_k = T_{k+1}$,process end, otherwise $k = k+1$, perform the previous step repeatly.

Finally, if the density of the image less than the threshold value, we make use of the seed fill method to fill the image and the density sets zero, wipe off the background.

4 Interpolation

Mutual information(MI) registration algorithms involve iteratively transforming image A with respect to image B while optimizing MI measure which is calculated from corresponding voxel values. While all samples are taken at grid points of the floating image A, Their transformed position will in general not coincide with a grid point of the reference image B so interpolation of the reference image is needed to obtain the image intensity value at this point. Three commonly used interpolation methods are Nearest Neighbor Interpolation, Trilinear Interpolation and Trilinear Partial Volume Interpolation. we use the method of Partial Volume Interpolation in image registration.

5 Pso Algorithm With Mutation Operation

We proposed Particle Swarm Optimization (PSO) [1][2]algorithm with mutation operation that outperforms traditional PSOs in search ability as well as having less parameter to control.

Each particle contains some information, as follow:

(1) $x_i = (x_{i1}, x_{i2}, \cdots x_{id})$:the current position of the particle;

(2) $v_i = (v_{i1}, v_{i2}, \cdots v_{id})$:the current velocity of the particle;

(3) $P_i = (P_{i1}, P_{i2}, \cdots P_{id})$:the best fitness value of the particle i , pbest;

(4) $P_g = (P_{g1}, P_{g2}, \cdots P_{gd})$: the best fitness value of the particle swarm, gbest。

The particles move according to the following equation:

$$\begin{aligned} v_i(t+1) &= \omega v_i(t) + c_1 * rand1() * (P_{id}(t) - X_i(t)) \\ &+ c_2 * rand2() * (P_{g,d}(t) - X_i(t)) \end{aligned} \tag{8}$$

$$X_i(t+1) = X_i(t) + V_i(t+1) \tag{9}$$

where ω is the weight factor, c_1 and c_2 are the acceleration constant, $rand1()$ and $rand2()$ v are the uniform random value in the range [0,1], $v_i(t)$ is the velocity of particle i at iteration t, $X_i(t)$ is the current position of particle i at iteration t.

The weight factor ω provides a balance between global and local explorations. The constants c_1 and c_2 represent the weighting of the stochastic acceleration terms that pull each particle toward the pbest and gbest position. Low values allow particles to roam far from the target regions before being tugged back. On the other hand, high values result in abrupt movement toward, or past, target regions .

Here, we also introduce a mutation operation exerted on gbest position to enhance the global convergence ability of the PSO algorithm. The process of PSO with mutation operation (PSO-MO) is outlined as follows.

```
Process of the PSO image registration:
For each particle
    Initialize population:(the value of each particle, the best fitness value of local, the best fitness value of global)
End
Do
    For each particle
        Use formula(2) to work out the value of mutual information and the best fitness value is the value of mutual information;
    If  the fitness value of current particle more than the best fitness value of particle(pbest)
    The fitness value of current particle replace the best fitness value of particle(pbest)
    End
    The best fitness particle is selected the maximum value from the particles, let this particle is the global best fitness value (gbest) .
    Exert a standard normal distribution on gbest.
    For each particle
        Use formula (8) to work out the component of velocity;
        Use formula (9) to update the best value of the particle;
    End
While (satisfy the condition)loop:
Not excess the maximum iteration (the maximum iteration is appointed by the user);
```

6 Experiments

Choose one MRI image of head (256×256),wipe off the background of the image. The image of wiping off the background rotates 30°,then down translation, rightward translation five pixels. The reference image is the image of wiping off the background, the floating image is the image of after transform. Finally, image registration wil be completed in four algorithm, and each algorithm will be ran ten. Image registration result and runtime is shown in table 1.

Table 1 MRI image registration result and runtime

Algorithm	$\Delta\theta$ RMS	Δt_x RMS	Δt_y RMS	T MRT
PSO	0.1025	0.5133	0.4755	17268
PSO-MO	0.0050	0.1115	0.1699	16925

Figure 1 is the reference image, Figure 2 is the reference image of wiping off the background, Figure3 is the floating image. Where, $\Delta\theta$ 、 Δt_x and Δt_y represent the error of rotation（measured in degrees）, the error of translation of X axis and Y axis（measured in pixels）separately, T represent the runtime of algorithms（measured in seconds）,RMS represent the average value of data, MRT represent the average value of runtime. In experiments, the iteration of PSO-MO is 300. Experiments depend on Matlab6.5.The computer is Pentium4,2.0MHz, EMS memory is 256MB。

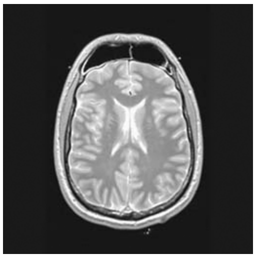


Figure 1 reference image

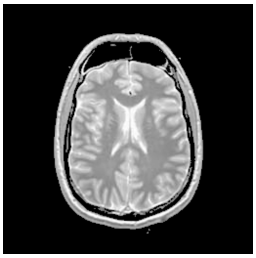


Figure 2 reference image of wiping off the background

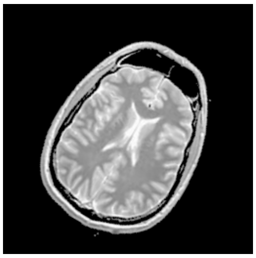


Figure 3 floating image

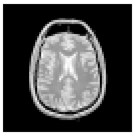


Figure 4 Two-stage WT of the reference image

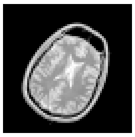


Figure 5 Two-stage WT of the floating image

The original parameter of PSO and PSO-MO is random. Error rate of registration using this method show in table 2.

Table 2 Error rate of registration

Algorithm	Error rate of registration（%）
PSO	20
PSO-MO	0

From the table 1 and table 2,we can find that the hybrid algorithm combined by PSO-MO algorithm is no error registration all.

7 Conclusion

The MI registration criterion presented in this paper allows for highly accurate, highly robust, and completely automatic registration of multimodality medical images. Because the method is largely data independent and requires no user interaction or preprocessing, the method is well suited to be used in clinical practice. Unfortunately, the mutual information function is generally not a smooth function but one containing many local maxima, which has a large influence on optimization.

This paper proposes a registration method based on wavelet transform. In this method the mutual information is used as the similarity measure and a PSO algorithm with mutation operation as the search technique. This method is applied to the 2D registration of MRI. Experiments shows that this registration method could efficiently restrain local maxima of mutual information function and it can improve accuracy and speed .Furthermore, this method can be used in mutimodality image registration, and three-dimension image registration.

References

[1] F .Maes, D .Vandermeulen, and P .Suetens, "Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information",*Medical Image Analysis*, 1999, 3(4), pp.373–386.

[2] J.P.W Pluim, J.B.A Maintz, and Viergever "M A,Mutual information based registration of medical images: A survey", *IEEE Trans. on Medical Imaging*, 2003, 22(8), pp.986–1004.

[3] Ximiao Cao, and Qiuqi Ruan, "A Survey on Evaluation Methods for Medical Image Registration. Complex Medical Engineering",2007.CME2007.IEEE/ICME International Conference May 2007, pp.718 – 721.

[4] Y. Yamamura, K. Hyoungseop, and A. Yamamoto, "A Method for Image Registration by Maximization of Mutual Information", SICE-ICASE, 2006. International Joint Conference Oct. 2006, pp.1469-1472.

[5] L.Seok, C.Minseok, K.Hyungmin, and F.C.Park, "Geometric Direct Search Algorithms for Image Registration",Image Processing,IEEE Transactions Volume 16(9) Sept. 2007,pp.2215 – 2224.

[6] C .Studholme, D.L.G .Hill, and D.J .Hawkes, "An overlap invariant entropy measure of 3D medical image alignment" *Pattern Recognition*, 1999, 32(1),pp. 71–86.

[7] F .Maes, A .Collignon, and D .Vandermeulen, *et al*,"Multimodality image registration by maximization of mutual information", *IEEE Trans. on Medical Imaging*, 1997, 16(2), pp.187–198.

[8] Rui Wan, Jiong Mei, and Minglu Li, "Medical image registration based on Maximization of Mutual Information and Correlation", Journal of Image and Graphics of china, 2003, 8（A）, spec, pp.834-838.

Wavelet Analysis and Processing Algorithm for Images Compressing and Data Fusion

Yulin Zhang[1] Baoguo Xu[1] Xia Zhu[2] Wenbo Xu[2]

1 School of Communication and Control Engineering, Jiangnan University, Wuxi, JiangSu, 214122, China

Email:zhangyulin@mail.hyit.edu.cn

2 School of Information Technology, Jiangnan University, Wuxi, JiangSu, 214122,China

Email:thanks2024@hyit.edu.cn

Abstract

Wavelet analysis, the newest time-frequency analysis tool with its unique advantages, was widely used in engineering. However, in the practical application, compressing images contain huge data, which made it difficult and impossible to process the image by the ordinary method. In this paper, an image compression and data fusion method is put forward based on wavelet transform. We analyze land surveying image “soil” by using two-dimensional wavelet and achieve good results.

Keywords：wavelet analysis; images compressing; data fusion; decomposition; reconstruction

1 Introduction

Everywhere around us are signals that can be analyzed. For example, there are seismic tremors, human speech, engine vibrations, medical images, financial data, music, and many other types of signals. Wavelet analysis is a new and promising set of tools and techniques for analyzing these signals.

In the late 1980s, Wavelet transform developed into the branch of the mathematical application. It is an analysis method of time-frequency and has the character of multi-resolution analysis. At present, the wavelet has been widely applied in many fields.[1] For example, J.Morlet used it into seismic wavelet analysis and signal processing; M.Frisch detected the unknown transient noise by the wavelet transform; P. Dutilleux deals with the languages by signal analysis, transformation and integration. H.Kim uses it in the time-frequency analysis.

Information fusion technology is a method which is to obtain, transfer and deal with varieties of information .Humans can obtain information through a variety of sensors , accurately identify the state of the environment or objects, and guide them to the next step .Usually ,data are collected from machinery or equipment ,the original signal will need huge storage capacity .In order to reduce the storage ,a special way is brought forward to memory some best data ,which need smallest memory, and we can restore the original signal from it.

In this paper, a wavelet-based image compression and data fusion [2] method is presented. According to a special principle, in order to achieve less amount of data memory, some the wavelet coefficients, which has small or no contribution, will be removed. And it only memories some coefficients that can contribute to other wavelet coefficients, that is, it is data compression.

2 Wavelet Transform

The continuous wavelet transform (CWT) is defined as the sum over all time of the signal multiplied by scaled, shifted versions of the wavelet function $\psi(t)$:

$$(W_{\psi} f)(a,b) = < f, \psi_{a,b} > = |a|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} f(t)\bar{\psi}(\frac{t-b}{a})dt$$ if it satisfies $f \in L^2(R)$ .The value of $\psi(t)$ depends on two

indices ‘a’ and ‘b’ with $\psi_{a,b}(t)=|a|^{-\frac{1}{2}}\psi(\frac{t-b}{a})$. The index ’a’ stands for the parameters of frequency character , indexed by $a\in R-\{0\}$, the index ‘b’ stands for parameters of time or space , indexed by $b\in R$.

Clearly, we may have noticed that wavelet analysis does not only use a time-frequency region, but a time-scale region. The fluctuation of index’ a’ changes not only the structure of the spectrum, but also its window size and shape. With the reduction of$|a|$, the spectrum of $\psi_{a,b}(t)$ moves to the part of high-frequency, and the width of $\psi_{a,b}(t)$ is narrower. And that just meets the request of signal frequency. The function $f(t)$ is said to be wavelet anti-transform if and only if its Fourier transform satisfies $C_\psi=\int_R|\omega|^{-1}|\psi^\wedge(\omega)|^2 d\omega<\infty$. The classical norm of $f(t)$ is given by $f(t)=\frac{1}{C_\psi}\int_{-\infty}^{+\infty}\left\{\int_{-\infty}^{+\infty}(W_\psi f)(a,b)\left[\frac{1}{\sqrt{a}}\psi(\frac{t-b}{a})\right]db\right\}\frac{da}{a^2}$.

For continuous wavelet transform, the indices ‘a’、‘t’ and ‘b’ are continuous. To calculate them by computer, we must deal with them by DWT. Normally, the indices ‘a’ and ‘b ‘are expressed in the form of exponential. That is: $a=2^j$, $b=2^jk$ while $j,k\in Z$ .The corresponding discrete wavelet is: $\psi_{j,k}(t)=a_0^{-\frac{j}{2}}\psi(a_0^{-j}t-kb_0)$. The indexes $C_{j,k}=\int_{-\infty}^{+\infty}f(t)\bar{\psi}(t)dt$ are called coefficients of$f(t)$. The small-scale changes of index ‘j’ will cause a lot of changes of index ‘a’. We usually denote$\mathrm{a}_0=2$,$b_0=1$and change ‘a’ and ‘b’ into $a=2^j$ and $b=2^jk$ .Then we can achieve dyadic wavelet:$\psi_{j,k}=2^{-\frac{j}{2}}\psi(2^{-j}t-k)$.

3 Algorithm Analysis

From an intuitive point of view, the wavelet decomposition consists of calculating ‘a’ "resemblance index" between the signal and the wavelet located at position ‘b’. If the index is large, the resemblance is strong, otherwise it is slight. Since the introduction of multi-differentiate, S. Mallat found a fast algorithm -Mallat algorithm[3-4]. Mallat devotes the signal decomposition and reconstruction algorithm. The notion behind compression is based on the concept that the regular signal component can be accurately approximated using the following elements: a small number of approximation coefficients (at a suitably chosen level) and some of the detail coefficients.

Scaling function $\varphi(x)$:

$$\varphi(x)=\sum_{k\in Z}a_k\varphi(2x-k)\,.$$

Wavelet function $\psi(t)$:

$$\psi(x)=\sum_{k\in Z}(-1)^k a_{1-k}\varphi(2x-k)\,.$$

Then two of the above can be rewritten as following based on band-pass filter $G(x)=e^{j\omega}\bar{H}(\omega+\pi)$:

$$\begin{cases}\varphi(x)=\sqrt{2}\sum_{k\in Z}h_k\varphi(2x-k)\\ \psi(x)=\sqrt{2}\sum_{k\in Z}g_k\varphi(2x-k)\end{cases}\tag{1}$$

Where $h_k=\frac{a_k}{\sqrt{2}}$, $g_k=(-1)^k\frac{a_{1-k}}{\sqrt{2}}$.

Then Eq.（1）can be changed into:

$$\begin{cases}\varphi_{j,k}=\sum_{n\in Z}h_k\varphi_{j+1,2k+n}\\ \psi_{j,k}=\sum_{n\in Z}g_k\varphi_{j+1,2k+n}\end{cases}\tag{2}$$

While $c_{j,k}=(f(x),\varphi_{j,k}(x))$ $d_{j,k}=(f(x),\psi_{j,k}(x))$, we can denote $C_{j,k}$ by using Eq.（2）:

$$\begin{aligned}c_{j,k}&=(f(x),\varphi_{j,k}(x))\\&=(f,\sum_{n\in Z}h_n\varphi_{j+1,2k+n})\\&=\sum_{n\in Z}\bar{h}_n(f,\varphi_{j+1,2k+n})\\&=\sum_{n\in Z}\bar{h}_{n-2k}\,c_{j+1,n}\;.\end{aligned}$$

And $d_{j,k}=\sum_{n\in Z}\bar{g}_{n-2k}\,c_{j+1,n}$.

In the end, we can get the following:

$$\begin{cases}c_{j,k}=\sum_{n\in Z}h_{n-2k}c_{j+1,n}\\ d_{j,k}=\sum_{n\in Z}g_{n-2k}c_{j+1,n}\end{cases}\tag{3}$$

We can express the above algorithm with infinite matrix:

Let $H_{k,n}=\bar{h}_{n-2k}$, $G_{k,n}=\bar{g}_{n-2k}$, $H=(H_{k,n})$, $G=(G_{k,n})$. $C_j=(c_{j,n})$, $D_j=(d_{j,n})$ .The Eq.（3）can be turned into :

$$\begin{cases} C_j = HC_{j+1} \\ D_j = GC_{j+1} \end{cases} \text{with j=J-1，J-2，，J-M} \qquad （4）$$

$$\begin{tikzcd}
C_J \arrow[r] \arrow[dr] & C_{J-1} \arrow[r] \arrow[dr] & C_{J-2} \arrow[r] \arrow[dr] & \bullet\bullet\bullet\bullet\bullet \arrow[r] \arrow[dr] & C_{J-M} \\
& D_{J-1} & D_{J-2} & \bullet\bullet\bullet\bullet\bullet & D_{J-M}
\end{tikzcd}$$

Figure 1　The decomposition algorithm

In signal processing, $f(t)$ is often expressed by V_j (if it contains j levels).

Because $V_{j+1}=V_j\oplus W_j$, then

$$\varphi_{j+1,k}(x)=\sum_{l\in Z}\alpha_l\varphi_{j,l}(x)+\sum_{l\in Z}\beta_l\psi_{j,l}(x) \qquad （5）$$

While $\{\varphi_{j+1,k}(x)\}_{k\in Z}$ is norms orthogonal basis, we can get the following equations through Eq.（2）:

$$\alpha_l=(\varphi_{j+1,k},\varphi_{j,l})$$

$$=(\sum_{k\in Z}\alpha_l\varphi_{j,l},\varphi_{j,l})=\int_R\varphi_{j+1,k}(x)\bar{\varphi}_{j,l}(x)dx$$

$$=\int_R\varphi_{j+1,k}(x)(\sum_{n\in Z}\bar{h}_n\bar{\varphi}_{j+1,2l+n}(x))dx=\bar{h}_{k-2l}.$$

And $\beta_l=\bar{g}_{k-2l}$, at last we can get:

$$\varphi_{j+1,k}=\sum_{l\in Z}\bar{h}_{k-2l}\varphi_{j,l}(x)+\sum_{l\in Z}\bar{g}_{k-2l}\psi_{j,l}(x)$$ Scalar product of $f(x)$ and $\varphi(x)$:

$$c_{j+1,k}=(f(x),\varphi_{j+1,k}(x))=\sum_{l\in Z}\bar{h}_{k-2l}c_{j,l}+\sum_{l\in Z}\bar{g}_{k-2l}d_{j,l}$$

That is:

$$c_{j+1,k}=\sum_{l\in Z}\bar{h}_{k-2l}c_{j,l}+\sum_{l\in Z}\bar{g}_{k-2l}d_{j,l} \qquad （6）$$

We also can express Eq.（6）with infinite matrix order:

$C_{j+1}=H^*C_j+G^*D_j$ With j=J-M, J-M+1….J.

While H^*,G^* are conjugated matrixes of H and G.

$$\begin{tikzcd}
C_{J-M} \arrow[r] & C_{J-M+1} \arrow[r] & \bullet\bullet\bullet\bullet\bullet \arrow[r] & C_{J-1} \arrow[r] & C_J \\
D_{J-M} \arrow[ur] & D_{J-M+1} \arrow[ur] & \bullet\bullet\bullet\bullet\bullet \arrow[ur] & D_{J-1} \arrow[ur] &
\end{tikzcd}$$

Figure 2　The reconstruction algorithm

4　Application

With a common sense that the image compressing must maintain the characteristics of original image, we can decide a global threshold value and compressing the high-frequency section or select separately thresholds to compress in every scale [5-10]. However, in order to get the approached anticipated result, the processing is slow and should be repeated many times.

4.1　The process of the wavelet analysis –image compression application

The compression procedure contains four steps:

Step one: Using two-dimensional Mallat decom position algorithm decomposes the original image. Assumption that we decompose the image into J layer, we will get three times high frequency and one time low frequency.

Step two: Quantize the coefficient of the wavelet transform. Quantize details of the scale wavelet coeffici ents. For each level, a threshold is selected and hard threshold is applied to the detail coefficients. We can quantize 1 ~ J layers by using the same or different threshold.

Step three: Symbol-stream will be changed into bit-stream to achieve the purpose of data compression.

Step four: Store or transmit the bit-stream. Compute wavelet reconstruction using the original approximation coefficients of level J and the modified detail coefficients of levels from 1 to J.

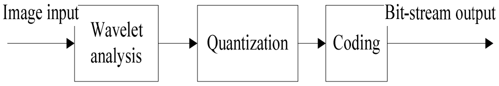


Figure 3　the flow chart of the compression

4.2 The realization of Image Compression

Let us give an example. We analysis the problem on the assumption that J=2 .The numerical form of image is $\left\{C^{0}{}_{k,m}\right\}=\left\{f(k,m)\right\}$. We can get the wavelet analysis from the Figure 4. This kind of algorithm leads to a decomposition of approximation coefficients at level J in four components: the approximation at level J + 1, and the details in three orientations (horizontal, vertical, and diagonal).That is $C^{j}{}_{k,m}, \alpha^{j}{}_{k,m}, \beta^{j}{}_{k,m}, \gamma^{j}{}_{k,m}$ where j=1, 2,3. Because the energy distribution in the "approximate" and the "details" is uneven, energy in the "details" are rarely. We can combine this character with human’s visual identification, which is "from coarseness to fine”. The wavelet transform of image is stored at the binary region. Then we can code and scan the image along with the painted order of Figure 4.The purpose in doing so is to outline image’s rough, and then gradually add the precise details to it. When it comes to large level, one part of the acceptance of images can intellectually tell if it has met the accuracy. If it is true, it will stop in time and get rid some of the details in order to accelerate the transmission speed.

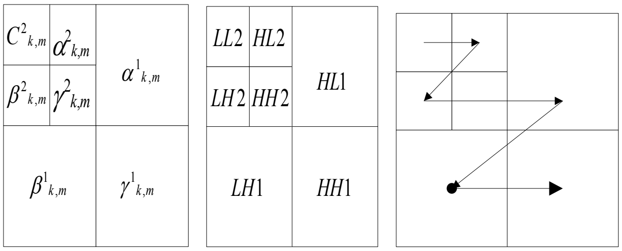


Figure 4 the analysis of wavelet

Figure 5 is the original image .Figure 6 to Fig 7 is decomposed images by the wavelet. Figure 6 is gotten from the first level; we can get the approximation section and detail section of the first level. Figure 7 is gotten from the second level; we can get the approximation section and detail section of the second level. Figure 8 is the reconstruction of Figure 6 and Figure 7.Comparing each pair of the two sets of the data form the two images, and we can analyze the first and the second data separately by the “energy function” of the wavelet.

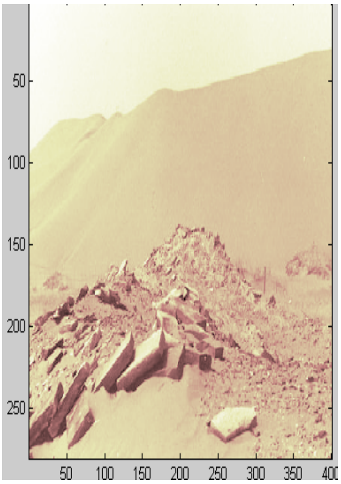


Figure 5 the original image

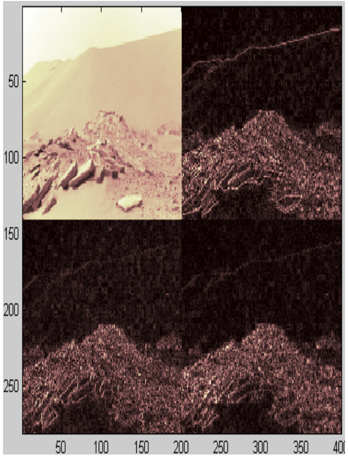


Figure 6 decomposed images by the wavelet

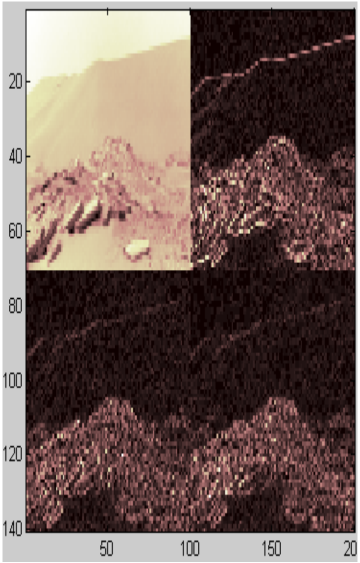


Figure 7 decomposed images by the wavelet

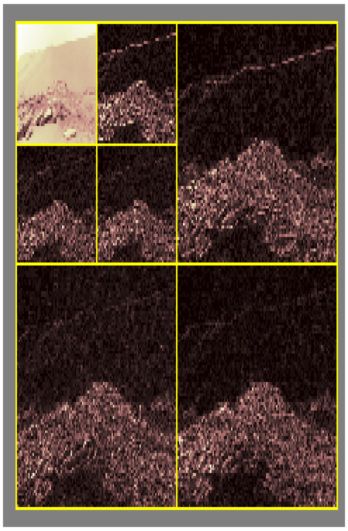


Figure 8 the reconstruction images

The results showed that, the reconstruction image is made up of the different characteristics of the original image. In the first compression (Figure 6), we can draw the low-frequency information of the first level .The effect is good at this time; compression ratio is about 1/3. In the second compression (Figure 7), we can extract the low-frequency information of first low-frequency part. The compression ratio is about 1/12. Figure 8 maintains as much as possible information of the original. Therefore, this method is suitable to Non-real-time occasions, such as multi-spectral image recognition, medical imaging, and ground investigation and so on.

References

[1] Stephane Mallat, A Wavelet Tour of Signal Processing. San Diego: Academic Press, 1998

[2] D. A. Yocky, “Image merging and data fusion by means of the discrete two-dimensional wavelet transform,” *J. Opt. Soc. Amer. A*,vol. 12, no9, pp. 1834–1841, 1995

[3] Xavier Otazu, Albert Prades, “Multiresolution-Based Image Fusion with Additive Wavelet Decomposition”, proceeding of the IEEE, vol.85.no 1, pp.1207-1211, 1997

[4] Mallat, S.G., 1989. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11, 674–693

[5] Whitcher, B.J., Guttorp, P., Percival, D.B., 2000. Wavelet analysis of covariance with application to atmospheric time series.Journal of Geophysical Research—Atmospheres 105 (D11), 14941–14962.R.M. Lark et al. / Journal of Hydrology290 272 (2003) 276–290

[6] Odeh, I.O.A., McBratney, A.B., 2000. Using AVHRR images for spatial prediction of clay content in the lower Namoi valley of eastern Australia. Geoderma 97, 237–254

[7] Lark, R.M., Webster, R., 1999. Analysis and elucidation of soil variation using wavelets. European Journal of Soil Science 50,185–206

[8] Lark, R.M., Webster, R., 2001. Changes in variance and correlation of soil properties with scale and location: analysis using an adapted maximal overlap discrete wavelet transform. European Journal of Soil Science 52, 547–562

[9] Polo A,CattaniF,VavassoriA,etal.MR and CT image fusion for postim plant analysis in permanent prostate seed implants[J].International Journal of Radiation Oncology Biology Physics,2004,60（5）:1572-1579

[10] Wu J,Huang H L,Tian J W,etal.Remote sensing image data fusion based on local deviation of wavelet packet transform [C].Proceedings of International Symposium on Autonomous Decen-tralized Systems,2005.372-377

Medical Image Segmentation Based on Gabor Filters and SOFM

Yao Wang[1] Wenbo Xu[2] Jun Sun[3]

1 School of Information Technology, Southern Yangtze University, Wuxi, 214122, China
Email: xiaoniu201@yahoo.com.cn

2 School of Information Technology, Southern Yangtze University, Wuxi, 214122, China
Email: xwb@sytu.edu.cn

3 School of Information Technology, Southern Yangtze University, Wuxi, 214122, China
Email: sunjun_wx@hotmail.com

Abstract

In this paper, presents a texture segmentation algorithm based on Gabor Filters and the Self-Organizing Feature Map. By filtering an input image with Gabor Filters tuned to the dominant frequency and orientation component of the textures, it is possible to locate each texture. The magnitude of the channel output should be large when the texture exhibits the frequency and orientation characteristics to which the channel's Gabor Filter is tuned, vice versa. There are weak intraclass dispersion and strong interclass dispersion in the filtered image, and the issue of texture segmentation is translated into that of traditional image segmentation. Finally, the pretreatment image data was clustered by combining Self-Organizing Feature Mapping network with clustering function.

Keywords: Gabor Filters; Self-Organizing Feature Map; medical image segmentation

1 Introduction

Medical image segmentation is very important for a medical diagnosis. But medical image segmentation techniques typically require some form of expert human supervision to provide accurate and consistent identification of anatomic structures of interest. Despite the great progress of medical image segmentation, it still lacks a portable and unified solution to address the diversity of textures. In accordance with the characteristics of medical images and medical observation, the method of interactive medical segmentation by combining Gabor Filters and neural network is stated.

In this paper, presents a texture segmentation algorithm based on Gabor Filters and the Self-Organizing Feature Map. By filtering an input image with Gabor Filters tuned to the dominant frequency and orientation component of the textures [1], it is possible to locate each texture. The magnitude of the channel output should be large when the texture exhibits the frequency and orientation characteristics to which the channel's Gabor Filter is tuned, vice versa. There are weak intraclass dispersion and strong interclass dispersion in the filtered image, and the issue of texture segmentation is translated into that of traditional image segmentation. Finally, the pretreatment image data was clustered by combining Self-Organizing Feature Mapping network with clustering function.

2 Gabor Filters

The Gabor Filters have received considerable attention because the characteristics of certain cells in the visual cortex of some mammals can be approximated by these filters. In addition these filters have been shown to posses optimal localization properties in both spatial and frequency domain and thus are well suited for texture segmentation problems [2, 3].

Gabor Filters have been used in many applications, such as texture segmentation, target detection, fractal dimension management, document analysis, edge detection, retina identification, image coding and image representation [4]. A Gabor Filter can be viewed as a sinusoidal plane of particular frequency and orientation, modulated by a Gaussian envelope. It can be written as:

$$h(x,y) = s(x,y)g(x,y) \qquad (1)$$

where $s(x,y)$ is a complex sinusoid, known as a carrier, and $g(x,y)$ is a 2-D Gaussian shaped function, known as an envelope. The complex sinusoid is defined as follows,

$$s(x,y) = e^{-2j\pi(u_0x+v_0y)} \qquad (2)$$

The 2-D Gaussian function is defined as follows,

$$g(x,y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2})} \qquad (3)$$

Thus the 2-D Gabor filter can be written as:

$$\begin{aligned} h(x,y) &= e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2})} e^{-2j\pi(u_0x+v_0y)} \\ &= g(x,y)e^{-2j\pi(u_0x+v_0y)} \end{aligned} \qquad (4)$$

The frequency response of the filter is:

$$\begin{aligned} & H(u,v) = G(u-u_0, v-v_0) \\ & \Rightarrow H(u,v) \\ & = 2\pi\sigma_x\sigma_y[e^{-2\pi^2[(u-u_0)^2\sigma_x^2+(v-v_0)^2\sigma_y^2]}] \\ & = \frac{1}{2\pi\sigma_u\sigma_v} e^{-\frac{1}{2}[\frac{(u-u_0)^2}{\sigma_u^2}+\frac{(v-v_0)^2}{\sigma_v^2}]} \end{aligned} \qquad (5)$$

where, $\sigma_u = \frac{1}{2\pi\sigma_x}, \sigma_v = \frac{1}{2\pi\sigma_y}$ This is equivalent to translating the Gaussian function by (u_0, v_0) in the frequency domain. Thus the Gabor function can be thought of as being a Gaussian function shifted in frequency to position (u_0, v_0) i.e at a distance of $\sqrt{u_0^2+v_0^2}$ from the origin and at an orientation of $\tan^{-1}\frac{u_0}{v_0}$. In the above 2 equations Eq.(4) and Eq.(5) , (u_0, v_0) are referred to as the Gabor Filter spatial control frequency. The parameters σ_x, σ_y are the standard deviation of the Gaussian envelope along *X* and *Y* directions and determine the filter bandwidth.

3 The Self-Organizin Feature Map

SOFM network is also called Kohonen feature map or topology saving map. It was brought forward by Finland scholar Kohonen in 1981. The establishment of the SOFM neural network model is stimulated from the modeling study of the biological system. There is a small extent area in the visualization layer of human brain, which responds to the external environment stimulation. According to such a feature of human brain, Kohonen built up the SOFM network model to stimulate the feedback feature of the brain visualization cells. He believed that the neighboring cells in the neural network interact and compete with each other, and finally self-adapt to the external environment to become the special detectors, which are capable of measuring different information [5]. It has powerful anti-interference ability.

This unsupervised neural network can learn to detect regularities and correlations in their input and adapt to their future responses to that input. SOFM network learns to classify input vectors according to how they are grouped in the input space by the competitive learning rule [6]. During being trained, SOFM network learns both the distribution and topology of the input vectors. The training result is that a neuron and its neighbors will be sensitive to a class.

The SOFM is composed of input layer and competitive layer. It is shown in Fig 1.

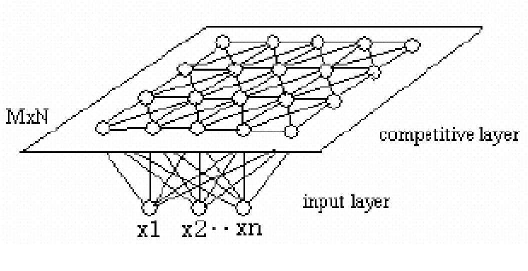


Figure 1 The Self-Organizing Feature Map

The input layer consisting of N neurons is a one-dimensional vector sequence. The nodes of the input layer have the same number as the dimension of the input space. The competition layer has m by n neurons that are arranged in a plane. The competitive layer is also the output layer. The different nodes in the output layer of the SOFM network represent the different classes after training. In this network, each element of the input vector P is connected to each neuron in the input layer through the weight matrix W. Another weight of each neuron can describe how close the neuron and the input vector are. The more similar they are, the smaller the distance is between them, and the easier the neuron will win in the competition.

The learning rule and working steps of SOFM[7, 8] as follows:

The Kohonen learning rule develops from the instar rule. For the instar model whose output is 0 or 1. The weight matrix is modified when output is 1, and then the Kohonen learning rule is found:

$$\Delta W_{ij} = lr \bullet (p_j - w_{ij}) \qquad (6)$$

Supposing that the input vector of the network is $P_k = (p_1^k, p_2^k, \ldots\ldots, p_n^k), k = 1,2,\ldots\ldots q$. The output vector of the competitive layer is $A_j = (a_{j1}, a_{j2}, \ldots\ldots, a_{jm})$, $j = 1,2,\ldots\ldots m$. Among them, P_k is a continuous vector, A_j is a numerical value. The weight matrix that connects the input neurons and the output neurons j is $W_j = (w_{j1}, w_{j2}, \ldots\ldots, w_{jN}), i = 1,2,\ldots\ldots N;\ j = 1,2,\ldots\ldots M$.

(1) Vector initialization. The linkage weight $\{W_{ij}\}$ is randomly assigned to a value within the range from 0.0 to 1.0, and the initial value of the learning rate $\eta(t)$ and neighborhood $N_g(t)$ are separately assigned to $\eta(0)(0 < \eta(0) < 1)$ and $N_g(0)$.

(2) Feed the network with an input vector P_k, and make them been normalized by Eq. (7).

$$\overline{P}_k = \frac{P_k}{\|P_k\|} = \frac{(p_1^k, p_2^k, \ldots\ldots p_n^k)}{[(p_1^k)^2 + (p_2^k)^2 + \ldots + (p_n^k)^2]^{1/2}} \qquad (7)$$

(3) Make the linkage weight vector been normalized by using Eq. (8), and then compute the distances d_j between the input vector and the linkage weight vector with Eq. (9).

$$\overline{W} = \frac{w_j}{\|w_j\|} = \frac{(w_{j1}, w_{j2}, \ldots\ldots w_{jN})}{[(w_{j1})^2 + (w_{j2})^2 + \ldots + (w_{jN})^2]^{1/2}} \qquad (8)$$

$$d_j = [\sum_{i-1}^{N} (\overline{P}_i^k - \overline{w}_{ji})^2]^{1/2}\ j = 1,2,\ldots\ldots M \qquad (9)$$

(4) Find out the winning neuron that has the minimum distance $d_g (d_g = \min[d_j], j = 1,2,\ldots\ldots M)$ to the input vector p.

(5) Adjust the linkage weights with Eq. (10). The linkage weights connect all the neurons of the neighborhood in the competitive layer with the input neuron.

$$\overline{w}_{ji}(t+1) = \overline{w}_{ji}(t) + \eta(t) \bullet [\overline{p}_l^k - \overline{w}_{ji}(t)] \qquad (10)$$

$j \in N_g(t) j = 1,2,\ldots\ldots, M (0 < \eta(t) < 1)$, Where $\eta(t)$ is the learning rate at time t.

(6) Feed the network with a new learning vector, then return to step (2), till all the vectors are inputted to the network and the network converges.

(7) Update the learning rate $\eta(t)$ and the neighborhood $N_g(t)$ separately by using Eq. (11) and Eq. (12).

$$\eta(t) = \eta(0)(1 - \frac{t}{T}) \qquad (11)$$

$$N_g(t) = INT[N_g(0)(1 - \frac{t}{T})] \qquad (12)$$

Where, t is the learning times, T is the total times of learning, $INT[x]$ is the sign of getting integer.

(8) Make $t = t+1$, go to step (2), till $t = T$.

4 The Application of the Arithmetic in Image Segmentation

According to the above-mentioned, we know that the Gabor filter $h(x, y)$ is determined by parameters $(u_0, v_0, \sigma_x, \sigma_y)$. So designing a best Gabor filter is to find a block of best parameters $(u_0, v_0, \sigma_x, \sigma_y)$. Filtering processing the input image is by Gabor filter, then the output image could be better adapt to the segmentation algorithm of self-organized neural network.

The frequency bandwidth of the vision cortex cell feeling wile is about 0.2 to 2.5 octave [9, 10], viz.

$$0.5 \le B = \log_2 \frac{\pi\sigma_x \sqrt{u_0{}^2 + v_0^2} + \sqrt{(\ln 2)/2}}{\pi\sigma_x \sqrt{u_0^2 + v_0^2} - \sqrt{(\ln 2)/2}} \le 2.5 \qquad (13)$$

and the Gabor filter is as the characteristic of frequency bandwidth, its crossbar ratio is among 1.5~2.0.

Hypothesis the input image $i(x, y)$ is composed by two kinds texture $t_1(x,y)$ and $t_2(x,y)$, and their energy density spectrum are $S_1(u,v)$ and $S_2(u,v)$ respectively. Then the representation of the filtering output total energy $P_1(u,v), P_2(u,v)$ are as follows:

$$P_1(u,v) = \iint S_1(u,v)\left|H(u,v)\right|^2 dudv \qquad (14)$$

$$P_2(u,v) = \iint S_2(u,v)\left|H(u,v)\right|^2 dudv \qquad (15)$$

The Gabor filter designed as follow:

(1) Calculate the energy density spectrum $S_1(u,v)$, $S_2(u,v)$ of two kinds texture $t_1(x,y)$ and $t_2(x,y)$.

(2) Determine the range of σ_x , viz. $\sigma_x \in [1, \text{width of the image} \times 0.2]$ .Let the initial value of σ_x is 1.

(3) For each determined σ_x , use Eq. (13) shrinking range of center frequency (u_0, v_0).

(4) Calculate each $P_1(u,v), P_2(u,v)$ among the range obtained on step 3 respectively .The center frequency (u_0, v_0) is the point of maximum value corresponding , viz. $(u_0, u_0) = \arg(\max_{(u,v)}\{p_1(u, v / P_2(u,v))\}$.

(5) Viz. $\sigma_x = \sigma_x + 1$, repeat steps 3 to 4 till σ_x reaches maximum, then go to the next step .

(6) From steps 3 to 5 obtaining some groups of data $(u_0, v_0, \sigma_x, \sigma_y)$, quantitative analysis according to the results, then choose the best Gabor filter.

Then use self-organizing feature map segment the output image $i_g(x,y)$ based on Gabor filter, the concrete implementation as follow:

(1) According to processed image determine the center nodes of each clusters, and by the results of the SOFM learning algorithm conserve the connection weight values corresponding the nodes.

(2) Make the each obtained weight values as the net weight of SOFM. Input value of each pixel point into the network to compete. If one pixel point closes with competition node, it is winning. Set the weight right vector corresponding to the node as the pixel point value.

(3) When all pixels complete the competition and judgment at step 2, the image segmentation has implemented.

5 Experimental Results

Figure 2 shows that output texture images of the Gabor filter at four orientations, corresponding orientations are 0, 45, 90 and 135.

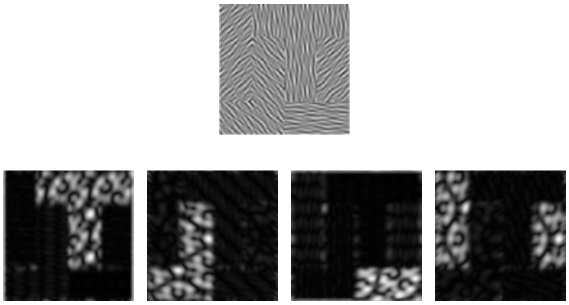
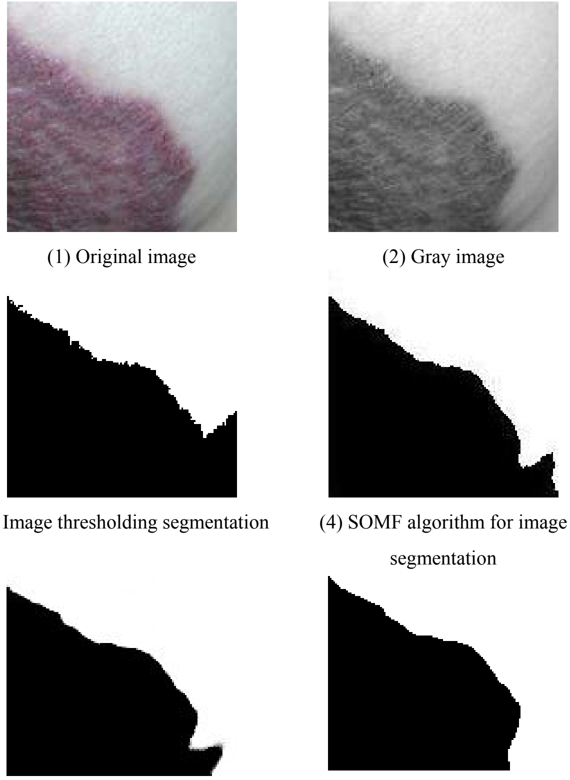


Figure 2 Output images by Gabor filter

The test image is a photo of lesions skin, the texture of lesions skin and healthy skin are different significantly .So use Gabor filter processing the image, from Figure 3 it shows that the proposed algorithm is effective segmenting the test image.



(1) Original image

(2) Gray image

(3) Image thresholding segmentation

(4) SOMF algorithm for image segmentation

(5) Image segmentation based on Gabor filter and threshold

(6) Image segmentation based on Gabor filters and SOFM

Figure 3 The experimental results

6 Conclusion

Experimental results indicate that the proposed algorithm outperforms conventional approaches in terms of both objective measurements and visual evaluation.

References

[1] ZHAO Yin-di, ZHANG Liang-pei, LI Ping-xiang, "A Texture Segmentation Algorithm Based on Directional Gabor Filters," *Journal of Image and Graphics*, 11(4), 2006, pp. 504-510

[2] A. Jain and S. Bhattacharjee, "Address block location on envelopes using gabor filters," *Pattern Recognition*, 25(12), 1992, pp. 1459-1477

[3] A. Jain, N. Ratha, and S. Lakshmanan, "Object detection using gabor filters," *Pattern Recognition*, 30(2), 1997, pp. 295-309

[4] T. P. Weldon, W. E. Higgins, and D. F. Dunn, "Gabor filter design for multiple texture segmentation," *Optical* Engineering, 35(10), Oct. 1996, pp. 2852-2863

[5] HASI Bagan, MA Jianwen, LI Qiqing… "Self-organizing Feature Map Neural Network Classification of the ASTER Data Based on Wavelet Fusion," *Science in China Ser. D Earth Sciences*, 47(7), 2004, pp.651—658

[6] Fangliang Xing, Guangyuan Wang. "Structural Choice Based on Knowledge Discovery System," *Journal of Harbin Institute of Technology (New Series)*, 9(3), 2002

[7] FECIT institute, *MATLAB 6.5 Assistant Analysis and Design in Neural Networks*, Beijing: Electronic Industries publishing company, 2003.1

[8] Reixiang Bai, Hongzhong Hui , Hui Song. "Clustering Analysis System Study Based on Self-organizing Feature Map Neural Net," *Control and Instruments in Chemical Industry*, 31(5), 2004, pp. 29-31

[9] Sheng Wen, Xia Bin. "Texture segmentation method based on Gabor filtering [J]", *Infrared and Laser Engineering*, 32(5), 2003, pp.485-488

[10] Pollen D A, Ronner S E. "Visual cortical neurons as localized spatial frequency filters[J]", *IEEE Transactions on System, Man and Cybernetics*, 13(5), 1983, pp.907-916

Digital Validation Code Recognition Algorithm Research and Design

Peng Cui Qingping Guo Pengpeng Duan

School of Computer Science and Technology, Wuhan University of Technology, Wuhan, Hubei 430063, China
Email:cys19873025@126.com

Abstract

Putting forward a method of extracting the image of validation code base on the amount of information[1]. First and foremost, the image of validation code is processed by Otsu Law, then removes the isolated point in the image and makes use of Median Filtering to denoise. After split the image up into several small digital images, we use the method of Finite element to divide the split digital image into a lot of basal unit. We should compute the amount of information in each basal unit, at the same time; the connecting relationships of each basal unit are recorded. We sort amount of information in each basal unit according to descending order. We choose a suitable threshold value so that we can remove the basal unit that amount of information is below the threshold value. At last we construct the relationship map of the digital image according to the connecting relationships of each basal unit in order to extract the image outline. Then we carry out pattern recognition [2][3] by invariant moments so that we can recognize the digital validation code.

Keywords: Amount of Information; Invariant Moment; Threshold Value; Connecting Relationship

1 Introduction

At present, there are so many validation code based on text in network. The validation code is a image containing a string generated randomly with some interferential pixels ,color and deformation and so on in it randomly .The user is required to recognize the characters in the image so that prevent the malicious registration. The thesis researches the image with some interferential pixels, color and deformation and so on in it randomly and the characters' center of mass should be in the middle of the image.

The technology of validation code recognition is required in so many application in network that it becomes more and increasingly complex; On the other hand ,recognition system just is suitable for some validation codes .There are some sample recognition algorithms during the public algorithms can realize the purpose of recognize part of validation code. We use the method of finite element to divide the split digital image into a lot of basal unit. We should compute the amount of information in each basal unit, at the same time; the connecting relationships of each basal unit are record. We sort amount of information in each basal unit according to descending order. We choose a suitable threshold value so that we can remove the basal unit that amount of information is below the threshold value. At last we construct the relationship map of the digital image according to the connecting relationships of each basal unit in order to extract the image outline .Then we carry out pattern recognition by invariant moments so that we can recognize the digital validation code.

2 Recognition Algorithm

2.1 Binaryzation processing of image

Because of the color interference, we process the image with Otsu Law. Otsu Law is proposed by Otsu in 1979.Let t is the segmentation threshold value of foreground and background, w0 is the percent of the foreground points in the image account for, u0 is the average of gray scale, w1 is the percent of the background points in the image account for.,u1 is the average of gray scale. The total average gray scale of the Image is u=w0*u0+ w1*u1, from the minimum to the maximum gray scale of t, t is the optimal threshold when t makes the equation of g=w0*(u0-u)*2+w1*(u1-u)*2 has the maximum value.

It costs a lot of time to compute by using Otsu Law directly, so in fact ,we adopt the equivalent formulas g=w0*w1*(u0-u1)2.

2.2 Image de-noising

2.2.1 Acnode Elimination

Let Γ is the set of pixel in the image, for each point p(p ε Γ), we should judge whether there are any points around p. If there is nothing, this point is acnode, we should delete p.

2.2.2 Median Filter

Median filter is the method which use a sliding window with odd points in it, then replace gray scale of center of the window by all the points' median value [10].

Let there is a one-dimensional series $f1, f2, ...fn$. choose a window which edge is m(m is odd),we use median filter to treat this one-dimensional series, we choose $ft-v,...,ft-1, fi, fi+1,..., ft+v$, in which fi is e gray scale of center of the window, v= $(m-1)/2$, then sort this points according to descending order and choose the center point as the result of filter. The mathematical formula is;

$$yi = Med\{fi - v..., fi..., fi + v\}$$

$$z \ni i \qquad \mathrm{v} = (m-1)/2$$

2.3 Image segmentation

Due to there are interval between the digital numbers in the validation code image, after median filter the background of the image is white, we use 0 to stand for white. At first we scan the image by horizontal scan line, find out the first black point in both the top and bottom of the image [8]. Then delete the white part of the image. Then we scan the image by vertical scan line; find out the first black point in both the left and right of the image. Then delete the white part of the image. We project the image to x-axis [6]; the projection toward the x-axis wouldn't be 0 if the pixel belongs to the image, unless the all of the pixel of the y-axis direction don't belong to the projection. According to the result of projection, we can find out the interval between the digital characters, and then split the characters [9].

2.4 Image contour extraction

Before extracting the image contour, we propose several definitions:

Definition 1: The segmentation block of the image is related to the image's distribution and proportion in the picture.

Definition 2: When the picture is divided by square grid, the square grid's edge length *g*'s formalized description is as follow:

$$g = \begin{cases} \lceil 1/w \rceil \quad only\,one \quad black \quad point \\ \lceil |\delta/(w*h) - 1/(w*h)| * w \rceil + \lceil |\delta/(w*h) - 1| * 1/w \rceil \\ \delta \quad the \quad number \quad of \quad black \quad points \\ w \quad po\mathrm{int}s\,are \quad all \quad black \end{cases}$$

W stands for the width of the image, h stands for the height of the image.

Definition 3: After being divided by a lot of square grids, let there is a counter, find out the other square grid around a grid，the counter increase one just we find one, then set the found grid invalidation. The final result of the counter is the grid's connecting length until there is no grid around the grid.

The image can be expressed by a m*n two-dimension

array img_{m*n}, $img_{m*n} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$, z϶ m, n，$a_{ij} = \begin{cases} 0 & \text{white} \\ 1 & \text{black} \end{cases}$.After being segmented, let

the image be split into k parts. Each part has a character image.

So the array img_{m*n} is consisted of k parts, we can describe img_{m*n} like this,

img_{m*n} = $(strimg_1, \ldots, strimg_{\lfloor n/k \rfloor})$.

$strimg_i$ stands for the information of the character image

which number is i . $strimg_i$ is a m* $\lfloor n/k \rfloor$ two-dimension array. Then we treat the $strimg_i$ as the following steps:

1) Grid the image of character by finite element. The grid should be chosen according to the definition 1.In order to compute easily, we choose square grid. The square grid's width should be chose according to definition 2. $strimg_i$ can be divided into v=$\lceil (w*h)/(g*g) \rceil$ parts,(v is the edge length of the character image, g's value is defined by definition 2), $subGrid_i$ stands for the ith grid(1 ≤ i ≤ v),The image of $strimg_i$'s edge length is w=width/k, the high is h=height(width is the width of the segmented image, height is the height is the height of the segmented image).

2) Find all of the pixels in the image of $subGrid_i$, compute the total pixels w, black pixels b, then gain the information Info=b/w, at the same time record each coordinate of the black point.

3) Repeat 2) until all the grids of the image $subGrid_i$ is considered. Delete the grid whose Info is 0, so there are l (0 ≤ l ≤ v) grids left. Sort the rest grids according to the information Info's descending order, choose a suitable threshold value so that we can remove the basal unit that amount of information is below the threshold value, record the grids whose connecting length is the largest.

4) Repeat the step 2) and 3) until all the character images are treated. At last we can gain the outline of the image.

2.5 Normalization

Because of existing certain difference on size after the image being cut, relatively speaking, it is better to unify the size of the characters so that we can recognize the characters easily and correctly [7]. Image unification is to make the characters having different size have the same size. We can do it like this:

1) Gain the width and height of each characters;

2) Compared the width and height to the given size in order to gain the transform coefficient.;

3) The points of the new image are mapped to the original image by interpolation method.

2.6 Target recognition

2.6.1 Invariant moment

When being amplified, translated and rotated, the object maintains its shape, we call it invariant moment. Geometric invariant moment [12] has been widely used for image recognition [5], edge detection [11], and template matching, digital watermarking and image analysis. In 1992, Hu put forward a method for image recognition by Hu moment. The method has been developed rapidly, new invariant moments method such as pluralism moment, rotation moment have appeared. After having selected a method to extract the shape of the image, we need to determine under what conditions the two images are similar; the distance similarity method is used commonly. The most commonly used measure of similarity is the Euclidean distance [14].

Definition 4: The Euclidean distance between the model sample vector X and Y is defined as:

$$D(x, y)= \| X-Y \| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Among them, n is the dimension of the feature space. Obviously, if samples X and Y in the same type of region, Euclidean distance D(x, y) is relatively small, if they are located in different types of regions, the Euclidean distance D(x, y) is relatively large. Moments feature is a very useful feature; it has several advantages such as good stability, being realized easily, and matching efficiency highly and so on. We can gain seven rotation invariant [13] moment of the image by the moment feature of the image. The seven rotation invariant moments are invariant for the image's translation [15], rotation and scale. In the process of target identification, respectively, based on regional and based on border of seven invariant moment of the image, we can get the feature vector of the image. Finally, we can gain the similar degree of the two images by eigenvectors of the two images. The feature vector of the object in object image and test image which are consisted of seven invariant moments can be computed according to invariant moment proposed by Hu. The result is stored in two array variant reign and line respectively. Then compute the Euclidean Distance D of the two vectors reign and line as the Euclidean Distance of the object image and test image's normal eigenvector. The similarity of them is determined by a pre-set threshold L, if D<L, the test image is the object we are finding, otherwise not.

3 Test and Results

Find a digital validation code picture in network (see Figure 1)

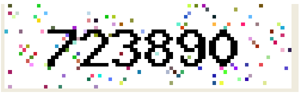
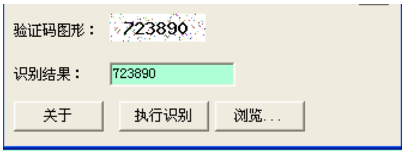


Figure 1 a Digital Validation Code

We gain the result by this algorithm (see Figure 2 the Result)


验证码图形：723890

识别结果：723890

Figure 2 the Result

4 Conclusions

The algorithm' innovation is making use of information theory and finite element to extract contours of the image, and then recognize the object by invariant moments. In the aspect of image extraction, the algorithm has a higher efficiency, compared with the method of handling individual pixels; the algorithm of dividing the image into a lot of grids is faster.

References

[1] Kohonen T.Self-organized formation of topologically correct feature maps [J].Biological Cybernetics, 1982, 43:59-69

[2] B.Moghad dam and A. Pentland, Probabilistie visual learning for object recognition, IEEE Trans. PattemAnal.Maeh.Iniell.Vol.19,No.7,1997,696-710

[3] Oivind Due T，Anil K Jain，Tiffin Taxt. Feature Extraction Methods for Character Recognition-A Survey. Pattern Recognition,1996,29(4):641-662

[4] Neikato, Shinchiro Omaehi. A handwriting character recognition system using directional element feature. IEEE Trans on Pattern Analysis and Machine Intelligenee,1999,21(3):258-262

[5] GAADER P. Recognition of handwritten digits using template and model matching [J].Pattern recognition, 1991

[6] Westall J M, Narasimha M S. " Vertex directed segmentation of handwritten numerals", Pattern Recognition, 1993, 26(10):1473-1486

[7] Lu Z, Chi Z, Siu W C, etc. "A background-thinning-based approach for separating and recognizing connected handwritten digit strings", Pattern Recognition, 1999, 32(6):921-933

[8] Jain A K, Yu B. Automatic Text Location in Images and Video Frames [J].PatternReeognition.1998, 31(12): 205-207

[9] Richard G. Casey and Eric Lecolinet. "A Survey of Methods and Strategies in Character Segmentation", IEEE Trans on Pattern Analysis and Machine Intelligenee, Vol.18, No.7, July, 1998

[10] Seeong, Whanlee. "A new Methodology for Gray-Scale Character Segmentation and Recognition", IEEE Trans on Pattern Analysis and Machine Intelligenee,Vol.18, No.10, 1996

[11] Marr，Hildreth. Theory of edge detention[M]，Proe. Roy. Soc. Lond.，1980，B207:187 一 217

[12] Hu M. K. Visual Pattern recognition by moment invariants [J], 1962, 8(l):179 一 187

[13] J. Flusser，On the independence of rotation moment invariant [J] ，Pattern Recognition, 2000, 33 (9):1405 一 1410

[14] Y. Li，Reforming the theory of invariant moments for Pattern recognition [J] ，Pattern Recognition，1992，25 (7):723-730

[15] P. Yap，R. Paramesran, S. Ong, Image analysis by Krawtchouk moments [J] ，IEEE Transactions on Image Processing，2003，12 (11):1367-1377

Peng Cui is a master degree candidate in the School of Computer Science and Technology, Wuhan University of Technology. He got his bachelor's degree in Wuhan Polytechnic University in 2006.He majors in computer application and his research interests are image processing, J2EE and network security.

Qingping Guo is a Full Professor and a head of Parallel Processing Lab, dean of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology. He graduated from Wuhan University in 1968; from Huazhong University of Science and Technology in

1981 with specialty of wireless technology. He is a holder of K. C. Wong Award of UK Royal Society (1994); was a visiting scholar of City University and University of West Minster (1986~1988), Visiting Professor of the UK Royal Society (1994), Visiting Professor of Queen Mary and Westfield College, London University (1997~2000), Visiting Professor of National University of Singapore (2000), Visiting Professor of University Greenwich (2003). He is one of the DCABES international conference founder, was the chairman of DCABES 2001, co-chair of DCABES 2002, and will be the chairman of DCABES 2004. He has published two books, over 80 Journal papers, edited two DCABES Proceedings. His research interests are in distributed parallel processing, grid computing, network security and e-commence.

Pengpeng Duan is a master degree candidate in the School of Computer Science and Technology, Wuhan University of Technology. She got her bachelor's degree in Zhengzhou Institute Of Aeronautical Industry Management in 2004.She majors in computer application and her research interests are J2EE and image processing.

A Novel Algorithm of Co-articulation Emotional Chinese Speech Synthesis Based on TD-PSOLA

Li Jing Wan Juan Li Lingling

School of Computer Science, Wuhan University, Wuhan, Hubei, 430079, China
Email: leejingcn@msn.com

Abstract

Speech synthesis technologies are the foundation of phonetic human-machine interaction. The tendency in the future development of speech synthesis technologies is how to further improve expressive ability of the machine synthesized speech and imitate the speakers' emotive state, which is still a difficult problem in the research field. In this paper, TD-PSOLA, based on decomposing the spectrum of speech signal, was used to deal with the co-articulation phenomenon in the syllable junction of speech synthesis, and those emotional parameters of speech synthesis were analyzed. CECTD-PSOLA, which is an algorithm of Co-articulation Emotional Chinese Speech Synthesis based on decomposing the spectrum of speech signal, was proposed in this paper. The co-articulation phenomenon between syllables, and emotional factors contained in speech was considered in the CECTD-PSOLA. The experimental results show that the algorithm could better increase naturalness and emotion of Chinese speech synthesis.

Keywords: Chinese synthesis speech, Pitch synchronous overlap add, Collaboration syllable, Emotional speech

1 Introduction

Speech synthesis has made dramatic development from the early time when using brain, ears and the main pronunciation organ to present's electronic computer in the speech processing. However, people have already not been satisfied with splicing syllable or phrase that storing in the computer to achieve “Machinery Mouth”, but higher requirements for naturalness and emotion of synthesized speech. What more, different age and sex characteristics, tone and speed of Speech Synthesis should be taken into account, make synthesized speech full of individual emotion. Namely lets the Speech Synthesis from the level of text to speech to the level of concept to speech.

In the present speech synthesis methods, the methods with good naturalness based on waveform is popular, but the merge for syllables after synthesizing is difficult for case of co-articulation phenomenon between syllables. The naturalness of synthetic voice will greatly increase if co-articulation phenomenon is taken into consideration [1]. And in order to enhance emotion of synthetic voice, the emotional and sentimental factors of speech signals should be taken into account [2].

Chinese is a language of widely used all over the world, with a population of more than 10 billions who uses it, the research of Chinese Speech Synthesis is very important. Therefore, CECTD-PSOLA, a Chinese speech emotional synthesis algorithm based on TD-PSOLA of spectral decomposition was proposed. Speech was first synthesized with TD-PSOLA technology, on the base of extracting prosody and spectrum parameters by analyzing original voice[3][4]. We synthesize emotional voice with STRAIGHT synthesis technology considering the changing of the parameters including prosodic features and spectral features.

2 Collaborative Speech Synthesis Based on TD-PSOLA

2.1 The selection of synthetic voices element

In Chinese, as for a syllable of the word flow, co-articulation depends on vowel at the end of the previous syllable and consonant at the beginning of the adjacent syllable. Monosyllabic words, two-syllable words and three-syllable words are the basic units of Chinese. The monosyllabic word was used as synthetic unit, which can reduce storage space, but the quality of the synthesized voice will be not very good. Furthermore, usage of two-syllable word can maintain co-articulation phenomenon between syllables, but greatly increase memory size of the speech database. In the present paper, therefore, a method for synthesized voice was established using the synthetic unit selection strategy, which chose mainly the monosyllabic words, and assisted by two-syllable words that have obviously co-articulation phenomenon.

2.2 The algorithm steps of TD-PSOLA

Three steps of TD-PSOLA algorithm are as follows[3][4]：

① Pitch synchronization analysis

The original speech signal $x(n)$ multiply by a series of the analysis window function $h(n)$ of pitch, we can obtain a series of short-term signals $x_m(n)$ that pitch synchronized and overlapped： $x_m(n)=h_m(t_m-n)$ $x(n)$, where, t_m is the position of the base unit. The centre of hamming window function $h_m(n)$ locate in t_m , its definition is given by:

$$h_m(n)=\begin{cases}0.5[1-\cos(2n/N)];n=0,1,2,\ldots,n-1\\0;n\in others\end{cases} \qquad (2\text{-}1)$$

② Pitch synchronization transform

Doing time domain transform on the short-term signals $x_m(n)$ that after pitch synchronization analyzing will gain a series of short-term signals $x_q(n)$ that synchronize with pitch curve. At the same, Time pitch mark of original signal t_m *is* corresponding changed to synthetic pitch mark t_q. From the mapping of $x_m(n)$ to $x_q(n)$, a segment signal $x_m(n)$ that selected will be converted to $x_q(n)$ according delay sequence $\delta_q=t_q-t_m$, as Eq.(2-2)：

$$x_q(n)=x_m(n-\delta_q)=x_m(n+t_m-t_q) \qquad (2\text{-}2)$$

③ Pitch Synchronous Overlap-Add synthesis

A series of short-term signals with window function processed are synchronous arrayed according to pitch period and do least squares overlap adding. In this way the voice waveform will be attained. The formula of least squares overlap add method is as follow：

$$x(n)=\sum a_q x_q(n)h_q(t_q-n)\Big/\sum h_q^2(t_q-n) \qquad (2\text{-}3)$$

Where, $h_q(n)$ are sequences of window, a_q are standardized factors that compensate the energy loss resulted from pitch transforming. $x_q(n)$ are a series of short-time signals processed [12][13].

3 Emotional Chinese Speech Syntheses

3.1 Acoustic characteristics of emotional speech

With the samples splicing synthesis technology of large speech database come to maturity gradually, the naturalness and understandability of synthesized speech have been great improved, however, but due to lack of emotional expression and speech changes, The applications of speech synthesis technology had greatly limited. The effective way to make Speech Synthesis fill with personal emotion is to analyze characteristics of parameters, and adjust the relevant parameters to change speech tone and intonation. Kawanami and other ones indicate that, Emotional Speech characteristics have divided into prosodic and spectral features. Spectral parameters need to be analyzed when analyzing prosodic parameters [4].

A three-dimensional coordinate for the Emotional Speech was proposed in this paper. A point(x1, y1, z1) in the coordinate is expressed as text information, spectral information and prosodic information. Among them, prosodic information is mainly composed of time length, pitch period, and magnitude strength. Spectral information is some sound source spectral parameters, the three-dimensional coordinate is as shown in Figure 1.

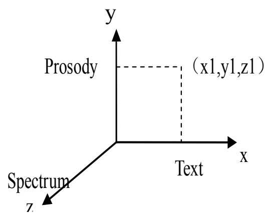


Figure 1　Coordinate of Emotional Speech

In this paper, we choose four representative emotions: happy, anger, surprise and sad for analysis Emotional Speech. In the aspect of speech speed , the order is: anger<surprise<happy<sad, while in the aspect of the length of pause, the order is: happy<anger<panic< grief .In the aspect of pitch mean, happy, anger and surprise are bigger than sad, but the change of pitch is more frequent than happy, anger and surprise. In a word, the pitch of anger and happy is high and fluctuate more frequent, the pitch of surprise is high but fluctuate less frequent, the pitch of sad is low and fluctuate less frequent.

3.2　STRAIGHT synthesis technology

STRAIGHT is a high performance analysis synthesized algorithm for speech signals, accurate spectral envelope of speech signal that remove the influence of pitch are extracted with adaptive interpolation smoothness for short-term spectrum of speech[7]. And time length, pitch and spectral parameters are adjusted with flexibility in the process of restoring voice. The whole analysis synthesis process is mainly composed of the following procedures:

(1) Spectral estimation of Speech signal that removed the influence of period.

In order to estimate the spectral envelope correctly that avoids the impact of pitch to adjust prosodic features flexibly. STRAIGHT technology adopts a smooth method of two-dimensional convolution triangular window.

$$s\left(w,t\right)=\sqrt{g^{-1}\left(\iint_{0} h_t\left(\lambda,t\right)g\left(\left|F\left(w-\lambda,t-\tau\right)\right|^2\right)d\lambda d\tau\right)} \qquad (3\text{-}1)$$

$$h_t\left(\lambda,\tau\right)=\frac{1}{4}\left(1-\left|\lambda/w_0\left(t\right)\right|\right)\left(1-\left|\tau/\tau_0\right|\left(t\right)\right) \qquad (3\text{-}2)$$

Where, $F\left(w,t\right)$ is the short-term spectrums that calculated, $S\left(w,t\right)$ is the spectral envelopes that smoothed. The function of $g(\ )$ is defined as a certain characteristic of retaining spectral parameters when smoothing [7][8].

(2) The extraction of smooth and reliable pitch track

STRAIGHT technology analyzes pitch of speech signal with wavelet analysis. First find the corresponding basic element of speech signal in the case of not knowing pitch, and then calculate instant frequency as pitch of speech signal.

(3) Implementation of Synthesis

Two-dimension spectrum and pitch track of speech signal that extracted in the above analysis are the inputting parameters for synthesis. In the process of synthesis a method based on Pitch synchronous overlap add and minimum-phase impulse response are used, and can implement adjustment of time length, pitch, spectral parameters. Their functions are given by:

$$y(t)=\sum_{t_i\in Q}^{n}\frac{1}{\sqrt{G(f_0(t_l))}}vt_i(t-T(t_i)) \qquad (3\text{-}3)$$

$$vt_i(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}v(w,t_i)\psi(w)e^{jwt}dw \qquad (3\text{-}4)$$

$$T(t_i)=\sum_{t_k\in Q,K<i}\frac{1}{G(f_0(t_k))} \qquad (3\text{-}5)$$

Where, Eq. (3-3) reflects the process of Pitch synchronous overlap add, and $y\left(t\right)$ is recovering speech signal, Q is set of Pitch Synchronous position for synthesis; Function $G(\ )$ is defined as the adjustment of pitch, which can be any form of mapping relations; Eq. (3-4) reflects the processing of obtaining each frame corresponding impulse response, and $v\left(w,t_i\right)$ is defined as Fourier transform of minimum-phase impulse response. $\psi\left(w\right)$ is activation with additional control phase to improve listening flu; Eq.(3-5) reflects the process of confirming pitch synchronous stations.

$v\left(w,t\right)$ can be calculated from the smooth spectrum that previously obtained. That is to say, We adopt a method based on cestrum to complete the transform from general phase of spectrum to minimum phase of spectrum [7][12][13].

The formula $v\left(w,t\right)$ is shown as follow:

$$v(w,t)=\exp\left(\frac{1}{\sqrt{2\pi}}\int_{0}^{\infty}h_t(q)e^{jwq}dq\right) \qquad (3\text{-}6)$$

$$h_t(q)=\begin{cases}0, & q<0\\ c_t(0), & q=0\\ 2c_t(q), & q>0\end{cases} \qquad (3\text{-}7)$$

$$c_t(q)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-jwq}\log A(s(\mu(w),r(t)))dw \qquad (3\text{-}8)$$

Where q denotes cestrum , $A(\ )$、$\mu(\ )$、$r(\ )$ are defined as the adjustment of magnitude, frequency and timelines for the smoothing spectrum $S(w,t)$ respectively [7][12][13].

3.3 The algorithm of CECTD-PSOLA STRAIGHT synthesis technology

Speech can be synthesized using TD-PSOLA technology and the changing of the parameters that including length of time, pitch period, amplitude strength. But expression of emotion is not enough. So we should do spectral decomposition of speech. In this paper, we introduce STRAIGHT synthesis technology to extract spectrum of speech signal and complete the synthesis from spectrum to voice.

The entire synthesis system framework is shown in Figure 2.

The process is described as follows:

Step1. *To synthesize voice:*

Text is synthesized to voice using speech synthesis system based on TD-PSOLA after text analyzing.

Step2. *Two segment voices is expressed in emotional speech coordinate:*

In emotional speech coordinate, One segment voice is expressed by dividing into three dimensions that text dimension、prosody dimension and spectrum dimension. Choose four pairs of voice signal including happy, anger, surprise and sad, which have same text dimension, and one segment is from natural voice, the other one is from synthetic voice.

Step3. *To extract emotional parameters including prosody and spectrum by prosodic analysis and short-term spectral analysis:*

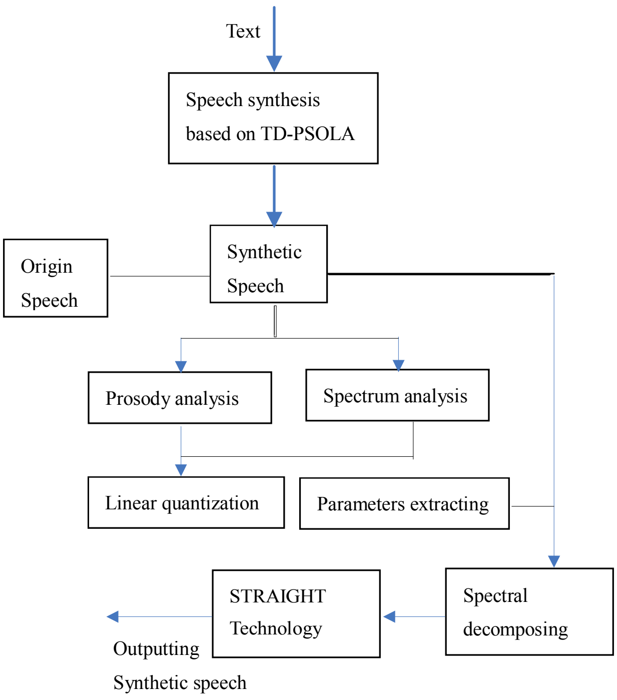


Figure 2 The entire synthesis system framework

In order to extract reliable parameters, speech signal is first processed by inverse filter. The filter of those transfer function is equal to the reciprocal of the channel filter is designed for speech signal to filter.

Step4. *Linear quantization:*

To research prosodic and spectral features of the different emotional voice, we use a sound transform algorithm based on mapping [6][9]. Three codebooks are produced from the training data of same text of the speech signal. Differences of prosodic and spectral features between natural speech and synthetic speech of the same emotion are gained by comparing these codebook vectors. On the base of analyzing four kind emotions, we can gain four pairs of mapping codebooks. Table 1 lists the conditions of the analysis.

Step5. *Parameters analysis of different emotions:*

① Prosodic features of different emotions

We extract Short-term energy and time length as prosodic feature. Below is analysis of short-term energy and time length of a representative of the speech material "哦，是这样" with different feelings, as Figure 3、4 follows：

Table 1　The conditions of Quantitative Analysis

Sampling frequency	12kHz	Clustering Measurement	Euclidean Distance of Cepstrum
Window length	256 dots	Clustering Training Samples	1200 Frames
Window shift	36 dots	Code size	256
LPC analyzing steps	14	Pre-emphasis	1-0.97Z-1
The necessary training words of mapping codebooks			20

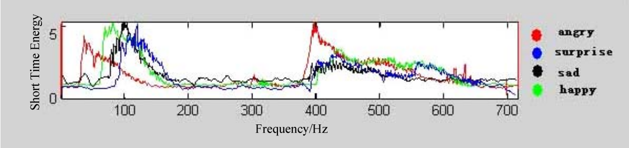


Figure 3　Short-term energy

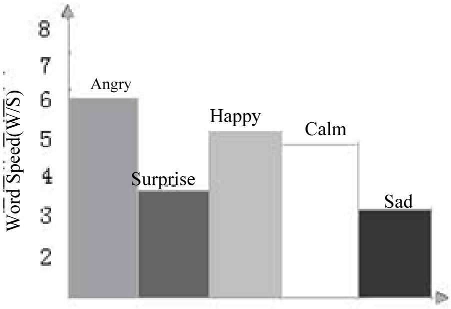


Figure 4　Emotional Speed

② Spectral features of different emotions

Formant is an important feature in the relevant spectral parameters. The first three formant means (F1, F2, F3) with different emotional states of speech material selected are to do statistical analysis, with which we found that F1 of natural speech and synthetic speech is very similar[10][11]. Formant frequency of anger is highest, happy is second, while surprise and sad is lower. There exists difference between F2 and F3 of natural speech and synthetic speech. In the synthetic speech, the formant of happy is obviously decreased, while the formant of surprise is obviously increased.

Step6. *Comparing differences between natural speech and synthetic speech*

After extracting prosodic and spectral features from natural speech and synthetic speech, these features will be marked in the three-dimensional emotional coordinate and form two three-dimensional vectors with which we are compared. In order to reduce the complexity of the algorithm, we only adjust those parameters with larger difference in the next phase of treatment.

Step7. *STRAIGHT Synthesis*

On one hand , If the difference of prosody dimension is larger, we should adjust the parameters of pitch、time length and pause of synthesis unit. On the other hand, if the difference of spectrum dimension is larger, pitch envelope and spectrum are gained by doing STRAIGHT analysis on natural speech, Sound Source parameters of synthetic speech of one specific emotion and pitch information forecasted are combined used for producing Sound Source spectrum of the corresponding emotion, and then superpose it on the Channel Response estimation that gained in the above step in order to acquire voice spectrum, At last, speech is synthesized to meet the adjusted pitch and spectral parameters using STRAIGHT Synthesis technology.

3.4　analysis of the experimental results

In this paper, the emotional speech materials used to test the effectiveness of the algorithm come from the actors' one hundred dialogues. And then use audio processing tool for marking pitch and other information. After processing, choose the representative sentences. In order to verify the quality of emotion loaded in synthetic speech, we will gain the emotional states with making judgments by four students. The results are statistically calculated by the following formula:

$$e = \frac{M_n^{'}}{M} \tag{3-9}$$

Where e is the correct rate, $M_n^{'}$ is the number of being accurately judged in the emotional states n, and M_n is the total number of emotional states n. The correct rate of CECTD-PSOLA is shown in Table 2.

Table 2　The correct rate of CECTD-PSOLA

Emotional state	The correct rate
happy	75%
angry	80%
surprise	58%
sad	63%

From the table, we find that in most cases the majority of emotional speech synthesis results through this system can be recognized. The correct rate of CECTD -PSOLA is better than that of TD-PSOLA.

4 Conclusions

In the present paper, CECTD-PSOLA, a co-articulation Chinese emotional Speech Synthesis Algorithm based on TD-PSOLA technology was proposed. The monosyllabic word was selected as major synthetic unit, and the two-syllable words which have obviously co-articulation phenomenon were used as supplementary synthetic unit. The general voice was synthesized using TD-PSOLA technology. After extracting prosodic and spectral parameters of original voice, the emotional speech was synthesized by STRAIGHT technology and modification of prosodic and spectral parameters of the emotional voice. The experimental results showed that the voice synthesized by co-articulation Chinese emotional speech synthesis algorithm based on TD-PSOLA technology has good naturalness and emotion.

References

[1] ZHANG Qin, LI Hui, DAI Bei-qian. A New Mandarin Speech Synthesis Basedon Coarticulation. MINI-MICROSYSTEMS. 2003,June,Vol.2,No.6

[2] Cowie, R.. Describing the emotional States Expressed in Speech. ISCA Workshop on Speech & Emotion. Northern Ireland 2000,pp.11-18

[3] Yunbo Zhu,Li Zhao.A Chinese Text To Speech System Based on TD-PSOLA. Proceeding of IEEE tencon 2002

[4] Xuejing Sun. Voice Quality Conversion in TD-PSOLA Speech Synthesis. Proceedings 2000 IEEE International Conference on,Volume 2,5-9 June 2000

[5] Wei Zha and Wai-Yip Chan, A Data Mining Approach to Objective Speech Quality Measurement. ICASSP 2004,IEEE 2004

[6] M.Abe, S.Nakamura, K.Shikano and H.Kuwabara. Voice conversion through vector quantization. In Proceedings, ICASSP88, 655-658, 1998

[7] Kawahara.h, Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. Acoustics、Speech and Signal Processing, 1997 IEEE International Conference

[8] Yamagishi.J、Kobayashi.T、Tachibana.M、 Ogata.K、Nakano.Y, Model Adaptation Approach to Speech Synthesis with Diverse Voices and Styles. IEEE International Conference , April 2007.p.1233-1236

[9] Murray I，etal. Towards the Simulation of emotion in Synthetic Speech:A review of the Literature on Human Vocal Emotion[J].in Journal of the Acoustic Society of America,1993:10 97-1108.

[10] Roddy Cowie,etal. Describing the emotional states' expressed in speech [J]. Speech Communication, 2003:5-32

[11] R.Cowie, E.Douglas-Cowie, N.Tsapatsoulis, GVotsis,S. Kollias, W.Fellenz and J.G.Taylor. Emotion Recognition in Human-Computer Interaction [J].IEEE Signal Processing Magazine, No.l, January 2001,pp.32 -80

[12] Chung-Hsien Wu、Chi-Chun Hsia,Voice Conversion using Duration-Embedded Bi-HMMs for Expressive Speech Synthesis. IEEE Transactions on Audio, Speech and Language Processing,Vol.14,No.4,July 2006

[13] Werner.S、Eichner.M.、Wolff. M.、Hoffmann. R, Toward spontaneous speech Synthesis-utilizing language model information in TTS.Speech and Audio Processing ,IEEE Transactions ,July 2004,pp.436-445

Design and Algorithm Research of Image Compression System Based on DSP Techniques

Chen Yuping[1] Ma Yan[2]

1 Jiangsu Wuxi Institute of Communications Technology, 214151, China

2 College Of Information Technology, Taishan University, Tai’an, Shandong, 271021, China

Abstract

The need for effective data compression is evident in almost all applications where storage and transmission of digital images are involved. For the simple hardware complexity it is so easy to implement in SILI. A logic scheme image processing system based on coordinated DSP and FPGA techniques is presented, in this system, FPGA is used as image acquisition unit and DSP is designated to image compression. In order to dispose computational complexity and search precision aiming at the black match algorithm, in this paper，a method of Block Match Quantum-behaved Particle Swarm Optimization(BMQPSO) is introduced in order to realize the image compression problem. During the process of image compression, an ordered representation frame of image is first searched by particles, and then the compressed code was optimized according to the particles astringency. Experimental results show that the compression efficiency of BMQPSO algorithm is much better than classical algorithms.

Keywords: image compression，image acquisition, DSP, FPGA, BMQPSO algorithm

1 Introduction

Nowadays, the development of the network and digitized technology is making the image processing require higher and higher. Development of multimedia technology requires file and image stored and transmitted to meet requirement of terminal user. For example, a image, 720 pixel×480 thread，each pixel in the each color weight use as 8bit, transmit 30 frames per second, require the transmission ability of the channel to reach 248Mb/S. If not compressing image, memory capacity, transmission rate and processing speeds of computer, etc. were caused great pressure. So, the image processing becomes the key technology of image communication [1], the video applications of the high-resolution put forward to very high requirement for the video computing capability of an encoder at the same time.

At present, CPU/DSP with single performance does not possess higher computing capability. In order to solve the problem, which the operation amount is great and real-time processing in the image compressing system. This text has proposed DSP design plan together with FPGA (Field Programmable Gate Array) [2], has solved the contradiction between the operation amount big and real-time processing well. Using FPGA gather image, compressing with high-performance DSP, Therefore DSP and FPGA can be operated successively for the line production operation .In this way it can improve the performance of the encoder effectively.

At the same time, in order to reduce the calculation complexity of the algorithm, this text introduced QPSO algorithm (BMQPSO) based on block matches. Since blocks of algorithm of matching calculation is very complex, usually account for more than 50% of system’s running capacity. In order to improve the efficiency of the motion estimation and make motion search and estimate algorithm course faster, more efficient, BMQPSO algorithm, first of all, calculates the block complexity (MBC), then values the present block code typing by comparing present MBC of block and SAD (sue for peace absolute error), reduces the calculation complexity of the algorithm greatly.

2 Systemetic Design Plan

In order to realize the reliable and widely adaptable hardware system, the text select FPGA (Virtex1000 of Xilinx Company), which is responsible to gather and control the outside data. DSP is TMS320C6000 of TI Company [5-6]; it is mainly responsible for the code compressing. Its system function structure is shown in Figure 1. Since the calculating amount of video compress code is high, in order to improve the video frame rate as much as possible, DSP especially responsible for video compress while other function of image gather are realized by FPGA. Video compress mode is first captured by FPGA module, and then makes an outside break off to DSP module. When DSP responds，it will read user orders which are stored in FPGA, and it will compress the codes according to user's requirement of video form, and frame image resolution and video compress speed .

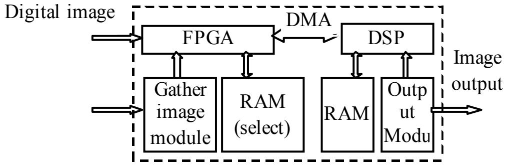


Figure 1 Image compress system function diagram

The functional diagram of FPGA module is shown in Figure 2. As shown in Figure 2, video decoder resulting digital signal and YUV (Y video luminance, UV color weight) dataflow were sent to gather systematic FPGA together. At first, to pre-process the out-dataflow, then these out-dataflow were deposited with table tennis frame way, end with compressing the systematic interface control with the image .Image Compress system release data deposited in frame and then release the control power. The FPGA that is used in the central controls part has systematic doors of 100,000 equivalents, and systematic clock frequency can reach 240MHz, the amount of I/O that user can spend in charge of foot have 196, nuclear voltage can reach 1.8V, peak value consumption lower [3-4].

DSP compresses and realizes the following Figure 3 partly:

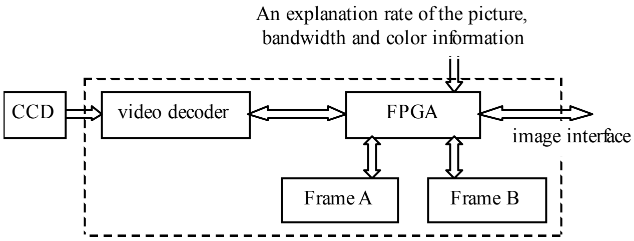

An explanation rate of the picture, bandwidth and color information
CCD
video decoder
FPGA
image interface
Frame A
Frame B

Figure 2 image gather systematic hardware structure chart

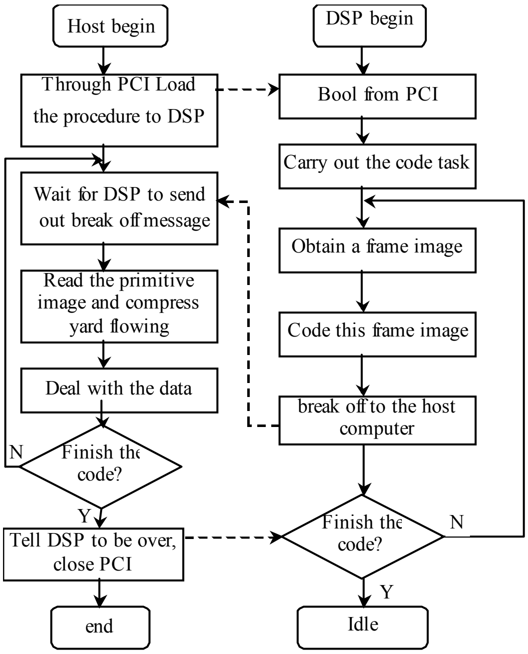

Host begin
Through PCI Load the procedure to DSP
Wait for DSP to send out break off message
Read the primitive image and compress yard flowing
Deal with the data
N
Finish the code?
Y
Tell DSP to be over, close PCI
end
DSP begin
Bool from PCI
Carry out the code task
Obtain a frame image
Code this frame image
break off to the host computer
Finish the code?
N
Y
Idle

Figure 3 system diagram

First of all, Host computer initialize DSP and load the procedure through PCI, then DSP begins to operate the coded program, obtain the video gathered from the video port at the simultaneous time. Digital video input in DM642VPORT that is VPOR export, VPORT output YUV is as the input in the coded program of form, DSP finish one frame code of image, send through PCI to the host computer break down, the host computer responds and breaks off, reads and deals with the primitive image data and dataflow compressed from the memory space of DSP.

3 The Optimization of the Compress Algorithm

The experiment shows, if the code and the data that code will visit are on the memory storage area of C6000

(PRAM and DRAM), its code carries the speed is 17 times on average in synchronous SDRAM outside chip (bus width is 256 in chip , the access to data is one CPU cycle). So putting the code and data in chip will improve running speed of procedure greatly.

Fast estimated algorithm has been adopted in MPEG-4 algorithm, but the calculating amount is obviously very big while identification a code type. For this reason, this text has proposed matching particle group algorithm with behavioral quantum and optimizing the algorithm on the basis of block (BMQPSO, Block Match Quantum-behaved Particle Swarm Optimization), reduce sport calculation amount to estimate course further.

3.1 Qantum-behaved particle swarm optimization [7,8]

Particle swarm optimization (PSO) is an evolutionary computation technique developed by Dr. Eberhart (http://www.engr.iupui.edu/~eberhart/) and Dr. Kennedy (http://users.erols.com/cathyk/jimk.html) in 1995, inspired by social behavior of bird flocking or fish schooling[5, 6]. However, the existing PSOs, make the particle only search in a finite space. So a Quantum-behaved Particle Swarm Optimization (QPSO) is proposed that outperforms traditional PSOs in search ability as well as guaranteeing global convergence algorithms. In the QPSO model[7，8，10], each individual is treated as a volume-less particle in the D-dimensional space. The particles move according to the following equations.

$$mbest = \frac{1}{M}\sum_{i=1}^{M} P_i = \left(\frac{1}{M}\sum_{i=1}^{M} P_{i1},......,\frac{1}{M}\sum_{i=1}^{M} P_{id}\right) \qquad (1)$$

$$p_{id} = \varphi * P_{id} + (1-\varphi) * P_{gd} \qquad \varphi = rand() \qquad (2)$$

$$x_{id} = p_{id} \pm \alpha * \left| mbest_d - x_{id} \right| * \ln(\frac{1}{u}), u = rand() \qquad (3)$$

Where, Vector $P_i = (P_{i1}, P_{i2}, ..., P_{id})$ is the best previous position (the position giving the best fitness value) of particle i called *pbest* , and vector $P_g = (P_{g1}, P_{g2}, ..., P_{gd})$ is the position of the best particle among all the particles in the population and called *gbest* . p_{id} , a stochastic point between p_{id} and p_{gd} , is the local attractor on the d*th* dimension of the i*th* particle, mbest is the mean best position among the particles. φ is a random number distributed uniformly[0,1]. u is another uniformly-distributed random number on [0,1] and α is a parameter of QPSO that is called Contraction-Expansion Coefficient. In experiment the α is used by the equation:

α = (1.0-0.5)*(MAXTIMES-T)/MAXTIMES+0.5

3.2 Model application of bmqpso algorithm

The application method of the algorithm is: According to certain match rule, utilize the procedure of searching to find the best movement vector to estimate through the image element between two frames. As shown in Figure 4, it take image element piece $M \times N$ (centre position in (x_0, y_0)) displacement, through search for frame $k-1$ (search for frame) of the same size one best to match piece come and confirm. From calculating factors consider, its search for range usually limit in ($M + 2M_1, N + 2N_1$), M_1 , N_1 value can follow the concrete estimation. Regard every image element piece as one particle, through constantly Generations and dynamic adjust to search for adjoin frame corresponding image element block, thus reach the best result of matching with the minimum calculation amount.

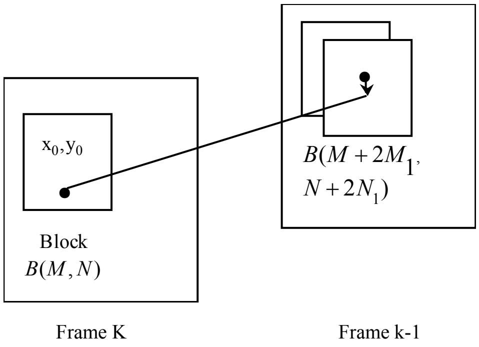


Figure 4 Principle diagrams of block match

So the whole efficiency of the algorithm embodies in picture quality, compresses yard of rates and searches speed (complexity) for this three respects. The more accurate of the sport is estimated, the higher the quality of the image is. The common match function is sued for absolute error; its definition is as follows:

$$SAD(d_x,d_y)=\sum_{(x_1,y_1)\in B}\left|f_k(x_1,y_1)-f_{k-1}(x_1+d_x,y_1+d_y)\right|$$

```
for each particle
Initialize an array of particles with random position
end
for T from 1 to MAXTIMES
clustering by the distance of Euclidean to equation
calculate fitness value to equation and determine the mean best position among the particles by $mbest = \frac{1}{M}\sum_{i=1}^{M} P_i = \left(\frac{1}{M}\sum_{i=1}^{M} P_{i1},......,\frac{1}{M}\sum_{i=1}^{M} P_{id}\right)$
for each particle
update the local optimization of particle *pbest* according equation (13), compare with the particle's previous best values: if the current value is less than the previous best value, then set the best value to the current value.
update the global particle *pbes* according equation (14), compare the current global position to the previous global: if the current global position is less than the previous global position, then set the global position to the current global
End
Repeat steps until a stop criterion is satisfied or a pre-specified number of iterations are completed.
end
```

4 Experimental Result

In order to prove performance of optimize algorithm, count to compare through every block average search point in every image array search speed. The main parameter adopted in the test, QCIF(176×144), SIF(352×240), Picture divide size of block is 16×16，Sport estimate search for area is 15×15，block matching criterion adopts SAD (the Sum of Absolute Differences), Starting point predict adopt prediction method based on adjoin incapability exercises vector。If at present all adjoin block vector is equal ,so regard it as present vector predicted value , Otherwise, use the starting point based on SAD value to predict, solve the present block and its adjoin block SAD value respectively, then choose SAD minimum movement vector as the predicted value. This method predicts precision is high, but calculate SAD time of value expenses is vast. Following parameter values for the results reported in this paper: popsize = 30, pchro =3, pmut = 0.01, Maxgen = 100,

The search points for Table 1 reflect the speed of the algorithm. The match probability of Table 2 reflects the accuracy of the algorithm. 4 optimization search algorithms demonstrate high search speed to all test arrays, and have reached the design object which searches for the speed of optimizing. Reveal from the data in the form, BMQPSO algorithm is equal to 4SS algorithm in block matched, but than FS algorithm and 3SS algorithm have more obvious result, so to deal with block match, search for algorithm is a new attempt of method.

Table 1 Different algorithms search for the average search points

	FS	3SS	4SS	BMQPSO
Foreman	873	46	13.6	12.8
Miss American	873	46	12.7	13.3
Mobile calendar	873	46	16.8	16.6

Table 2 Different algorithmos search for the match minimum probability

	FS	3SS	4SS	BMQPSO
Foreman	91.57	87.62	88.96	89.13
Miss American	89.91	88.69	88.74	9.95
Mobile calendar	90.03	87.82	88.68	87.97

5 Conclusion

In the system using FPGA as logic control to realize the image gathered, can improve systematic flexibility and expansibility, at the same time reflected less hardwires complexity, to different image input data , only need revise partly buffer in FPGA and work again. Compressing code in DSP, two can run side by side, work well, reduce hardware debug difficulty. Adopt BMQPSO algorithm reduce calculation complexity to compress algorithm effectively, compress punish, having better attempt application in image.

References

[1] He, X.H. Communication of image [M].xi’an：Electronic publishing house of University of Science and Technology，2005:100-106

[2] Niu, J.W., Hu,J.P.，Mao,S.Y.. Design optimizing realizing with the algorithm together with video encoder of FPGA on the basis of DSP [J]. Aviation journal，2005，26(1):90-93

[3] Zhu, H.G.，Ding,W.R.. The video compresses research and design of the encoder [J]，The signal and information processing，2006，36(6):36-38

[4] Zhao, B.J.,Shi,C.C.. Compression of picture when FPGA and what DSP realize are real of the base [J], Electronic journal，2003，31(9):1317-1319

[5] Texas Instrument Co. H.263 encoder TMS320C6000 implementation [Z]. http://www.ti.com./

[6] Texas Instrument Co. TMS320C6000 imaging developer kit user guide [Z]. http://www.ti.com./

[7] Sun, J., and Xu W.B. *.A Global Search Strategy of Quantum-behaved Particle Swarm Optimization*[C]. Proceedings of IEEE conference on Cybernetics and Intelligent Systems, 2004:111 – 116

[8] Sun, J., Feng B. ， and Xu W.B..*Particle Swarm Optimization with Particles Having Quantum Behavior*[C]. Proceedings of 2004 Congress on Evolutionary Computation, 2004:325-331

Combined Forecast Model and Application Research of Tobacco Sales Based on Group Method of Data Handling and Auto Regression Integrated Moving Average

Weimin Liu[1, 2] Aiyun Zheng[2] Sujian Li[1] Jiangsheng Sun[1]
Fanggeng Zhao[1] Zhihong Li[3]

1 Department of Logistics Engineering, University of Science and Technology Beijing Beijing 100083, China

Email: lzh-tsh@163.com

2 Department of Mechanical Engineering, Hebei Polytechnic University ,Tangshan, Hebei Province 063009, China

Email: zay@heut.edu.cn

3 Department of Industrial Engineering, Tsinghua University,Beijing 100084, China

Abstract

Tobacco sales prediction is important to policy formulation and management upgrading of Chinese Tobacco. In this paper, the characteristics and impact factors about tobacco sales forecast system were detailed analyzed. Particularly, aiming at the long-term growth trends and seasonal fluctuations of monthly sales, a hybrid method which combined both ARIMA and GMDH models were proposed. The proposed method took advantage of each model's strength in linear and nonlinear modeling. Real sales data (Jan 2002 to Mar 2007) of one municipal tobacco commercial company were used to test this model. The tested model was used in sales prediction of first three months 2007. By analyzing the PE (Percentage Error) and MAPE (Mean Absolute Percentage Error) of forecast results, it indicates that the accuracy of combination forecast model can fit the need of forecast process and the proposed method is an effective way to tobacco sales prediction.

Keywords：Tobacco sales forecast; GMDH; ARIMA; Combined forecast; data mining

1 Introduction

Chinese Tobacco's logistics operating system is like "raw and supplementary materials suppliers-industrial enterprises-commercial companies-retailers-consumers", which is a multilevel supply chain model. In this chain, industrial enterprises organize production under the market demands and the macro-control. The commercial enterprises are the profit producing points which up against the market directly and hold the market information. The accurateness of market information is the basement to construct core competitiveness of Chinese Tobacco, which includes three main aspects: acute market insight, far-reaching marketing net and efficient logistic delivery system. Accordingly, the prediction of tobacco sales is very important for Chinese Tobacco to formulate long term development policies and to optimize the logistics system. Now, Chinese Tobacco is carrying out experimental work named "Organizing the supplement according to orders", which includes sales tracking, orders forecasting, delivery time programming, brands cultivating and information real-time interactive online between industrial enterprises and commercial companies, etc. The aim is to change the supply chain management style from "pushing" to "pulling". The former is driven by production and the latter by orders. In this process, tobacco sales prediction is the foundation and emphasis.

The forecast theories, models and algorithms and their application researches have made solid foundations and technical supports to tobacco sales prediction. The classic methods include moving average, exponential smoothing and regression analysis, etc. Recent years, the methods such as grey system, support vector machine and combined forecast, have achieved many successes. Therefore, it has practical significance that selecting applicable models and algorithms to tobacco sales prediction and developing them in application.

2 Tobacco Sales Forecast System and Impact Factors Analysis

Tobacco sales prediction includes long term, medium term and short term prediction. Each type has individual forecast object and service target but associated each other also. Next, each type of prediction and its impact factors will be detailed analyzed.

Long term prediction means the annual one whose object is all brands' yearly total sales. The result is very important to policy formulation, brand planning and macroscopic controlling of Chinese Tobacco. It's also the basement of yearly order planning between industrial enterprises and commercial companies. Annual sales forecast is based on analysis of historical records and impact factors such as social economy, retailer terminal form and marketing policy. First, the social economic factors include: a. Regional consumption structure factors. Such as consumer market constitute and consumer preferences. The main aspect is distinction between rural and urban. b. Supply factor. Because of the restriction of industrial production and contract between industrial and commercial, retailers' demand can't be entirely satisfied. So, they will order excessively which lead sales forecast to be expanded and distorted. c. Population and smoking ratio factor, which is a key indicator to total sales forecast. d. Economy developing factors, which include per capita disposable income, per capita GDP and Gross industrial output value, etc. In forecasting process, comprehensiveness of these factors must be considered and their duplicate related must be avoided. Second, retail terminal factors: mainly include the influence of counterfeit cigarettes and unauthorized retailers. Though both forms are non-normal-form and the law enforcement capabilities of Chinese Tobacco are enhanced all the time, but they really exist and considerably impact the forecast. Last, marketing policy factors: including brand programming, sales and market structure controlling, etc. From above, the accuracy of tobacco annual forecast is not only depended on the quantitative statistic forecast methods but also influenced by human experience and the qualitative factors.

Medium term prediction means semi-annual forecast, which based on annual forecast results and the quarterly impact factors. Combined with the market departments' experience, semi-annual forecast predicts single brand sales.

The results can help brand marketing programming and adjusting. In practice, it directs the semi-annual supply agreements between industrial enterprises and commercial companies.

Short term sales prediction includes monthly, weekly and daily forecast. The objects of monthly forecast are both total forecast and single brand forecast. Monthly forecast provides support to monthly supply and demand planning between industrial and commercial. Because the holidays such as "May 1st", "October 1st" and Spring Festival greatly disturb tobacco sales, the monthly impact factors must be considered. Weekly and daily forecasts are single brand forecast. The results are foundation to inventory management, sorting scheme, marketing service management and logistic delivery management. Because of different brand and different origin, the sales are greatly divers. So the brand characteristic analyzing is decisive to weekly and daily forecasts.

From above, we can see that the factors influence tobacco sales are complex. How to consider them integrally and closely analyze the related data, finally, select the applicable models and methods to each type of prediction is the key point. In practice, except for analyzing the operation database, market departments' internal investigate and market investigate must be carried out to grasp the market characteristics. All the data must be eliminated noise and distortion.

3 Tobacco Sales Forecast Models

In this paper, the models and algorithms are proposed specially to monthly tobacco sales forecast. We use real sales data from Jan 2002 to Mar 2007 to test our method, which offered by one municipal tobacco commercial company, Zhejiang province. Through observation, we find that monthly sales have long-term growth trends and near the Spring Festival present seasonal fluctuation. To such complex system which has double characteristics, the classic moving average and regression methods are unable to reach satisfactory results. Therefore, we propose a combined forecast model based on Group Method of Data Handling (GMDH) and Auto Regression Integrated Moving Average (ARIMA).

3.1 ARIMA model

ARIMA is the improved form of the Box-Jenkins method [1], which get satisfying results in short term time series forecasting. The basement of ARIMA is that the unstable auto correlative time series being changed to stable one. Statistic methods such as difference are used to do this change. Then, the data characteristics can be grasped through regression analyze. ARIMA is the application of regression thinking in dynamic time series analyzing. To time series y_t, the general model of ARIMA is in the form of

$$y_t = \theta_0 + \sum_{i=1}^{p} \varphi_i y_{t-i} - \sum_{j=1}^{q} \theta_j y_{t-j} + \varepsilon_t \qquad （1）$$

Which is known as ARIMA (p, d, q). It assumed to be a linear function of several past observations and random errors. Where y_t is the actual value and ε_t is the random error at time period t, respectively. φ_i are self regression coefficients. p is the regression orders as integer. θ_j are moving average coefficients. q is moving average orders. $\varepsilon_t \sim N(0，\sigma^2)$ are random errors. d is one length difference times. If series have seasonal characteristic, seasonal difference must be done, then the model become ARIMA (p, d, q) (P, D, Q)$_s$. Where P,D,Q are the corresponding coefficients and s is the seasonal difference length.

ARIMA process includes three iterative steps:

Step1. Pattern recognition and parameter estimation. In this stage, time series' auto regression characteristics are corresponded to empirical model. The essential is to let the series stationary and estimate model's coefficients such as p, d, q and P, D, Q. Series' autocorrelation function and partial autocorrelation function analyzing are generally used methods. The initial model usually needs to be adjusted several times through further experiments.

Step2. Diagnostic check of model adequacy. After recognition and estimation, the model has to be adequacy checked. If the residuals examined as white noises, it shows that the noise is a pure stochastic series and the model contains almost all the trends and fluctuations. So it can be used in forecasting. Otherwise, a new tentative model should be identified.

Step3. Prediction. To different data, the above steps maybe need to be repeated several times until satisfactory model is finally selected. When the model is accepted and used in practice, the forecast results must be checked and analyzed closely.

3.2 GMDH model

The basement of GMDH[2] is that one time series can be entirely described by Kolmogorov-Gabor polynomial:

$$y(t) = \sum_{i=1}^{N} a_i x_i + \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij} x_i x_j + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} a_{ijk} x_i x_j x_k + \ldots \qquad （2）$$

Theoretically, when the data are enough, all the parameters of this polynomial can be fitted. But practically, following the increasing of variables numbers and function powers, the calculation times increase too rapidly to reach a result. GMDH algorithm resolves this problem, which has outstanding capability in complex system forecasting.

GMDH builds the mathematical model through simulating the process of organism evolution, which is based on the thinking of ANN (Artificial Neural Network) and GA (Genetic Algorithm). The model constructer only needs to give the test data with their grouping ways and the external criterion to check the intermediate models. At first iteration, based on initially simple regression analysis, the GMDH algorithms analyze the input units of test data and automatically across them to synthesize new active nerve cells for next iteration. The high-degree regression polynomial models are produced by this net. Then, use the external criterion to filter these polynomials to derive more accurate presentations in the next iteration. These iterations repeat until the new model's external criterion value increase which means the best complexity model has reached [3]. GMDH is superior to ANN in model explain because of its regression analyze principle. With the reducing of parameters need to be artificial decided, human biases and misjudgements also be avoided. Partial induction method let GMDH have better performance in search efficiency and time complexity than GA.

The internal and external criterions used in whole model identification process are the most notable characteristic of GMDH. Fist, the dataset is divided into training set and test set. In iterations, the internal criterion is used on training set to produce under selected models and the external criterion used on test set to choose the model with best data fitting capacity. With the repeat of these processes includes genetic variation, selection and evolution, the complexity of models increases all the way until the best model reached. In practice, the partial quadratic polynomials in the form:

$$Y = A + Bx_i + Cx_j + Dx_i^2 + E_j^2 + Fx_i y_j \text{ (i≠j)} \quad (3)$$

often be used to represent Kolmogorov-Gabor polynomial and to data fitting [4].

4 The Realization of Forecasting

Recent five years real sales records coming from one municipal tobacco commercial company, Zhejiang province, are used in our study. After data filtering, clear outing and other checking, we get sixty monthly samples. The models of ARIMA and GMDH are constructed separately, and then a hybrid model based on them is designed. After applicability checking, use these models to forecast tobacco sales of former three months 2007. The powerful statistic software SAS is used to realize ARIMA model and VC++ system is used to write GMDH algorithm.

4.1 Process of ARIMA forecast

Monthly tobacco sales data appear obviously long-term growth trends and seasonal fluctuations. After one length time difference and one length seasonal difference, series become stationary. Then we have d=D=1 initially. With monthly data, s=12. The Figure s of autocorrelation and partial autocorrelation functions exhibit the complexity of the series. In ACF Figure , we find the points k=35 and k=36 exceed the confidence interval, so we assume that p, q∈[0,1,2] and P, Q∈ [0,1,2,3]. Then, we get the models group in form of ARIMA $(p,1,q)(P,1,Q)_{12}$.

In our experiment, AIC criterion is used to assistant decide the coefficients of model [5]. We select ARIMA$(1,1,1)(1,1,1)_{12}$ and ARIMA$(2,1,1)(1,1,1)_{12}$ whose value of AIC are lower.

These models also past t check and the values of R-Square check are 0.961 and 0.953 which mean the fitting capacities of models are acceptable. From the ACF Figure s of residuals series e_t of these models, the autocorrelation coefficient of e_t are obviously near zero which show that the residuals series is white noise series and the models are acceptable to forecasting. The mean absolute percentage error (MAPE) of ARIMA$(1,1,1)(1,1,1)_{12}$ and ARIMA$(2,1,1)(1,1,1)_{12}$ are 5.58 and 7.15. It is obvious that the former is more accurate. In practice, other models are also checked with sales dataset and the MAPE are all higher than our selected model. This also testifies our selection from another side.

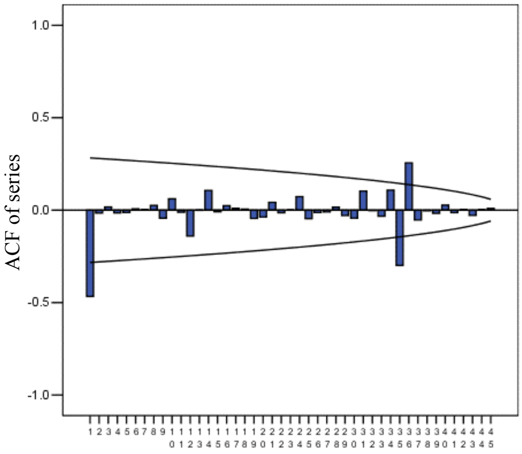


Figure 1 Autocorrelation function analyze

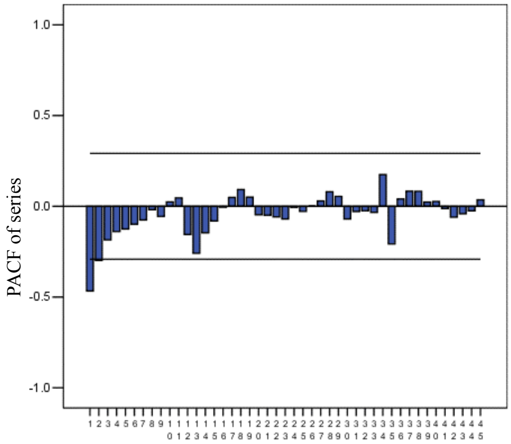


Figure 2 Partial autocorrelation function analyze

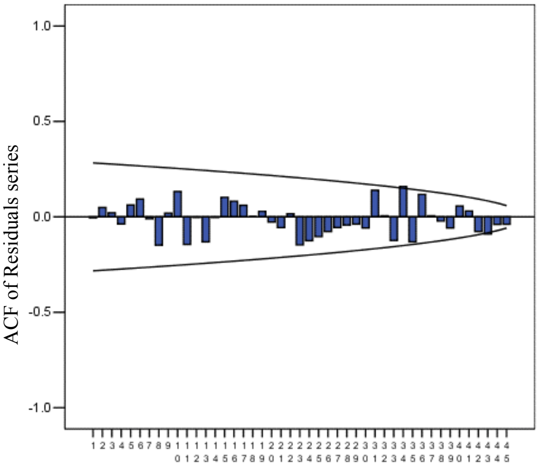


Figure 3 Autocorrelation analyze of residuals series

4.2 Process of GMDH forecast

The GMDH net has six input units; it means the former six month data are used to forecast the next month sales, so, 54 data groups are gotten. We select former 44 groups as training set and the residual 10 groups as test set.The steps of GMDH algorithm:

Step1. Dataset θ (the number of data are N_θ) are divided into training set A and test set B and forecast set C, which satisfy

$$\theta = A \cup B \cup C, N_\theta = N_A + N_B + N_C \qquad （4）$$

The dataset matrix is

$$X_\theta = \left[X_A X_B X_C \right]^T \qquad （5）$$

The export vector is

$$Y_\theta = \left[Y_A Y_B Y_C \right]^T \qquad （6）$$

We use the formula in form of

$$x_i^{'} = \frac{x_i - \overline{x}}{\overline{x}} \qquad （7）$$

to standardize data, where $\overline{x}$ is the mean of dataset, i=1,2,…,n.

Step2. The partial quadratic polynomial as Eq.（3）is used in data fitting. Each pair of input variables from test set A are used in parameters estimation and the coefficients of polynomial are computed through least square method.

Step3. All the polynomials generated are checked by test data set B with the external criterion. In our study, the regularity selection criterion based on the root mean squared error r_k is used for this purpose, where

$$r_k = \left| \sum_{i=1}^{N_B} (y_i - z_{ki})^2 / \sum_{i=1}^{N_B} y^2_i \right|^{1/2} \text{(k=1,2,…,m(m-1)/2)} \quad （8）.$$

In which, z_{ki} is forecast value of y_i; r_s is eliminate threshold. The polynomials and intermediate variables meet $r_k < r_s$ are hold, which used as initial value to next iteration. The minimum r_{min} is recorded each iteration.

Step4. Repeat step2 and step3. r_{min} of each iteration is checked. When r_{min} increasing, the iteration is stopped. The model selected at last iteration is the optimal nonlinear regression model. Use the model to forecast.

4.3 Process of combined forecast

Because the principle and method is different, Single forecast method is distinct in data mining viewpoint and deepness. So, when unique methods are combined together, the new hybrid one may get more usable information from the data. The combined forecast method just based on this thinking. 1969, Bates and Granger [6] proposed the combined forecast method, after that, this theory become research hotspot. Despite someone was doubtful to the validity of combined forecast, the affirmative attitudes were in majority [7] [8]. The typical view was like Hibon's [9], through mass experiments, they got the conclusion: the result of combined forecast must not be superior to the best individual one. But in practical application, individual method can hardly grasp all characteristics of the data, therefore, when we don't know which one is the best, selecting among combinations leads to a choice that has, on average, significantly better performance than that of a selected individual method.

The key points of combined forecast are the selection of individual methods and the combined weights of each one. Individual methods selection must base on the data catachrestic analysis and experiments. There are two ways of weights selecting. First one is through linear and nonlinear programming and this also divided into fixed weight and dynamic weight methods. Second one is using the result of individual forecasts as the inputs for complex methods such as ANN. The weights are calculated by the ANN. When imported the complex method, it's also increase the complexity of model design. Besides, though model's fitting capability to observed values is greatly enhanced, the model become easy to be overfitting and therefore unlikely to behave well on new data. So, we adopt the first method to calculate combined weights.

At model construction process, assuming Y_t is the actual value of observation samples where t=1,2,…,n; the respective forecast values of respective ARIMA and GMDH is $Y_{1t}$and Y_{2t}; combined forecast values is ;

$$\hat{Y} = k_{1t}Y_{1t} + k_{2t}Y_{2t} \qquad (9)$$

k_{1t},k_{2t} are the combined weights of ARIMA and GMDH, which meet $k_{1t},k_{2t} \geqslant 0$ and $k_{1t}+ k_{2t}=1$; The absolute error is used as objective function, then the linear programming model to calculate the combined weights is:

$$\min\sum_{t=1}^{n}|e_t| = \sum_{t=1}^{n}|Y_t - \hat{Y}_t| = \sum_{t=1}^{n}|k_1(Y_t - Y_{1t}) + k_2(Y_t - Y_{2t})|$$

$$\text{s.t. } k_{1t}+ k_{2t}=1 \text{ and } k_{1t},\ k_{2t} \geqslant 0. \qquad (10)$$

To remove the absolute value in the function, following transformation is done:

$$u_t = \frac{1}{2}|k_1(Y_t - Y_{1t}) + k_2(Y_t - Y_{2t})| + \frac{1}{2}\left(k_1(Y_t - Y_{1t}) + k_2(Y_t - Y_{2t})\right) \qquad (11)$$

$$v_t = \frac{1}{2}|k_1(Y_t - Y_{1t}) + k_2(Y_t - Y_{2t})| - \frac{1}{2}\left(k_1(Y_t - Y_{1t}) + k_2(Y_t - Y_{2t})\right) \qquad (12)$$

The model is changed to:

$$\min\sum_{t=1}^{n}|e_t| = \sum_{t=1}^{n}|Y_t - \hat{Y}_t| = \sum_{t=1}^{n}(u_t + v_t)$$

$$\text{s.t. } k_{1t}+ k_{2t}=1 \text{ and } k_{1t},\ k_{2t} \geqslant 0. \qquad (13)$$

At forecast process, there are no observation values, but the combined weights can get from the results calculated in model construction process. Assuming forecast steps are m (m=1,2,…), then the weight calculation formulas of ARIMA and GMDH are as follow:

$$k_{1m} = \frac{1}{n}\sum_{t=1}^{n+m-1} k_{1t},\ k_{2m} = \frac{1}{n}\sum_{t=1}^{n+m-1} k_{2t} \qquad (14).$$

4.4 Results

The forecast results list in Table 1. We use PE (Percentage Error) to examine the accuracy of single point fitting and prediction and use MAPE (Mean Absolute Percentage Error) to compare different forecast methods [10].

From the results, the MAPE of ARIMA and GMDH is not greatly distinguished. But GMDH gives better forecast to monthly disturbance than ARIMA forecast. The reason is that based on the assumption of the data series is linear, ARIMA is suitable to find the linear characteristics, while GMDH is more susceptive to nonlinear characteristics. The accuracy of proposed combination forecast model is highest. The proposed method can fit the need of forecasting process and is an effective way to tobacco sales prediction.

Table 1 Forecast results and Comparison of methods (part)

Period	Sales	ARIMA(111)(111)$_{12}$		GMDH		Combined forecast	
		Forecast	PE	Forecast	PE	Forecast	PE
2005.1	16020	15996	0.15	17256	-7.71	15996	0.15
2005.2	11276	10677	5.31	9970	11.58	10555	6.40
2005.3	12132	12399	-2.20	12740	-5.01	12454	-2.66
2005.4	12231	12580	-2.85	12701	-3.84	12623	-3.20
2005.5	12037	12099	-0.52	11374	5.51	12093	-0.46
2005.6	12404	12569	-1.33	11960	3.58	12495	-0.73
2005.7	12147	12512	-3.00	12292	-1.20	12322	-1.44
2005.8	12470	12791	-2.57	12787	-2.55	12789	-2.56
2005.9	12359	13124	-6.19	12871	-4.14	12949	-4.77
2005.10	12286	12955	-5.44	12598	-2.54	12661	-3.06
2005.11	12333	12948	-4.99	13087	-6.12	13004	-5.44
2005.12	10780	12263	-13.7	11966	-11.0	11680	-8.35
2006.1	20076	17060	15.02	17840	11.14	18113	9.78
2006.2	10459	12002	-14.7	10656	-1.88	10677	-2.09
2006.3	12341	13226	-7.17	13319	-7.92	13268	-7.51
2006.4	12514	13271	-6.05	13044	-4.24	13119	-4.83
2006.5	12673	12692	-0.15	12784	-0.88	12695	-0.17
2006.6	12705	13067	-2.85	13181	-3.74	13108	-3.18
2006.7	12718	13028	-2.44	12875	-1.23	12906	-1.48
2006.8	13133	13354	-1.68	13249	-0.88	13272	-1.06
2006.9	13529	13312	1.60	13040	3.62	13267	1.93
2006.10	13298	13154	1.08	12880	3.14	13125	1.30
2006.11	13123	13333	-1.60	12978	1.10	13092	0.23
2006.12	11549	11995	-3.86	11211	2.93	11497	0.45
2007.1	20498	19587	4.44	19730	3.75	19836	3.23
2007.2	15101	11741	22.25	14315	5.20	14335	5.07
2007.3	12822	13235	-3.22	13097	-2.14	13093	-2.11
MAPE		5.58		5.35		3.66	
MAX PE		16.55		11.64		9.78	
MIN PE		-14.75		-11.00		-8.35	

5 Conclusions

It's hardly to decide which method suitable for the real time series forecasting because it's difficult to confirm the series is linear or nonlinear. In practice, a great deal of experiments must be tried to select more preferable model. Furthermore, coming from the complexity of real world, time series always mix both linear and nonlinear trend and influenced by other unexpected factors, therefore, single forecast model usually can't adequately get all the characteristics of series. The combined models always have better performance as take advantage of the unique strength of single model. Our research is just enlightened by this thinking and the experiment results correspondingly confirm it.

Tobacco monthly sales present linear and nonlinear hybrid complexity. Because it influenced by lots of factors such as tobacco industrial policies, marketing specialty and seasonal requirements, and so on. In forecasting process, to obtain more accurate forecast results, it's important to select appropriate models. Besides, the combination of human experience is also a noticeable approach.

References

[1] G.E.P. Box, G. Jenkins, Time Series Analysis, Forecasting and Control, Holden-Day, San Francisco, CA, 1970

[2] Nikolay Y. Nikolaev, Hitoshi Iba, “Polynomial harmonic GMDH learning networks for time series modeling”, Neural Networks, Vol.16, No.10, December 2003, pp.1527~1540

[3] A.G.Ivakhnenko and G.A.Ivakhnenko, “The review of problems solvable by algorithms of the group method of data handling”, Pattern Recognition and Image Analysis, Vol5, No.4, 1995, pp. 527~535

[4] R.E. Abdel-Aal, “Predictive modeling of mercury speciation in combustion flue gases using GMDH-based abductive networks”, Fuel Processing Technology, Vol.88, No.5, May 2007, pp. 483~491

[5] Clifford M.Hurvich and Chih-Ling Tsai, "Bias of the corrected AIC criterion for underfitted regression and time series models", Mathematics & Physical Sciences, Vol. 78, No. 3, 1991, pp. 499~509

[6] J.M.Bates and C.W.J.Granger, “The Combination of Forecasts”, Operations Research Quarterly, Vol. 20, No. 4, December 1969, pp. 451~468

[7] Thomas D. Russell and Everett E. Adams, Jr, "An Empirical Evaluation of Alternative Forecasting Combinations", Management Science, Vol. 33, No. 10, October 1987, pp. 1267~1276

[8] Ilan Fischer and Nigel Harvey, "Combining forecasts: What information do judges need to outperform the simple average?", International Journal of Forecasting, Vol. 15, No. 3, July 1999, pp. 227~246

[9] Michěle Hibon, Theodoros Evgeniou, “To combine or not to combine: selecting among forecasts and their combinations”, International Journal of Forecasting, Vol. 21, No. 1, January-March 2005, pp. 15~24

[10] H.W.Mui and C.W.Chu, "Forecasting the spot price of gold: combined forecast approaches versus a composite forecast approach", Journal of Applied Statistics, Vol. 20, No. 1, 1993 , pages 13~23

Empirical Study on Stock Return of Shanghai through Arch with DQPSO Algorithm

Halqam Ablat[1] Wenbo Xu[2]

1 School of Mathematics and Information Technology, Xinjiang Education Institute Urumqi, Xinjiang, China

E-mail:halqam@tom.com

2 School of Information Technology Southern Yangze University, Wuxi , jiangsu, China

E-mail:xwb@sytu.edu.cn

Abstract

Due to the disadvantages of traditional estimating methods of ARCH model, we estimate the parameters in ARCH model accurately with a new enhanced quantum-behaved particle swarm optimization algorithm with diversity. Then the ARCH model for stock return is established empirically with algorithm and forecast of the return is given.

1 Introduction

Most generally the traditional econometrics is taken as a major tool to research the mathematical finance, a big number of models of which is based on the assumption that sample satisfy the conditions of homoscedasticity, yet, along with the development of finance theory and deepening forwarding of demonstration, the assumption is thought to be unreasonable. For instance, in some actual economic objects as Shanghai Composite Index, there is heteroskedasticity in the revenue. If using homoscedasticity method to calculate and deduce, it would produce serious errors. Meanwhile, the relative intensive changes during the variances of a number of economic objects always occurs in some period of time, and the smaller scale change is centered in anther period, this is to say there is the feature of "peak and thick tail, waving and clustering." Mr. Engel proved that the root cause is that there is a heteroscedasticity existing. Therefore, how to measure the wave accurately and present the dynamic the revenue heteroscedasticity has profound significances. In 1982 Engel carried out a Inflation on Britain and brought out ARCH (Auto-Regressive Conditional Heteroskedasticity Model) that may be regarded as new model to solve such problems. This paper is a study that intends to use PSO(Particle Swarm Optimization), QPSO (Quantum-behaved Particle Swarm Optimization),and QPSO in the RCH model to research the heteroskedasticity in the revenue of the Shanghai Composite Index. After experiment, DQPSO performance is more excellent than general PS0 and QPSO.

2 Quantum-behaved particle swarm optimization algorithm

PSO was ever brought out early in 1995 by James Kennedy, an American social psychologist and Russell Eberhart, an electrical engineer. The two scientists were ever enlightened by the results of bird group behavior in an emulating environment as well as the biotic population model established by biologist Frank Heppner.

Mr. Sun and several others was published a paper on IEEE CEC 2004 in which the research achievement of convergence proposed by Clerc, etc. are studied, and a new particle evolution model based on quantum behavior is proposed. Taking DELTA potential well as a base, Clerc considered that particle's behavior is similar to quantum's, and brought out the algorithm of the QPSO

2.1 Particle Swarm Optimization Algorithm

PSO algorithm is established on the basis of an enlightenment of aves group behavior, proposed by James Kennedy and Russell Eberhart at the IEEE in 1995, the core idea is that they considered each partial is a substance in the N-dimensional search space without any weight and volume, flying at a certain speed that is regulated according to both individual experiences and group experiences of flight.

$$\begin{cases} X_i(t) = (x_{i1}(t), x_{i2}(t), \ldots x_{iD}(t)) \\ V_i(t) = (v_{i1}(t), v_{i2}(t), \ldots v_{iD}(t)) \end{cases} \quad i = 1, 2, \ldots S \qquad (1)$$

Of which, $X_i(t)$ represents the location of the particle “i” at time “t”; $V_i(t)$ represents the speed of the particle “i” at time “t”; “D” represents the dimensional number of particles; “S” represents the number of particles .The speed and location of particle is refreshed through （2）as follows:

$$\begin{cases} v_{ij}(t+1) = v_{ij}(t) + c_1 r_{1j}(t)(p_{ij}(t) - x_{ij}(t)) \\ \qquad + c_2 r_{2j}(t)(p_{gj}(t) - x_{ij}(t)), \\ \qquad v_{ij} \in [-v_{\max}, v_{\max}] \\ x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \end{cases} \qquad (2)$$

Of which c_1 and c_2 are compressibility factors; r_1 and r_2 are random numbers of D-dimensions between 0 and 1. At 1998 IEEE, Mr. Y. Shi and R.C. Eberhart proposed to add weighting to the evolution equation, the range of which is fro 0 to 1.4. Therefore the Formula 2 becomes as follows:

$$\begin{cases} v_{ij}(t+1) = v_{ij}(t) + c_1 r_{1j}(t)(p_{ij}(t) \\ \qquad - x_{ij}(t)) + c_2 r_{2j}(t)(p_{gj}(t) \\ \qquad - x_{ij}(t)), \ v_{ij} \in [-v_{\max}, v_{\max}] \\ x_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1) \end{cases} \qquad (3)$$

Particle searches an overall optimization by means of changing its own location. From the Formula 3 we may see that the speed gives the difference of location this time and next time, and the change of speed depends on the most optimized value of location this time, the overall group, and the most optimized difference of particle this time. Therefore the particle evolution depends ①, the experiences of particle of position and speed in the past, ② individual cognitive ability (particles itself optimal value), and ③ social acceptability (the optimal value of PSO), the three parts. It is the sharing group information and summary of self-experiences of particle that particle is evolved in the complete search space and the optimal solution is reached.

At present, most of study on PSO algorithm takes inertia weight PSO algorithm as a base, on which it is extended and revised. Thereupon, most of documents and materials the inertia weight PSO algorithm is called as a Standard PSO algorithm (SPSO).

2.2 QPSO Algorithm

PSO is a metapopulation based evolution search technology, yet, all fundamental or modified PSO algorithms can ensure not an overall convergence because the evaluation formula of PSO lets all particles search in a limited sample space. Sun and the others, inspired by the characteristics of convergence of PSO as well as the basic theory of quantum mechanics, brought out QPSO algorithm (Quantum-behaved Particle Swarm Optimization). This is a strategically change to the PSO search algorithm. The new formula needs no speed vector, and less parameter, simpler, easier to be controlled. The QPSO algorithm is much better preceded in every aspect than the PSO.

In order to ensure convergence of the algorithm, each particle must be converged onto each self-location p, $p = (p_1, p_2, \ldots p_d)$, the d-Dimensional coordinate of the particle “i” at point “p” is as follows:

$$p = (\varphi_1 * p_{id} + \varphi_2 * p_{gd}) / (\varphi_1 + \varphi_2)$$

$$\varphi_1 = rand(0,1) \qquad \varphi_2 = rand(0,1)$$

$$\text{Or} \quad p = \varphi * p_{id} + (1-\varphi) * p_{gd}, 0 < \varphi < 1 \qquad (4)$$

We introduce an overall point mbest into PSO to calculate the next step of variable L which is defined as the average vale of the best locations among all particle locations, the Formula is as follows:

$$mbest = \frac{1}{M}\sum_{i=1}^{M} P_i = \left(\frac{1}{M}\sum_{i=1}^{M} P_{i1}, \ldots\ldots, \frac{1}{M} P_{id}\right) \qquad (5)$$

Of which M is the number of particles; p_i is partially the best location of the particle “i”, then the Formula is as follows:

$$L(t+1)=2*\beta*|mbest-x(t)| \quad (6)$$

Therefore, particle iterative equation becomes:

$$x(t+1)=p-\beta*|mbest-x(t)|*\ln(1/u) \quad (7)$$

$x(t+1)=p+\beta*|mbest-x(t)|*\ln(1/u)$ of which u = rand (0,1) （8）

Of which β is called contracting and expanding coefficient, and regulating its value may control convergence rate. Generally speaking, when β is reduced from 1.0 to o.5, it may reach better results, that is

$$\beta = (1.0\text{-}0.5)*(\text{MAXITER-T})/\ \text{MAXITER}+0.5 \quad (9)$$

Of which MAXITER is the max., Formula 7 and Formula 8 are known as the quantum behavior of particle swarm algorithm, referred to as the QPSO.

Recently, Mr. Sun, et al. proposes the Diversity-Guided Model of QPSO, the basic principle of which is that the algorithm performance is improved by means of adjusting the by adjusting the diversities among PSO.

3 An introduction to arch model

In the ARCH(p)model, the revenue sequence X_t is expressed as follows:

$$X_t=\sigma_t\varepsilon_t \quad (10)$$

Of which, the usually assumed disturbance is ε_t ~N(0,1)i.i.d, and conditional heteroskedasticity difference is linear function of p period of disturbance delaying.

$$\sigma_t^{\,2}=\alpha_0+\sum_{i=1}^{p}\alpha_i X^2_{\,t-i} \quad (11)$$

In order to make Formula 11 have a clear process of definition, $\sigma^2_{\,t}\geq 0$ and X_t must have a clear definition, a sufficient condition is as below:

$$\alpha_i \geq 0 \qquad \text{i=0,…p}$$

And

$$\alpha_1+\ldots+\alpha_p<1$$

ARCH model is a breakthrough to the traditional econometrics of the limitation of homoscedasticity and is more accurate for describing the process between the reality and the risks possible in finance market.

4 Traditional algorithm in ARCH model and its shortage

From the ARCH model, we may know that $X_t|\xi_{t-1} \sim \text{N}\,(0,\sigma_t^{\,2})$ has The conditional probability density function described as below:

$$f(x_t|\xi_{t-1})=(2\pi\sigma^2_{\,t})^{-\frac{1}{2}}\exp(-\frac{1}{2\sigma^2_{\,t}}x^2_{\,t}) \quad (12)$$

Of which $\sigma_t^{\,2}=\alpha_0+\sum_{i=1}^{p}\alpha_i X^2_{\,t-i}$. Substitute σ_t^2 with（12）to gain:

$$f(x_n,\cdots,x_1|x_0)=f(x_n|x_{n-1},\cdots,x_0)\ldots f(x_2|x_1,x_0)f(x_1|x_0) \quad (13)$$

Getting math. log from both sides of the equation, log-likelihood function may be gained as below:

$$\log f(x_n,\cdots,x_1|x_0)=\sum_{t=1}^{n}\log f(x_t|\xi_{t-1}) \quad (14)$$

The estimating method of econometric parameter model is no more than MLE and ME, the two categories, and in the ARCH model, the parameter estimating adopted is mainly BHHH algorithm and GMM method.

The above Formula （14）may be optimized by MLE, but MLE requires a given variable of objective prior probability distribution, which is often unrealistic. Although traditional ME has overcome the restrictions on distribution it is often too rough and does not have good statistical nature.

Traditional numerical method is generally poor in convergence. The BHHH algorithm proposed by Mr. Berndt and GMM brought out by Hanson have an improvement in consistency and progressive normality, yet, both would come to the optimization of a certain multivariate nonlinear function. To treat it, calculus feasible direction method is mostly adopted, which may cause defects as follows：

（1）Poor robustness. If the objective function is not smooth and intermediate results shake, it would cause a complete failure of the algorithm;

（2）Single-mode search space is easy to cause partial converge optimum solution;

（3）Small dimension of search space and less efficacy.

5 PSO improvement and DQPSO improvement of ARCH model

5.1 PSO Improvement of ARCH Model

This paper employs PSO to solve the optimization problem, and the specific algorithm processes using PSO to estimate ARCH model are as follows:

（1）Set initial location and speed of PSO and calculate the adaptive value of each particle, and assume that all are at the best positions, and record as pbest. Compare those adaptive values and record the locations with best adaptive values to be current best overall positions, expressed as gbest

（2）PSO evolutionary process conducted on the evolution and adaptation of each of the particles, if they are good in the best position pbest current, updated pbest value, in the best position to identify the optimal location of the particles, if it good in the current global best position gbest, updated gbest value.

（3）If current results has reached pre-set accuracy or iterative number have reached pre-set circling number, stop iteration and input the optimal solution, otherwise turn to （2）.

5.2 DQPSO Improvement of ARCH Model

This paper employs DQPSO as a solution of optimization, and the specific algorithm processes to estimate ARCH model are as follows:

（1）In solution space, initialize location value x (i) of a particle swarm, the size of the swarm is M;

（2）Calculate diversities of particle swarms. Evaluate adaptive value of individual particle by means of optimized value. If current information is more excellent than previous best value of an individual, replace current value with individual best vale (pbest), on the contrary, not replace it;

（3）Evaluate adaptive values of all particles by optimizing function, and obtain gbset. Calculate mbest.

（4）Make use of Equation [6], Equation [7] and Equation [8] to refresh particle information.

（5）Check whether it reaches the pre-set beset value or maximum accuracy, if not, return to （2）, and on the contrary, end the iteration.

Form above we may see, in the DQPSO, it needs only position vector to describe particle state, and in the algorithm there is only one flare factor β, selection and control of which is vital, which relates to the convergence property of overall algorithm.

6 PSO and DQPSO use of the proceeds of index cards on the empirical estimates ARCH

The model researched in the paper is ARCH （2）. The experimental data from closing stock index of Shanghai Composite Index that is expressed as I_t, the sampling interval is based on week and the span from January 5, 1996 to September 29, 2005, the sample volume is 485, and the revenue is defined as $X_t = \ln I_t - \ln I_{t-1}$, and the simple statistical properties are shown as following Table 1:

From Table 1 we may know that revenue close to the traditional normal distribution and maximum likelihood function value of MLE is 313.67.

The purpose of this paper is to make used of PSO and DQPSO to conduct parameter estimation and compare the results with the results of basic PSO algorithm and improved algorithm. The specific experimental parameters are as follows:

Table 1 Statistical Properties of Earnings Data

X_t	
Sample mean	0.0016
Sample variance	0.0013
Maximum	0.1606
Minimum	-0.2263
Kurtosis	4.9060
Skewness	-0.1340

Let number of particle swarms be 30, initial speed is randomly produced between -10 to +10, initial location info is produced randomly between 0 to 1 or setting it to be 1.9, Iterations is set to be 100. Of which w in inertia weight is set to linear reducing from 0.9 to 0.4, and, for the QPSO algorithm it is set to be a linear reducing from 1 to 0.5, thus the experimental results are as follows:

Table 2 Experimental Result

	Near normal distribution	Basic PSO	Inertia Weight PSO	DQPSO
Iterations	None	100	100	100
Parameter α_0	None	0.0017289	0.0015718	0.0015186
Parameter α_1	None	0.51417	0.45546	0.66719
Parameter α_2	None	0.20168	0.33899	0.42168
Maximum Likelihood function	313.67	335.26115	335.62900	337.215

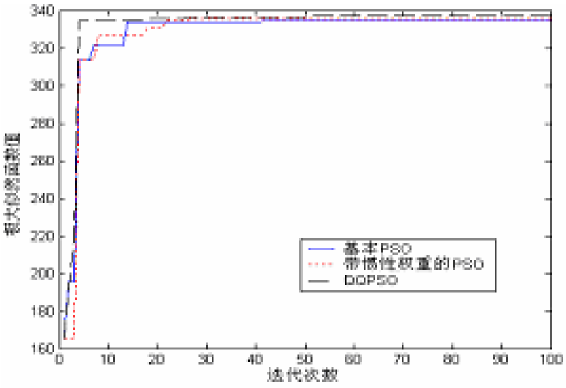


Figure 1 Value Change of Maximum Likelihood function in Iterative Process

From Table 2, the experimental results may reach the conclusions as follows:

（1）From Table 2 we may know, when checking maximum likelihood, and making use of DQPSO, basic PSO and inertia weight PSO, the results of ARCH model obtained are all better than normal, and the results of DQPSO obtained is the most optimal. This is to say that ARCH model is capable of explaining the non-normal of the peak and thick tail of revenue of Shanghai Composite Index.

（2）From Figure 1 we may know that the iteration is about 50 times, and the maximum likelihood in the three experiments are a fixed converging value, this is to say the speeds of the three algorithms are fast and meanwhile we may know the convergence speed of DQPSO is the fastest and the best one.

Table 3 Actual Value and Estimated Value of Experiment

	2005-10-14
Actual Value	1139.55
Basic PSO	1129.47
Inertia Weight PSO	1132.58
QPSO	1141.04

From Table 3 we may know that, when conducting actual value estimating, the results from the model that makes use of DQPSO to estimate is better than basic PSO as well as inertia weight PSO. Although it is impossible to estimate each value change accurately, we can obtain approximately value change trends.

7 Conclusion

In this paper, the author of this paper makes use of DQPSO, PSO, and improved algorithm to establish a ARCH model for Revenue of Shanghai Composite Index. The different algorithms are adopted to estimate accurately the parameters in the models. The meliority of DQPSO and improved algorithm is verified. Also, the paper, making use of estimating model obtained, conducts a forecast on index revenue and the results of which is approximately close to the actual trend. This is to say the ARCH model is indeed capable of describing heteroskedasticity of securities market.

References

[1] Chen Yiheng , Financial Data Analysis and Time Series [M], Beijing China Statistics Press, 2004. P. 112-- 113

[2] Mandelbrot B.The variation of certain speculative prices. Journal of Business[J]1963.94~419

[3] Zeng Jianchao, Jie Jing, Cui Zhihua , PSO [M], First Edition, Beijing Science Press, 2004

[4] Sun, J. and Xu W.B.. A Global Search Strategy of Quantum-behaved Particle Swarm Optimization[C]. Proceedings of IEEE conference on Cybernetics and Intelligent Systems. 2004,111～116

[5] J Kennedy, RC Eberhart.Particle Swarm Optimization[J].

Proceedings of the IEEE International Joint Conference on Neural Networks,1995,4:1942～1948

[6] Angeline, P.J..Using Selection to Improve Particle Swarm Optimization[C]. Proceedings of IEEE International Conference on Evolutionary Computation,1998: 84～89

[7] Clerc , M., Kennedy, J.. The Particle Swarm:Explosion, Stability and Convergence in a Multi-Dimensional Complex Space[J]. IEEE Transaction on Evolutionary Computation, 2002（6）: 58～73

[8] Wenbo Xu, Jun Sun: Adaptive Parameter Selection of Quantum-Behaved Particle Swarm Optimization on Global Level. ICIC （1）2005: 420-428

[9] Jun Sun,Bin Feng,Wenbo Xu. Particle Swarm Optimization with Particles Having Quantum Behavior[C]. Portland,Oregon :IEEE Press,Proceedings of the 2004 IEEE Congress on Evolutionary Computation,2004.325-331

[10] Sun J, C.H.Lai, Wenbo Xu. Quantum-behaved Particle Swarm Optimization and Its Application. Journal of Computer and Mathematics with Applications, U.K

Research on E-Learning Effects Evaluation Based on Information Entropy

Lan li[1] Minjie xiao[2] Liu rong[3]

1 Software college NanChang University East Nanjing No 235, Nanchang, Jiangxi, P.R.China
lilan@ncu.edu.cn
2 Information center Local Taxtaion of Nanchang Bureau No.229 Tuanjie street，Xihu district, Jiangxi province

3 Training center Local Taxtaion of Nanchang Bureau No.229 Tuanjie street，Xihu district, Jiangxi province

Abstract

Although there are a large amount of platforms and tools providing distributed online learning services, it is rather difficult to identify and evaluate which solution gets the best effect. Therefore, it is definitely necessary to make assumptions about e-learning effects in distributed system which are not explicitly available. One possible choice is to assume that the better e-learning effects are, the more difficult questions web-based learners can answer. In this paper, a feature approach from the point of entropy, according to students' answers to on-line tests and exams, is proposed. Results show that it is a dependable and practical method for evaluating the e-learning effects especially for multiple choice questions.

Keywords: Information entropy, e-learning effects, decision tree, discrete variable, leaf node

1 Introduction

Today's learners, immersed in a rapidly changing environment, need to keep pace with large amount of technological information. Consequently, learning systems must evolve and adapt to meet learners' requirements and be able to face up to an extensive, massive and diversified learning demand. It's difficult for traditional learning systems to satisfy those demands. However, E-learning could be a solution to this problem. During the last several years, web-based learning has played an important role in the pedagogical field and emerged as one of the fastest-moving trends in education.

However, it is obvious that the effect, acceptance and understanding of web-based learning is difficult to evaluate. In order to realize and to verify the learners' acceptance to the e-learning, we used the entropy model and decision tree to gain understanding of the e-learning system effects. As we all known, in the past years, the teacher was expected to act as a coach; they diagnoses mistakes and supports students. Nowadays, students who take courses on-line can access the courses whenever and wherever convenient. They can download the lecture note from the Web, communicate with each other and their instructor through e-mail, and took exams by responding to questions on computer screens. In this article, we proposed an entropy based analysis evaluation approach, which works on the web-based exams and tests, to quantity the e-eduating and the e-learning effects.

The paper is structured as follows: Section 2 explains the fundamental idea of entropy and its interpretations. A discussion of implementation details about evaluation of on-line learning effects is given in section 3. Sections 4 give concluding remark and discuss related work.

2 Entropy-Based Evaluation

In information theory, the Shannon entropy or information entropy is a measure of the uncertainty associated with a random variable. We can say that the uncertainty in a set of discrete outcomes is the entropy. Entropy is also related to how difficult it is to guess the value of a random variable. That it is to say, the entropy is the average minimum number of yes-no questions necessary to identify an item randomly drawn from a known, discrete probability distribution. Assume that p is possibility, then we apply our information measure I(p) to indicate the information quantity. Because information is a non-negative quantity, which is also monotonic and continuous, we can derive that I(p) = $\log_b$ (1/p) = -log $_b$(p). If we observe the symbol a_i, we will get $\log(1/p_i)$ information from that particular observation. In a long run of observations, assume that the number of the observations is N. we will see approximately N*p_i occurrences of symbol a_i . Thus, in the N independent observations about symbol a_i, we will get total information I of a_i as following:

$$I = \sum_{i=1}^{n} (N * p_i) * \log(1/p_i) \tag{1}$$

Therefore, the average we get for each observed symbol will be:

$$I/N = \sum_{i=1}^{n} p_i * \log(1/p_i) \tag{2}$$

Since we have defined the information strictly in terms of the probalilities of events, we quantify the information about the observed events. Let us suppose that we have a probability distribution P(p_1,p_2,…,p_n). We define the information quantity, which is called information entropy, about the set of probalilities as follows:

$$H(P) = \sum_{i=1}^{n} p_i * \log(1/p_i) \tag{3}$$

According to the assertion about the entropy discussed above, it is easy to infer that a larger number of potential outcomes have larger uncertainty. It is obvious that the more uncertain we are about an event, the harder it is to guess the outcome. Here we only talk about the on-line test, especially the selection questions.

Figure 1 point out the information content of one bit of data—for example, the answer to a yes-no question. If the answer is certain to be "yes" or certain to be "no," otherwise, it conveys no information. The information content is greatest when the probability of “yes” and “no” are both equal to 0.5. We modify this kind of method to enable a entropy-based measure to determine which test questions are most effective at classifying data into desired categories.

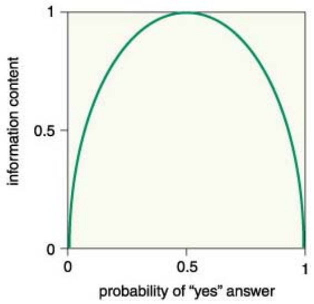


Figure 1 Entropy function.

Suppose we wanted to identify a particular question by serial information on this question’s several choice. Consider the random variable X, which is the correct answer of the question, with following distribution (4):

$$Pr(X = A) = X1,$$
$$Pr(X = B) = X2,$$
$$Pr(X = C) = X3,$$
$$Pr(X = D) = X4.$$

On average, how many yes-no questions will it take a e-learner to figure out the correct answer of the question? We have: Average Number of Yes-no Questions to Determine

$$X = \sum_{i=1}^{4} X_i * |\log_2 X_i| \tag{5}$$

It turns out that the entropy of the distribution given in (4) is exactly equal to (5).we can know that:

H[X]<= Average Number of Yes/No Questions to Determine X <=H[X] + 1[3].

So what is the relation to our e-learning effects? It is easy to figure out and quantify the e-learning effects when we use the entropy. According to the above discussion about entropy, the harder a question is, the more yes-no questions will be provided. The connection between entropy and e-learning effects is that the more difficult questions are more uncertain than normal or easy questions in some respects and a more random in others.

3 Implementations

In order to demonstrate the effects of e-learning on entropy statistics, we will now discuss the algorithm and considerations in the implementation of the proposed entropy-based evaluation. Given a set *T* (whose number of elements is denoted |*T*|) and the ability to ask yes-no questions about a particular element of the set *x*. All we need to know is what is the minimum number of questions we need to ask, that is, we need to determine what *x* is. Intuitively, our answer is $\log_2|S|$. It is obviously that the best we can do at each point is split the data in half, that is, we could immediately identify the object as part of a minority subset, sharply reducing the remaining possibilities. According to that, we applied a decision tree to evaluate whether a question is good or not.

Decision tree is a classifier in the form of a tree structure where each node is either a leaf node, which indicates the value of the target, or a decision node, which specifies some test to be carried out on a single attribute-value, with one branch and sub-tree for each possible outcome of the test.

The decision tree is essentially a branched list of questions in which the answer to one question determines the next question asked. Each question is called a "node" on the tree, and each answer leads to a separate "branch." At the end of each branch lies a "leaf node," where the decision tree assigns a classification to the data.

As we have known, a decision tree is generated from a set of training data, which we can call D. (Initially, D would be the whole training data set; at later steps, it represents only a part of the training data.) Each observation in D has been assigned by an expert to one of several categories, which we denote by C1, C2 and so on. The job of the decision tree is to reproduce this assignment as accurately as possible. To measure how good the effects of a on-line test is , we can use decision tree to figure out the entropy of each question. Once the decision tree is produced, the entropy of the question can be computated. According to this, it is easy to quantify the difficulty of a question. Consequently, it is convenient to evaluate the effects of a web-based test. The decision tree produce process is gave below.

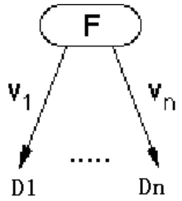


Figure 2 Decision-tree

```
function entropy-based evaluation
Input:    (C1: a set of non-target questions,
               C2: the target questions,
               D: a training set of questions)
returns a decision tree;
begin
If all instances in D are in class C1, create a node C1 and stop, otherwise select a feature or attribute F and create a decision node.
Partition the training instances in D into subsets according to the values of V.
  Recursively apply entropy-based evaluation to subsets {D_i| i=1,2, .., m}
        until they are empty
Compute entropy(D)
end
```

Given a set D, containing only positive and negative examples of some target concept (a 2 class problem), the entropy of set D relative to this simple, binary classification is defined as: Entropy(D) = - $R\log_2 R - F\log_2 F$

where R is the proportion of positive examples in D and F is the proportion of negative examples in D. In all calculations involving entropy we define. To illustrate, suppose D is a collection of 30 examples, including 10 positive and 20 negative examples. Then the entropy of D relative to this classification is Entropy(D)=- $(10/30)\log_2(10/30) - (20/30)\log_2(20/30) = 0.980$.

Otherwise, thus far we have discussed entropy in the special case where the target classification is binary.

If the target attribute takes on *n* different values, then the entropy of D relative to this n-wise classification is defined as: Entropy(D)= $\sum_{i=1}^{c} -p_i \log_2 p_i$, Which is discussed in the above algorithm.

4 Conclusions and Future Work

This paper applied the information entropy principles to describe the decision-tree construction of the on-line test and proceeded to verify the evaluation of web-based learning according to the construction. In the end it replied to the reader by providing the implementation of the target decision-tree. According to the implementation results, the proposed method can quickly and correctly describe the evaluation of the on-line learning. Furthermore, with the use of entropy, the proposed method can check the difficulty of the given selection question, and provide a learning mechanism. Finally, to make the evaluation of the distance learning, some of the characteristics of the questions should be defined in the future. In this way, the proposed method can come up with better and different responses based on these characteristics.

However, every coins have two sides. Although there are many strengths for entropy-based decision tree methods, such as, it is able to generate understandable rules, it performs classification without requiring much computation, it provides a clear indication of which fields are most important for prediction or classification, there still have weaknesses of decision tree methods. The result is often a very large, complex tree that attaches too much significance to the error or noise in the data. Thus, it is necessary for us to continue to make some efforts on the entropy computation algorithm to make sure that the entropy-based decision tree evaluation is quick and noise avoided.

References

[1] Robert M.Gray: Entropy and Information Theory [B] Information Systems Laboratory.Electrical Engineering Department.Stanford University. 2007

[2] Anderson, M. D. . Individual Characteristics and Web -Based Courses[A]. Christ opher R. Wolfe . Learning and Teaching On theWorldWideWeb [C]. San Diego: Academic Press . 2001 .

[3] T. M. Cover and J. A. Thomas. Elements of Information Theory. John Wiley & Sons, Inc., 1991.

[4] G.W. Rowe and P. Gregor, “A Computer Based Learning System for Teaching Computing: Implementation and Evaluation”, Computers and Education, 33 (1999). pp. 65-76.

[5] T.K. Shih, et al., "An Adaptive Tutoring Machine Based on Web Learning Assessment", Proc. ICME 2000, IEEE International Conference on Multimedia and Expo, 2000. pp. 1667-1670.

[6] Moore, M.G. and Kearsley, G., Distance Education: A Systems View, 1996, Belmont: Wadsworth

[7] Hillman Daniel C. A., Deborah J. Willis, and Charlotte N. Gunawardena, “Learner-Interface Interaction in Distance Education: An Extension of Contemporary Models and Strategies for Practitioners”,The American Journal of Distance Education, Vol.8, No.2, 1994.

[8] Dragos D. Margineantu and Thomas G. Dietterich. Pruning adaptive boosting┌ In Machine Learning: Proceedings of the Fourteenth International Conference,pages 211-220, 1997

[9] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using condence-rated predictions. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, pages 80-100,1998. To appear, Machine Learning.

[10] HONG Yixin. On the Laws of Information-Knowledge-Intelligence Transforms [J] . Journal of Beijing University of Posts and Telecommunications, 2007,30 (01):1-8.

Profile

Li lan the master of computer applied technology and the lecturer of software college of Nanchang university, whose research interesting is computer network and information security.

Framework for Efficient Letter Selection in Genetic Algorithm Based Data Mining

Xiaoyan Chen[1] Shijue Zheng[1] Tao Tao[2]

1 Department of Computer Science , Hua Zhong Normal University, Wuhan, Hubei, 430079, China

Email:xiaoyan0205@126.com

2 Sanya Garrison, Sanya, Hainan, 572011, China

Email: taoxinzhou @126.com

Abstract

We present the design of more effective and efficient genetic algorithm based data mining techniques that use the concepts of letter selection. Data mining involves nontrivial process of extracting knowledge or patterns from large databases. Genetic Algorithms are efficient and robust searching and optimization methods that are used in data mining. In this paper we propose a letter selection in genetic algorithm based data mining, Explicit letter selection is traditionally done as a wrapper approach where every candidate feature subset is evaluated by executing the data mining algorithm on that subset. In this article we present a GA for doing both the tasks of mining and letter selection simultaneously by evolving a binary code along side the chromosome structure used for evolving the rules. Results from applying the above techniques to a real world data mining problem show that combining both the letter selection methods provides the best performance in terms of prediction accuracy and computational efficiency.

Keywords: letter selection; GA; data mining; framework; self-adaptive

1 Introduction

Data mining is a process of extracting nontrivial, valid, novel and useful information from large databases.Hence Data mining can be viewed as a kind of search for meaningful patterns or rules from a large search space, that is the database. In this light, Genetic Algorithms are a powerful tool in data mining, as they are robust search techniques. Genetic Algorithms (GA) are a set of random, yet directed search techniques.They process a set of solutions simultaneously and hence are parallel in nature. They are inspired by the natural phenomenon of evolution. They are superior to gradient descent techniques as they are not biased towards local optima[1].

Genetic Algorithms have found a wide gamut of application in data mining, where knowledge is mined from large databases. Genetic algorithms can be used to build effective classifier systems, mining association rules and other such datamining problems. Their robust search technique has given them a central place in the field of data mining and machine learning.

This work is partially supported by “The research of foreign Chinese distance visualized teaching model” (The National Society Science Fund of P.R.China) No: 07BYY033

GA can be viewed as an evolutionary process where at each generation, from a set of feasible solutions, individuals or solutions are selected such that individuals with higher fitness have greater probability of getting chosen.At each generation,these chosen individuals undergo crossover and mutation to produce a population of the next generation. This concept of survival of the fittest proposed by Darwin is the main cause for the robust performance of GAs. Crossover helps in the exchange of discovered knowledge in the form of genes between individuals and mutation helps in restoring lost or unexplored regions in search space.

In this paper we propose frameworks for data mining using genetic algorithms, implement these, and evaluate their performance using examples. We chose genetic algorithm due to its simplicity and its capability as a powerful search mechanism. We present the design of more effective and efficient genetic algorithm based data mining techniques that use the concepts of self-adaptive letter selection together with a wrapper letter selection method based on Hausdorff distance measure. [2]A genetic algorithm uses a population of individual solution structures called chromosomes. The fitness of an individual solution is its performance measure.This measure is used to favor selection of successful parents for new offspring, such that the whole population of solutions incrementally evolves towards greater fitness. Offspring solutions are produced from parent solutions by the application of crossover and mutation operators. Theory shows that the knowledge about desirable solutions is advantageously stored in the population itself, implicitly contained in the surviving chromosomes. We take advantage of this principle in developing modified frameworks essentially using the genetic algorithm at its core.

2 Letter selection

There have been many approaches to letter selection based on a variety of techniques, such as statistical, geometrical, information-theoretic measures, mathematical programming, neurofuzzy,Receiver Operating Curves, discretization, among others. Each of these has its advantages and disadvantages that are context-specific, and there is no one universally best letter selection method. As gathered,data used in any data mining application tends to not be in a form that is suitable for data mining. A first step is to clean or pre-process this data to a form that is appropriate for data mining that follows.

Letter selection has been traditionally used in data mining applications as part of the data cleaning and preprocessing step where the actual extraction and learning of knowledge or patterns is done after a suitable set of letters is extracted. If the letter selection is independent of the learning algorithm it is said to use a filter approach. Ref. presents one such filter approach to letter selection using genetic algorithms. If the letter selection method works in conjunction with the learning algorithm it is using a wrapper approach.The problem with thefilter approach is that the optimalset of letters maynot be independent of the learning algorithm or classifier. The wrapper approach provides a better approach, however it is computationally expensive as each candidate letter subset has to be evaluated by executing a learning algorithm on that subset. When used with GAs, the wrapper approaches become even more prohibitively expensive. In this article we present a wrapper approach to letter selection using a GA as the learning algorithm and show how the computational complexity of the approach can be reduced by incorporating a second self-adaptive letter selection approach where the learning and letter selection are done simultaneously.[3]

Letter selection concept presented in this article is related to two aspects of work done earlier: self-adaptation and use of non-coding material in the chromosome structure motivated by the existence of non-encoding DNA in biological systems. In biological systems an intron is a portion of the DNA that is not transcribed into proteins. Introns can become an important part of the evolution process by providing a buffer against the destructive effects of the genetic algorithm. At the same time introns have been shown to be useful as a source of symbols that can be effectively used to evolve new behaviors through subsequent evolution. Self-adaptation refers to the technique of allowing [4]characteristics of the search to evolve during the search rather than be specified by the user. Most of the work done on self-adaptation has focused on choices related to search operators. Aspects of these choices are encoded along with each member of the population and they are allowed to vary and adapt on an individual basis.

3 Genetic algorithm based data mining

Genetic algorithms are adaptive methods, which may be used to solve search and optimization problems (Beasley and Bull, 1993). They are based on the genetic process of biological organisms. Over many generations, natural populations evolve according to the principles of natural selection, i.e. *survival of the fittest*, first clearly stated by Charles Darwin (1859) in *The Origin of Species by Natural Selection*. By mimicking this process, genetic algorithms, if suitably encoded, are able to *evolve* solutions to real world problems. Before a genetic algorithm can be run, an *encoding* (or *representation*) for the problem must be devised. [5]A *fitness function*, which assigns a figure of merit to each encoded solution, is also required. During the run, parents are *selected* for reproduction and *recombined* to generate offspring.The genetic algorithm described here is an adaptive GA where between evaluation of individuals and applying the three operators of GA (selection, crossover and mutation), the parameters used for these operators are updated based on the evaluation of individuals.[6]

3.1 Chromosome representation

The genetic algorithm described in this paper uses a random key alphabet which is comprised of random numbers between 0 and 1. The evolutionary strategy used is similar to the one proposed by Bean (1994), the main difference occurring in the crossover operator. The important feature of random keys is that all offspring formed by crossover are feasible solutions. This is accomplished by moving much of the feasibility issue into the objective function evaluation. If any random key vector can be interpreted as a feasible solution, then any crossover vector is also feasible. Through the dynamics of the genetic algorithm, the system learns the relationship between random key vectors and solutions with good objective function values.[7]

A chromosome represents a solution to the problem and is encoded as a vector of random keys. In a direct representation, a chromosome represents a solution of the original problem, and is usually called *genotype*, while in an indirect representation it does not and special procedures are needed to derive a solution from it usually called *phenotype*.

$$\text{Chromosome} = \left(\underbrace{\text{gene}_1, \cdots, \text{gene}_n}_{\text{Pr iorities}}, \underbrace{\text{gene}_{n+1}, \cdots, \text{gene}_{2n}}_{\text{Delaytimes}}, \underbrace{\text{gene}_{2n+1}, \cdots \text{gene}_{2n+m}}_{\text{Re leasedates}} \right)$$

The first *n* genes are used to determine the priorities of each activity. The genes between *n* + 1 and 2*n* are used to determine the delay time used at each of the *n* iterations of scheduling procedure which schedules one activity per iteration. The last *m* genes are used to determine the release dates of each of the *m* projects.

3.2 Problem definition

In this section we present the design of a genetic algorithm for rule learning in a data mining application. Assume that the data mining problem has k attributes and we have a set of training examples

$$\theta = \{(\mathrm{L_i}, \mathrm{c}) \mid \mathrm{i} = 1, \cdots, \mathrm{L}\} \qquad (1)$$

where L is the total number of examples, each example $\mathrm{L_i}$ is a vector of k attribute values

$$\mathrm{L_i} = [\mathrm{l_{i1}}, \mathrm{l_{i2}}, \cdots, \mathrm{l_{ik}}] \qquad (2)$$

and c is its classification value. The goal of data mining is to learn concepts that can explain or cover all of the positive examples without covering the negative examples.

The representation of a concept or a classifier used by the GA is that of a disjunctive normal form. A concept is represented as

$$\varphi = \tau_1 \vee \tau_2 \vee \cdots \vee \tau_{\mathrm{q}} \qquad (3)$$

where each disjunct τ_{i} (also referred to as a rule) is a conjunction of conditions on the k attributes,

$$\tau_{\mathrm{i}} = (\delta_{1,\mathrm{i}} \wedge \delta_{2,\mathrm{i}} \wedge \cdots \wedge \delta_{\mathrm{k,i}}) \qquad (4)$$

The above concept φ is said to have a size of q. In order to handle continuous attributes each condition $\delta_{\mathrm{j,i}}$ is in the form of a closed interval [$\mathrm{a_j}$,$\mathrm{b_j}$]. We say that a disjunct τ_i covers an example L_i if

$$(a_j \le l_{ij} \le b_j)\forall j = 1\cdots k \qquad （5）$$

Each member of the population in the GA is a single disjunct and the GA tries to find the best possible disjunct. At each generation it retains the best disjunct and replaces the rest through the application of the genetic operators. After the genetic algorithm converges, the best disjunct found is retained and the positive examples it covers are removed. The process is repeated until all the positive instances are covered.

3.3 Self-adaptive value P_c and P_m

Crossover probability P_c , control the used frequency of crossover operation,higher crossover probability can achieve a greater space of solution,thereby reducing the chance for non-optimal solution. Variation probability P_m control a new gene into groups of proportion, if the probability is too low, a lot of useful genes can not be chosen; if the probability is too high, which induces random changes, future generations may loose good characters that from their parents.In this paper,we adopt self-adaption to adjust P_c and P_m ,which can avoid excellent genes be damaged .[8]

When the max fitness value F_{max} and the average fitness value F_{av} are closed, the colony close to the convergence, should increase P_c and P_m . Contrarily,the diversity of the colony is strong, should reduce P_c and P_m .

$$P_c = \begin{Bmatrix} P_c * \dfrac{F_{\max} - X_1}{F_{\max} - F_{av}} \text{当} X_1 \ge F_{av} \\ P_c \qquad\qquad \text{当} X_1 < F_{av} \end{Bmatrix}$$

$$P_m = \begin{Bmatrix} P_m * \dfrac{F_{\max} - Y}{F_{\max} - F_{av}} \text{当} Y \ge F_{av} \\ P_m \qquad\qquad \text{当} Y < F_{av} \end{Bmatrix}$$

Where X_1 is the bigger self-adaptive value in crossover operation ,Y is the self-adaptive value in aberrance.

3.4 The condition of convergence

When the value (F_{max} －F_{av}) is more smaller, the possibility of local optimal solution is bigger, the possibility of prematurity is bigger too. Augment the value of P_c and P_m to enhance the ability of colony producing new individuals.Contrary, (F_{max} －F_{av}) is more bigger, the colony is emanative, in order to enhance the individual ability of convergence,shoud reduce the value of P_c and P_m.Thereby, we should let P_c and P_m with (F_{max} － F_{av}) is inversely proportional.Secondly,in order to protect excellent mode,the different ivdividual in the same generation, we should let P_c and P_m with (F_{max} － F_{av}) is proportional.[9]

4 Simulation experiment

We carried out several experiments to first test the effectiveness of using the binary selection vector for performing letter selection. Since we are using a population-based method (GA), one of the best ways to study the effectiveness of such a technique is to study the evolution of proportions having different letters. Figure.1 shows the letter selection speed without GA, and Figure2. shows the letter selection with GA, comparing the two figures, we can find that GA can improve the speed of search, which is very important for our project.

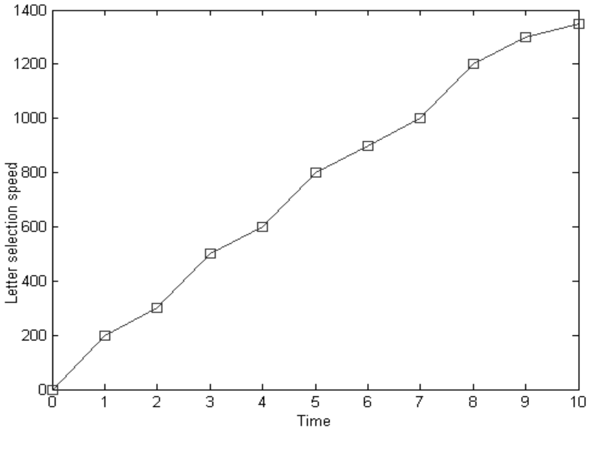


Figure 1 The letter selection speed without GA

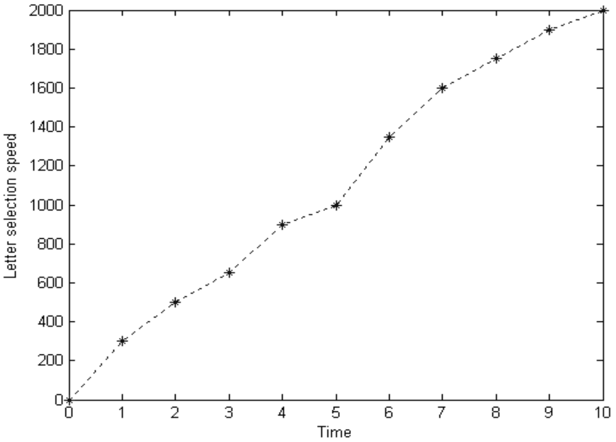


Figure 2 The letter selection speed with GA

5 Conclusions

Datamining algorithms perform well when concepts to be learned are mapped to space that is not complex.However, most real world data mining applications do not fall into this category, Evolutionary algorithms are excellent for performing search, especially when the search landscape is complex.[10]However, one of the reasons why the use of evolutionary algorithms (especially GAs) has not received widespread attention in the data mining community is that the GAs tend to be computationally expensive, and become prohibitively so, once the problem size increases. These techniques include improvement in the learning algorithm itself, input data preprocessing to reduce the complexity of learning concepts of interest, and incorporation of data preprocessing in the learning algorithm, among others.

References

[1] Ashish Ghosh, Bhabesh Nath, "Multi-objective rule mining using genetic algorithms", Information Sciences 163 (1–3) (2000) 123–133

[2] F. Herrera, M. Lozano, "Gradual distributed real-coded genetic algorithms", IEEE Transactions on Evolutionary Computation 4 （1）(2000) 43–62

[3] J.M. Benitez, J.L. Castro, C.J. Mantas, F. Rojas, "A neuro-fuzzy approach for feature selection", in: Proceedings of IFSA World Congress and 20th NAFIPS International Conference, vol. 2, 2001, pp. 1003–1008

[4] K. Deb, H.G. Beyer, "Self adaptive genetic algorithms with simulated binary crossover", Evolutionary Computation 9 （2）(2001) 197–221

[5] J.H. Lobo, "The parameter-less genetic algorithm: rational and automated parameter selection for simple genetic algorithm operation",Ph.D. thesis, University of Lisbon, Portugal, 2000

[6] Oscar Montiel, Oscar Castillo, Roberto Seplveda, Patricia Melin, "Application of a breeder genetic algorithmnext term for finite impulse filter optimization", Information Sciences 161 (3–4) (2004) 139–158

[7] W. Heinzelman, A. Chandrakasan, H. Balakrishnan, "Energy -efficient communication protocol for wireless micro-sensor networks", in: Proc. International Conference on System Sciences, 2000

[8] R. Madan, S. Lall, "Distributed algorithms for maximum lifetime routing in wireless sensor networks", in: Global Telecommunications Conference (GLOBECOM'04), IEEE, vol. 2, 2004

[9] L. Qi, D. Sun, "Smoothing functions and a smoothing Newton method for complementarity and variational nequality problems", Journal of Optimization Theory and Applications 113 (2002) 121–147

[10] P.von Rickenbach, R. Wattenhofer, "Gathering correlated datain sensornetworks",in:DIALM-POMC'04:Proceedings of the 2004 Joint Workshop on Foundations of Mobile Computing, ACM Press, New York, NY, USA, 2004, pp.60–66

K-harmonic Means Data Clustering with Particle Swarm Optimization

Kezhong Lu[1] Wenbo Xu[2] Guangqian Xie[3]

1 Department of Computer Science, Chizhou College, 247100, China

Email: luke76@163.com

2 School of Information Technology, Southern Yangtze University, Wuxi, 214122, China

Email: xwb@sytu.edu.cn

3 School of Computer Information and Engineering, Changzhou Institute of Technology, 213002, China

Email: xgqmail2000@163.com

Abstract

Unlike K-means, the K-Harmonic means (KHM) is less sensitive to initial conditions. However, KHM as a center-based clustering algorithm can only generate a local optimal solution. In this paper, we develop a new hybrid clustering algorithm combining Particle Swarm Optimization and K-Harmonic Means (HPSO) for solving this problem. This algorithm has been implemented and tested on several real datasets. The performance of this algorithm is compared with KHM and PSO. Our computational simulations reveal the HPSO clustering algorithm combines the ability of global searching of the PSO algorithm and the fast convergence and less sensitive to initial conditions of the KHM algorithm. The HPSO is a robust clustering algorithm.

Keywords: Clustering; K-Harmonic Means; Particle Swarm Optimization; Hybrid Clustering Algorithm

1 Introduction

Cluster analysis is the grouping of similar feature vectors into clusters that in some sense belong together because of similar characteristics. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering algorithms are used in many applications, such as data mining[1], data compression[2], image segmenta- tion[3], machine learning[4], etc.

The existing clustering algorithms can be simply classified into the following two categories: hierarchical clustering and partitional clustering. These algorithms try to minimize certain criteria (e.g., a square error function) and can therefore be treated as optimization problems. Partitional clustering techniques are more popular than hierarchical techniques in pattern recognition [5], hence, this paper will concentrate on partitional techniques.

The most popular class of partitional clustering methods is the center-based clustering algorithms. K-means (KM), Fuzzy K-Means (FKM) and K-Harmonic means (KHM) are three center-based algorithms. K-means has been used as a popular clustering method due to its simplicity and high speed in clustering large data sets. However, K-means has two shortcomings: dependency on the initial state and convergence to local optima. FKM is proposed to smooth the hard nature of KM algorithm. The FKM employs fuzzy partitioning such that a data point can belong to all clusters with different membership grades between 0 and 1. Like KM, FKM does not ensure that it converges to a solution which is global optimum. The clustering result of FKM is dependent on initial membership degrees, because of cluster centers are initialized using membership grades which are randomly initialized. KHM is a more recent algorithm proposed by Zhang at 1999[6] and 2000[7] and modified by Hammerly and Elkan at 2002[8]. This algorithm minimizes the harmonic average from all points in N to all centers in K. Moreover, KHM has a soft membership function and a varying weight function. Contrary to KM and FKM, KHM is less sensitive to initial conditions. Experiments conducted by Zhang et al. showed that KHM outperformed KM and FKM[6,7,8]. However, like both KM and FKM, KHM is also a center-based clustering method. One of the most important problems in center-based clustering is convergence to local optimum. Recently, evolutionary and metaheuristics like, tabu search[9], genetic algorithms [10] ,simulated annealing[11] and Particle Swarm Optimization[3,12] have been successfully employed to overcome this problem. In this study, we recast the recently proposed Particle Swarm Optimization algorithm to suit the need for data clustering with KHM.

The remainder of this paper is organized as follows: Section 2 provides a general overview of the KHM and PSO optimal algorithm. The PSO clustering algorithms are described in Section 3. Section 4 provides the detailed experimental setup and results for comparing the performance of the hybrid PSO algorithm with the KHM and PSO algorithms. The discussion of the experiment's results is also presented. The conclusion is in Section 5.

2 Background

2.1 K-Harmonic Means Clustering

K-Harmonic means clustering is a center-based clustering method proposed by Zhang at 1999 and 2000 and modified by Hammerly and Elkan at 2002[8].

For the purpose of this paper, define the following symbols:

- ***K*** : number of clusters
- ***Z*** : data set matrix
- ***N*** : number of data samples (patterns)
- ***C*** : cluster centers Matrix
- ***U*** : membership matrix
- ***W***: Weight matrix

The harmonic average is defined as

$$HA(\{a_1 \ldots\ldots\ldots a_k\}) = \frac{K}{\sum_{k=1}^{K} \frac{1}{a_k}} \tag{1}$$

This function has the property that if any one element in $a_1 \ldots a_k$ is small, the Harmonic Average will also be small. If there are no small values the harmonic average will be large. It behaves like a minimum function but also gives some weight to all the other values. When we assigning the patterns to the clusters we use minimum of the harmonic average in K-Harmonic means algorithm:

$$HA\left\{\|z-c\|^2 \middle| c \in C\right\} = \frac{|C|}{\sum_{c \in C} \frac{1}{\|z-c\|^2}} \tag{2}$$

And the objective function for the K-Harmonic means algorithm is then;

$$J_{KHM}(X,C) = \sum_{i=1}^{N} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|z_i - c_j\|^p}} \tag{3}$$

Where p is a user-specified parameter. Zhang [7] introduces a class of KHM with parameter p that is power associated with the distance calculation. It was found that KHM works better with values of $p \geq 2$.

The membership and weight functions for KHM are:

$$u(c_j / x_i) = \frac{\|z_i - c_j\|^{-p-2}}{\sum_{j=1}^{K} \|z_i - c_j\|^{-p-2}} \tag{4}$$

$$w(x_i) = \frac{\sum_{j=1}^{K} \|z_i - c_j\|^{-p-2}}{(\sum_{j=1}^{K} \|z_i - c_j\|^{-p})^2} \tag{5}$$

Hence, KHM has a soft membership function and a varying weight function. KHM assigns higher weights for patterns that are far from all the centroids to help the centroids in covering the data [8].

The KHM algorithm is summarized as:

1. Randomly initialize the *K* cluster centroids

2. Repeat

(a) For each pattern, z_p, in the data set do

Compute its membership $u(c_j / x_i)$ to each centroid c_j and its weight $w(z_\mathrm{p})$

endloop

(b) Recalculate the *K* cluster centroids, using

$$c_j = \frac{\sum_{\forall z_p} u(c_j \mid z_i) w(z_i) z_i}{\sum_{\forall z_p} u(c_j \mid z_i) w(z_i)} \qquad （6）$$

until a stopping criterion is satisfied.

The KHM clustering process can be stopped when any one of the following criteria are satisfied: when the maximum number of iterations has been exceeded, when there is little change in the centroid vectors over a number of iterations, or when there are no cluster membership changes. For the purposes of this study, the algorithm is stopped when a user-specified number of iterations has been exceeded.

2.2 Particle Swarm Optimization

Particle swarm optimization (PSO) is one of the evolutionary computational techniques. Since its introduction[13,14], PSO has attracted much attention from researchers around the world. It is a population-based search algorithm and is initialized with a population of random solutions, called particles. Each particle in PSO moves over the search space at velocity dynamically adjusted according to the historical behaviors of the particle and its companions.

Suppose that the search space is *D*-dimensional, and the position of *i*th particle of the swarm can be represented by a *D*-dimensional vector, $x_i=(x_{i1},\ldots, x_{id},\ldots,x_{iD})$. The velocity (position change per generation) of the particle x_i can be represented by another *D*-dimensional vector $v_i=(v_{i1},\ldots, v_{id},\ldots,v_{iD})$. The best position previously visited by the ith particle is denoted as $p_i=(p_{i1},\ldots, p_{id},\ldots,p_{iD})$. If the topology is defined such that all particles are assumed to be neighbors and *g* as the index of the particle visited the best position in the swarm, then p_g becomes the best solution found so far, and the velocity of the particle and its new position will be determined according to the following two equations:

$$\text{vid=}\omega\text{vid+ c1*rand1()*(pid - xid)+ c2*}$$
$$\text{rand2()*(pgd - xid)} \qquad （7）$$

$$\text{xid = xid + vid} \qquad （8）$$

Where c_1 and c_2 are positive constants, and *rand*1() and *rand*2() are the uniform random value in the range [0,1]. The weight factor *ω* provides a balance between global and local explorations. The constants c_1 and c_2 represent the weighting of the stochastic acceleration terms that pull each particle toward the p_i and p_g position [13].

The PSO is usually executed with repeated application of Eq.（7）and （8）until a specified number of iterations has been exceeded. Alternatively, the algorithm can be terminated when the velocity updates are close to zero over a number of iterations.

3 Description of the PSO clusteringa-lgorithm

3.1 The Basic PSO Clustering Algorithm

In the PSO clustering algorithm, the multi-dimensional vector space of a dataset is modeled as a problem space. Each term in the dataset represents one dimension of the problem space. Each vector of a dataset can be represented as a dot in the problem space. The whole dataset can be represented as a multiple dimension space with a large number of dots in the space.

One particle in the swarm represents one possible solution for clustering a given dataset. Therefore, a swarm represents a number of candidate clustering solutions for the problem space. Each particle maintains a matrix $x_i = (C_1, C_2, \ldots, C_\mathrm{i}, .., C_\mathrm{k})$, where C_i represents the *i*th cluster centroid vector and *K* is the number of clusters. According to its own experience and those of its neighbors, the particle adjusts the centroid vector' position in the vector space at each generation. The average harmonic distance of datasets to the cluster centroid is used as the fitness value to evaluate the solution represented by each particle. The fitness value is measured by the Eq.（3）.

The PSO algorithm can be summarized as:

```
(1) At the initial stage, each particle randomly chooses $K$ different data vectors from the given dataset as the initial cluster centroid vectors.
(2) For $t$=1 to $t_{\max}$ do
    (a) For each particle $i$ do
    (b) For each data vector $z_p$ do
        calculate the fitness using Eq.（3）
    (c) Update the global best and local best positions
    (d) Update the cluster centroids using Eq.（7）and Eq.（8）
```

where $t_{\max}$ is the maximum number of iterations.

3.2 The Hybrid PSO clustering

In this section, we propose a hybrid clustering algorithm combining PSO and KHM algorithm. The KHM algorithm can enhance the hybrid clustering algorithm's local search ability, and the PSO algorithm can enhance the hybrid clustering algorithm's global search ability. So the performance of the hybrid clustering algorithm can further be improved.

The Hybrid PSO clustering algorithm is:

```
(1) At the initial stage, each particle randomly chooses $K$ different data vectors from the given dataset as the initial cluster centroid vectors.
(2) For $t$=1 to $t_{\max}$ do
    (a) For each particle $i$ do
    (b) For each data vector $z_p$ do
        calculate the membership function using Eq.（4）
        calculate the weight function using Eq.（5）
        calculate the KHM cluster centroids using Eq.（6）
        iv. calculate the fitness using Eq.（3）
    (c) Update the local best positions:
        If the new $p_i$ is better than the old $p_i$, update local best positions and replace the current particle $i$ with KHM cluster centroids
    (d) Update the global best positions
    (e) Update the cluster centroids using Eq.（7）and Eq.（8）
```

where $t_{\max}$ is the maximum number of iterations.

4 Experiments and results

4.1 Datasets

We used three different document collections to compare the performance of the KHM, PSO and hybrid PSO clustering algorithms (HPSO). The main purpose is to compare the quality of the respective clustering. The three datasets are described below.

- Iris. The dataset consists of N = 150 samples of three iris flowers (K = 3). Each object is defined by four attributes, n = 4.
- Wine. This dataset contains chemical analysis of N = 178 wines, derived from three different cultivars, K = 3. Wine type is based on 13 continuous attributes, n = 13.
- Breast-cancer. The Wisconsin breast cancer database contains 9 relevant inputs and 2 classes. The objective is to classify each data vector into benign or malignant tumors.

4.2 Experimental setup

The the KHM, PSO and HPSO are applied on the three datasets, respectively. For all algorithms, $t_{\max}$=100 and the parameter p=2. In the PSO and HPSO clustering algorithm, we choose 10 particles, the inertia weight w is initially set as 0.72 and is reduced by 1% at each generation to ensure good convergence, the acceleration coefficient constants c_1 =c_2 =1.49.

4.3 Results and Discussions

The fitness equation 3 is used not only in the all algorithms for fitness value calculation, but also in the evaluation of the cluster quality. The smaller the fitness value, the more compact the clustering solution is. Table 1 demonstrates the experimental results by using the KHM, PSO and HPSO respectively. Thirty simulations are performed for each algorithm. The average fitness values (F_{avg}), the best fitness values (F_{best}) and the worst fitness values (F_{worst}) are recorded in Table 1.

As shown in Table 1, the KHM algorithm is less sensitive to initial conditions. Because the KHM algorithm is easy to converge to local optimum, the clustering results are not satisfied. The PSO are stochastic algorithms. It is greatly dependent on the generation of initial solutions. So the fitness values of the algorithm change obviously. The hybrid algorithms keep up the merit of KHM algorithm, and are also less sensitive to initial conditions. Because 100 iterations is not enough for the PSO algorithm to converge to an optimal solution, the result values in the table 1 indicate that the PSO algorithm has less improvements compared to the results of the KHM algorithm. However, the HPSO algorithm improves greatly. The HPSO algorithm generates the clustering result with the lowest fitness value for all three datasets.

Figure 1 illustrates the convergence behaviors of these algorithms on the Iris problem. In Figure 1, the KHM algorithm converges quickly but prematurely. As shown in Figure 1, the fitness value of the KHM algorithm is sharply reduced from 576 to 183.52 within 5 iterations and fixed at 183.52. The HPSO algorithm has lower convergence, but to best fitness function value. As indicated in Figure 1, the hybrid algorithms converge after about 10 function evaluations. The PSO algorithm has lowest convergence. It converges after about 40 function evaluations.

Table 1 Comparison of KHM, PSO and HPSO algorithms

	Method	F_{best}	F_{avgt}	F_{worst}
Iris	KHM	183.5158	183.5158	183.5158
	PSO	182.0762	203.3161	307.6339
	HPSO	181.7359	181.9476	182.6335
Wine	KHM	5.4262e6	5.4262e6	5.4262e6
	PSO	5.3894e6	5.3954e6	5.4094e6
	HPSO	5.3883e6	5.3886e6	5.3980e6
Breast-cancer	KHM	3.0187e4	3.0187e4	3.0187e4
	PSO	3.2348e4	3.6444e4	4.0996e4
	HPSO	3.0041e4	3.0076e4	3.0138e4

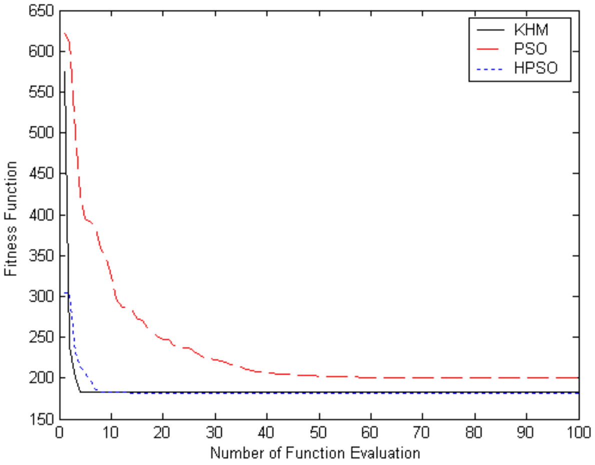


Figure 1 Algorithm convergence for Iris problem

5 Conclusion

In this study, we propose a hybrid clustering algorithm combining KHM and PSO algorithm. In the hybrid clustering algorithms, the clustering behavior can be classified into two stages: the global searching stage by PSO algorithm and the local refining stage by KHM algorithm. So The hybrid PSO clustering algorithm combines the ability of global searching of the PSO algorithm and the fast convergence and less sensitive to initial conditions of the KHM algorithm and avoids the drawback of both algorithms. The hybrid PSO clustering algorithm is compared against KHM and PSO algorithm, which showed that the hybrid PSO clustering algorithm generally have better convergence to lower fitness value.

Future studies will do more elaborate tests on higher dimensional problems and large number of patterns. The hybrid PSO clustering algorithm will also be extended to dynamically determine the optimal number of clusters.

References

[1] IE Evangelou, DG Hadjimitsis, AA Lazakidou, C Clayton, “Data Mining and Knowledge Discovery in Complex Image Data using Artificial Neural Networks”, Workshop on Complex Reasoning an Geographical Data, Cyprus, 2001, pp.134-143

[2] HM Abbas, MM Fahmy , “Neural networks for maximum likelihood clustering”, Signal Process, 36(1), 1994, pp.111–126

[3] M Omran, AP Engelbrecht, A Salman, “Particle Swarm Optimization Method for Image Clustering”, International Journal on Pattern Recognition and Artificial Intelligence, 19（3）, 2005,pp.297–322

[4] C Carpineto, G Romano, “A lattice conceptual clustering system and its application to browsing retrieval”, Machine Learning, 1996,24（2）:95–122

[5] AK Jain, R Duin, J Mao, “Statistical pattern recognition:a review”, IEEE Trans Pattern Anal Mach Intell 22（1）, 2000,pp.4–37

[6] B Zhang, M Hsu, U Dayal, “K-harmonic means - a data clustering algorithm”, Technical Report HPL-1999-124, Hewlett-Packard Laboratories, 1999

[7] B Zhang, M Hsu, U Dayal, “K-Harmonic Means”, International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, TSDM2000, Lyon, France, 2000,pp.327-333

[8] G Hammerly, C Elkan, “Alternatives to the k-means algorithm that find better clusterings”, in Proc. of the 11th International Conference on Information and Knowledge Management,2002,pp.600-607

[9] PS Shelokar, VK Jayaraman, BD Kulkarni, “An ant colony approach for clustering”, Analytica Chimica Acta , 509, 2004, pp.187–195

[10] P Scheunders, “A Genetic C-means Clustering Algorithm Applied to Image Quantization”, Parttern Recognition, 30（6）,1997, pp.651-659

[11] DA Bell, “Application of Simulated Annealing to Clustering Tuples in Databases”, Journal of the American Society for Information Science, 41（2）, 1990, pp.98-110

[12] VD Merwe, AP, Engelbrecht, “Data clustering using particle swarm optimization”, Proceedings of IEEE Congress on Evolutionary Computation 2003 (CEC 2003), Canbella, Australia, 2003,pp.215-220

[13] J Kennedy, RC Eberhart, “Particle Swarm Optimization”, Proceedings of IEEE International Conference on Neural Networks, Piscataway, NJ, 1995, pp.1942-1948

[14] RC Eberhart,Y Shi, “Particle Swarm Optimization: Developments, Applications and Resources”, Proceedings of the IEEE Congress on Evolutionary Computation, Seoul, Korea, 2001,pp. 81-86

An Efficient Structure Similarity Measure Method for XML documents based on Vector Space Model

Hongcan Yan*[1,2] Minqiang Li[1] Dianchuan Jin[2] Dazhuo Zhou[3] Shaohong Yan[2]

1 School of Management, Tianjin University, Tianjin, 300072, China

2 College of Science, Hebei Polytechnic University, Tangshan, Hebei, 063000, China

3 Computer Center, Hebei University of Economics and Trade, Shijiazhuang, Hebei, 050061, China

Email: yanhongcan@heut.edu.cn

Abstract

A novel way of similarity measure for XML documents structure based on frequency structured vector model is proposed against the detects of the methods in existence. In this model, all frequent subtrees of documents are viewed as structured characteristic space; the expression of document structured vector and weight function are derived and the angle cosine between Eqtwo vectors is applied to measure similarity of the two documents. At the same time, the algorithm TreeMiner is reformed to improve the efficiency of mining frequency subtrees in a forest from data structure and mining process, which entitled TreeMiner+. The experimental results show that this method acquires very high precision and accuracy, the time cost of algorithm TreeMiner+ is reduced three times when minimum support is 70% or higher.

Keywords: Document Structure Similarity; Embedded Subtree; Frequent Structured Vector Model; Frequent Subtree; Class Extension; Scope-List Join

1 Introduction

In recent years, XML has become a popular way of storing and exchanging[1-3] many data sets because the semi-structured nature of XML allows to modify, extend and transplant structure data[4-7]. Increasing domains adopt XML as standard for expressing, such as MathML, NewsML, OWL, LOGML, ebXML, cnXML[8-9]. XML data thus forms an important data mining domain, and it is valuable to develop techniques that can be used to extract patterns from such data. For example, mining XML document structure provides useful information for biology informatics, network log analysis, Web structure analysis and classification. XML document structure similarity measure is basic core question of XML structure analysis. The existing document structure similarity measure methods have distance editing path matching method[10-13], time analysis method[14]. These methods all view XML document as a tag tree. Distance editing method uses the smallest cost of operation (rename, delete, insert) for translating a document into an another document to measure two document similarity by graph matching arithmetic, but if the document is large, the cost of calculation is high. The path matching method use Jaccard coefficient to measure similarity of two document by side matching, but cannot get high precision. Time analysis method views a document as a time sequence, view a side as an impulse, use Fourier transform to calculate similarity, but this method cannot resolve the influence made by label repetition.

The references[15,16] provided an efficient mining frequent subtree iterative arithmetic TreeMiner. It can find all frequent subtrees satisfying the smallest support threshold. The references[17] gave an efficient mining

* This work was supported by a grant from the Ph.D. Programs Foundation of Ministry of Education of China (No. 20020056047)

unordered frequent subtree arithmetic. According to these, this paper combines Vector Space Model of text classification, puts forward the frequent structure vector model of XML document, and defines the similarity of document by the frequency of frequent structure and weight. At the same time, we improve on the arithmetic TreeMiner from two ways: ①in order to judge rapidly the relation of ancestor-descendant for scope-list, add the hierarchy code information to data structure of node. If and only if the nodes of £(y) are descendant nodes of £(x), then in-scope and out-scope test can be carried out. Thus the arithmetic get rid of a lot of redundant operations and improve the capability. ② according to the idea of “calculating support first, then do class extension”, adjust the order of nodes joins within class extension and scope-list joins within frequent subtrees support; avoid the unwanted linking for inexistent or not frequent subtrees, thereby improve the mining efficiency.

The experiment makes it clear, the arithmetic TreeMiner$^+$ is excel TreeMiner in time complexity. This advantage is especially distinct with the increase of support, and this method is fit to large documents. Frequent structure vector model has very high precision in measuring XML document structure similarity with lower calculating cost.

2 Frequent structured vector model

The references [18,19] used Vector Space Model to express XML document. Especially, the reference[9] provided the structured link vector model. It views each structure unit of document as a vector which is similar to a document of VSM model, but it did not give the method of getting structure unit, and only view each node as a basic structure unit, which could not express maturity structure information of XML document. This paper will provide the expression of frequent structure vector model (FSVM) and similarity measure method of structure, and give an efficiency algorithm TreeMiner+ of mining frequent subtree-structure unit of structured vector space.

2.1 The concept related to frequent subtree

We denote a XML document as a ordered tree with label T=(r,V,B,L), where r is root node, V is the set of nodes, and

$$\pi\,(D,S)=\sum\nolimits_{T\in D} dT(S)/|D|\ ,\qquad （1）$$

$$\text{where:}dT(S)=\begin{cases}1 & \text{When}'T(S)>0\\ 0 & \text{when}'T(S)=0\end{cases}$$

B the set of branches, L the set of all labels. Then XML documents set D (or database) is viewed as set of ordered trees, D={ t0,t1,t2,……}, entitle document forest, where each tree has an exclusive identifier tid, tid∈{0,1,2,…… }, express the tree t0,t1,t2 and so on separately.

Definition 1 (**Ordered Tree**) XML document ordered tree T=(r,V,B,L), where V＝{$n_0,n_1,n_2,\ldots n_{|T|-1}$}, nid∈{0,1,2,…… }. Nodes are coded according to depth first, |T| express the size of tree, or count of nodes. The nodes are described as a circle, and needn’t be distinguished from element or attribute. Each node comes from the numbering scheme L={1,2,3,……,m}, where different node can have same label. Each branch b=<x,y> ∈B is an ordered couple, where x is parent node of y. The figure 1 is an ordered tree, where r=n0, V={n0,n1,n2,n3,n4,n5,n6}, B={<n0,n1>,<n1,n2>, <n2,n3>,<n2,n4>,<n1,n5,><n0,n6>}，L={1,2,3,4}.

Definition 2 (**Topology Encoding of tree**) Travel all nodes according to depth first and then the ordered sequent of labels is entitled Topology Encoding of tree., where express the sequence of backdating to parent node from the child node with -1, for example, the topology encoding of the tree in figure 1 is “1131- 12-1- 14- 1-12-1”.

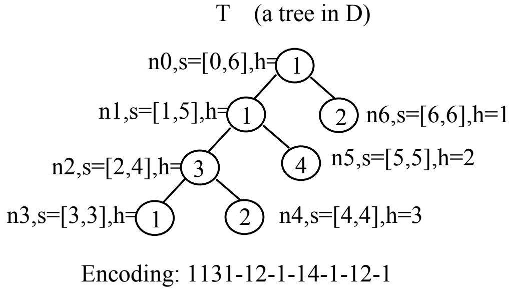


Figure 1 XML document node encoding

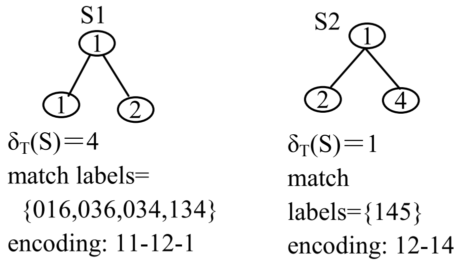


Figure 2 Embedded subtree and match labels

Definition 3 (**Triple Encoding**) Each node of ordered tree adopts Triple (tid,scope,high) Encoding, where tid∈D; scope is popedom scope of node, described as interval [l,u]. l is code of node itself; u is code nid of the last node of right sub-tree. Figure 1 is an example of node code of tree. For example, the code of root node n0 is nid=0; its right sub-tree last node is nid=6. So that scope s=[0,6]. The hierarchy (or height) of root is 0, thereby the Triple Encoding of root is (tid,[0,6],0).

Definition 4 (**Embedded Subtree**) We say that a tree $S=(r_s,V_s,B_s,L_s)$ is an embedded subtree of T=（r_t,V_t,S_t,L_t）, denoted as $S\preceq T$, provided i) $L_s \subseteq L_t$; ii) b=(x,y)∈Bs if and only x is ancestor of y in T. In other words, we require that a branch appears in S if and only if the two vertices are on the same path from the root to a leaf in T. In figure 2, S1 and S2 are embedded subtrees of T. Note that x is not always parent node of y, so as it is ancestor of y. This is difference in traditional definition of induced subtree, just ensure the completeness of mining frequent subtrees in large documents.

Definition 5 (**Support of Subtree**) the count of occurrences of embedded subtree S in a tree T is entitled support of S to T, denote $\delta_T(S)$. For example, $\delta_T(S1)=4$, $\delta_T(S2)=1$. The support of subtree S to database D is defined as:

|D| is expressed as the total of the ordered tree in database

Definition 6 (**Frequent Subtree**) If the support of subtree S to database D π(D,S) is greater or equal to smallest threshold minsup gave by user, then we say S as frequent subtree.

Definition 7 (**Match Labels**) the label sequence of all nodes of frequent subtree corresponding the tree T is titled match labels. In general, if D is forest of document sets, $\{n_1,n_2,\ldots\ldots,n_n\}$ is nodes of T, where T∈D，$\{s_1,s_2,\ldots\ldots,s_m\}$ is node set of frequent subtree S in T, then mach labels of S is $(n_{i1},n_{i2},\ldots\ldots,n_{im})$, where 1) $L(S_k)=L(n_{ik}),k=1,2,\ldots\ldots,m$; 2) each side (S_j,S_k) of S correspond to branch (n_{ij}, n_{ik}) of T, for instance, the match labels of subtree S1 and S2 is expressed in figure 2.

2.2 Structured vector notation of XML document

Construct a high dimension space with character words in the vector space model. Each word is a dimension of the space, thus XML document is viewed as a vector, $d_x=(d_{w1},d_{w2},\ldots\ldots d_{wn})^T$, where n is count of difference words of document. TFIDF（Term Frequency Inverse Document Frequency）is a common method of making document vector in vector space model. It takes the frequency of each word occurring in a document and in all document sets into account. $d_{wi}=TF(w_i,d_x)*IDF(w_i)$, where $TF(w_i,d_x)$ is count of the word w_i occurring in document d_x; $IDF(w_i)=\log(|D|/|DF(w_i))$. Where |D| is total of document sets, $DF(w_i)$ is count of document containing word w_i ; $IDF(w_i)$ is global feature of word w_i, which embody the ability of distinguishing documents.

Vector space model can express effectively document content, but cannot embody document structure information, and that XML document has abundant of structure information, which is contained in frequent subtrees. Based on this, we evolve traditional vector space model, construct a structured feature space with all frequent subtrees of documents, each frequent subtree is a dimension of this space, document (tree) d_t is viewed as a vector of structured space, $d_t=(d_{s1},d_{s2},\ldots\ldots d_{sn})^T$, where s_i is different frequent subtree; d_{si} is weight of frequent subtree s_i in document tree d_t, $d_{si}=TF(s_i,d_t)*IDF(s_i)$.Let $TF(s_i,d_t)= \delta_t(s_i)*B(s_i)$, where $B(s_i)$ is count of sides in frequent subtree; $IDF(s_i)$ is global feature of structure, whose value is smaller, the distinguishing ability of structure is better, denoted as $\log(|D|/|DF(s_i))$. $DF(s_i)$ is count of tree containing structure subtree s_i . In order to make use of all frequent subtrees, edit it to log(|D|/|DF(s_i)+0.5). In this way , we get the weight function of each frequent subtree:

$$dsi = \delta t(si)*B(si)*\log(|D|/|DF(si)+0.5) \qquad （2）$$

Document similarity measuring

Similar to VSM, FSVM model use horny cosine of frequent structure to measure document structure similarity, where Tx,Ty represent two document tree, dx(si) and dy(si) represent separately structured weight of si in document dx and dy. Seen from Eq.（3）, the process of calculating similarity of unknown document and known document is:

step1: mining all frequent subtrees in document sets;

step2: ordering the encoding frequent subtrees according to support, using Eq.（2）calculate their weight, constructing the structured feature space;

step3: calculating the structured feature vector of unknown document;

step4: using Eq.（3）to calculate similarity of two document.

The similarity of unknown document and certain document is bigger, the possibility of being classified as same class by structure is bigger, whereas the possibility is smaller. The key step is the first and the second step, now we describe below how to mining frequent subtrees.

3 The algorithm treeminer$^+$ of mining frequent subtrees

The algorithm of mining frequent subtrees is based on right-most extension increase by degrees. The basic idea is: firstly, get 1-Subtrees (contain one node tree), and choose candidate frequent subtree by calculating the support of each 1-Subtree, then generate 2-Subtrees by sharing prefix of right class extension. In order to get all possible candidate class, the algorithm must calculate the support of all subtrees build by scope-list join. According to this, we get (k+1)-frequent subtrees from k-subtrees, until build all frequent subtrees.

3.1 Related concept of frequent subtree extension

When building (k+1) frequent subtree from k frequent subtree, we firstly need confirm extension class (or candidate extension node sets). Let P be a prefix of (k-1)-subtree before k-subtree. Let $[P]_k$ refer to its class, where the form of each element is (x,i), x is label of extension node, i is the position of extension.

Theorem 1 （Class Extension）: Let P be a prefix class with encoding P, and let (x,i) and (y,j) denote any two elements in the class. Let Px denote the class representing extensions of element (x,i). Define a join operator ⊗ on the two elements,denoted (x,i) ⊗(y,j) as follows:

Case I-(i=j):

a) if P≠ϕ, add (y,j) and (y,j+1) to class [Px];

b) if P=ϕ, add (y,j+1) to class [Px];

Class II-(i>j): add (y,j) to class [Px];

Case III-(i<j): no new candidate is possible in this case.

Theorem 2 (Scope-List Join): Each element of [Px] class has own scope-list, denote £(x), where each element uses triple encoding. The operate £(x) $\cap_{\otimes}$£(y) of scope-list join is defined as :

1) if ty=tx=tid, then the triples both occur in the same tree with tid;

2) if high(x)>=high(y), then x is not ancestor of y in the tree T;

3) if lx<=ly and ux>=uy, or scope(y)⊂ scope(x), then show that y is a descendant of x in tree T. We next extend the match label m_y of the old prefix P to get the match label for the new prefix Px, and add the triple (ty,{matchlabels∪lx} ,scope(y)) to the scope-list of (u,j+1) in [Px]. (here add match label to scope-list, but delete the hierarchy information). We refer to this case as an in-scope test.

4) If ux<=ly, or scope(y)<scope(x) , then show that y is a brother of x, we add the triple (ty,{match labels∪lx} ,scope(y)) to the scope-list of (y,j) in [Px]. We refer to this case as an out-scope test.

TreeMiner$^+$ algorithm: TreeMiner+(D, minsup)

Input: triple code set D of document trees, minimum support of frequent subtree

Output: prefix encoding of all frequent subtrees S andδ_T(S)

```
1: compute support of each label, choose frequent labeled nodes F1;
2: built candidate class extension set P0={(F1,-1)};
3：prefix[P0]={};encoding[P0,k]= (F1,-1);
4 ： for each element of [P0] do Enumerate-subtrees([P0]);
5：Enumerate-subtrees([P])
6：{ For each element (x,i)∈[P] do
7：    { [Px]=ϕ;
8：        For each element (y,j)∈P do
9：            if high(x)>high(y) then
10：               {£(S) =£(x) $\cap_{\otimes}$£(y);
11：                if S is frequent then
12：                  { R=（x,i）⊗(y,j)
13：                      [Px]= [Px]∪{ R};
14:                       k+=1;
15:                       build prefix[Px] and encoding[Px,k];
16：                  }
17：               }
18：        Enumerate-subtrees([Px]);
19：    }  }
```

3.2 Actualizing example of TreeMiner+

In figure 3, there are three documents in database D, minsup=70% (minimal support), the process of mining frequent subtrees is described as below:

step 1: get triple encoding of each node of document trees by scanning database, build adjacency list of same label, entitled as scope-list. Storage data structure is shown in figure 3(b); let L={1,2,3,4,5}. Easy to calculate the support of node label 1,2,3,4 by scope-list, they are 100%, and they are frequent nodes, the support of label 5 is 1/3, is not frequent node.

Step 2: transfer recursively the process Enumerate-subtrees(), and build (k+1)-subtrees by k-subtrees. For example, 2-Subtrees in figure 4, there are four nodes in P0. Any two of P0 can be joined to build 2-subtree, so the algorithm (the sixth, eighth row) enumerate all pairs of any two element, including self-joins. But some pairs cannot build 2-subtree, such as (4,-1) and (1,-1), because label 1 is ancestor of label 4. In order to reduce the count of scope-list join, the algorithm can judge directly the possibility of constructing subtree by hierarchy of node (the ninth row). Only the nodes of being likely to building subtree are allowed to join, which avoid redundant calculation. At the same time, the algorithm limit the condition of class extension (the eleventh row): only frequent subtrees take part in class extension. In short, the algorithm reduces greatly the cost of time. Figure 4 shows the process of building 2-subtrees.

Step 3: process and store the prefix encoding and topology encoding of frequent subtrees (the fourteen, the fifteen row).

$$\mathrm{Sim(Tx,Ty)=cos(Tx,Ty)=}\ \sum\nolimits_{i=1}^{n}(dx(si)*dy(si)/\sqrt{\sum\nolimits_{i=1}^{n}dx^2(si)*\sum\nolimits_{i=1}^{n}dy^2(si)} \qquad （3）$$

Figure 4 shows the full process of mining frequent subtrees in database D. According to count of nodes and the value of support, the order of frequent subtrees is:

S1:12-134-1-1，S2:12-14-1，S3:12-13-1，S4:134-1-1，S5:12-1，S6:14-1，S7:13-1，S8:34-1

4 Similarity calculating example of document structure

On the supposition that database D is known documents in figure 3, figure 1 is known document tree, the frequency of frequent subtrees showed in table 1.

According to Eq.（2）, calculate the weight of each frequent subtree. Table 2 shows the weight value:

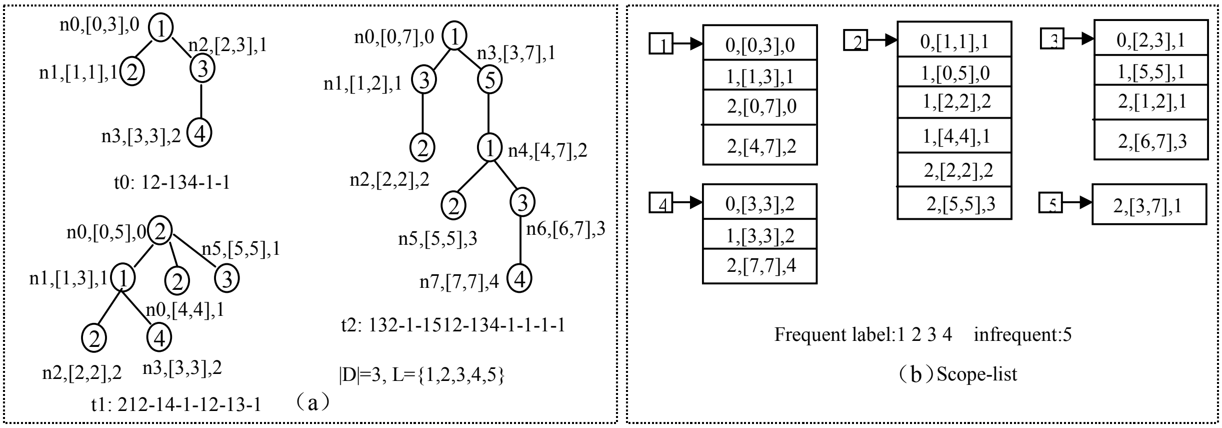


Figure 3 Document forest and data structure of node

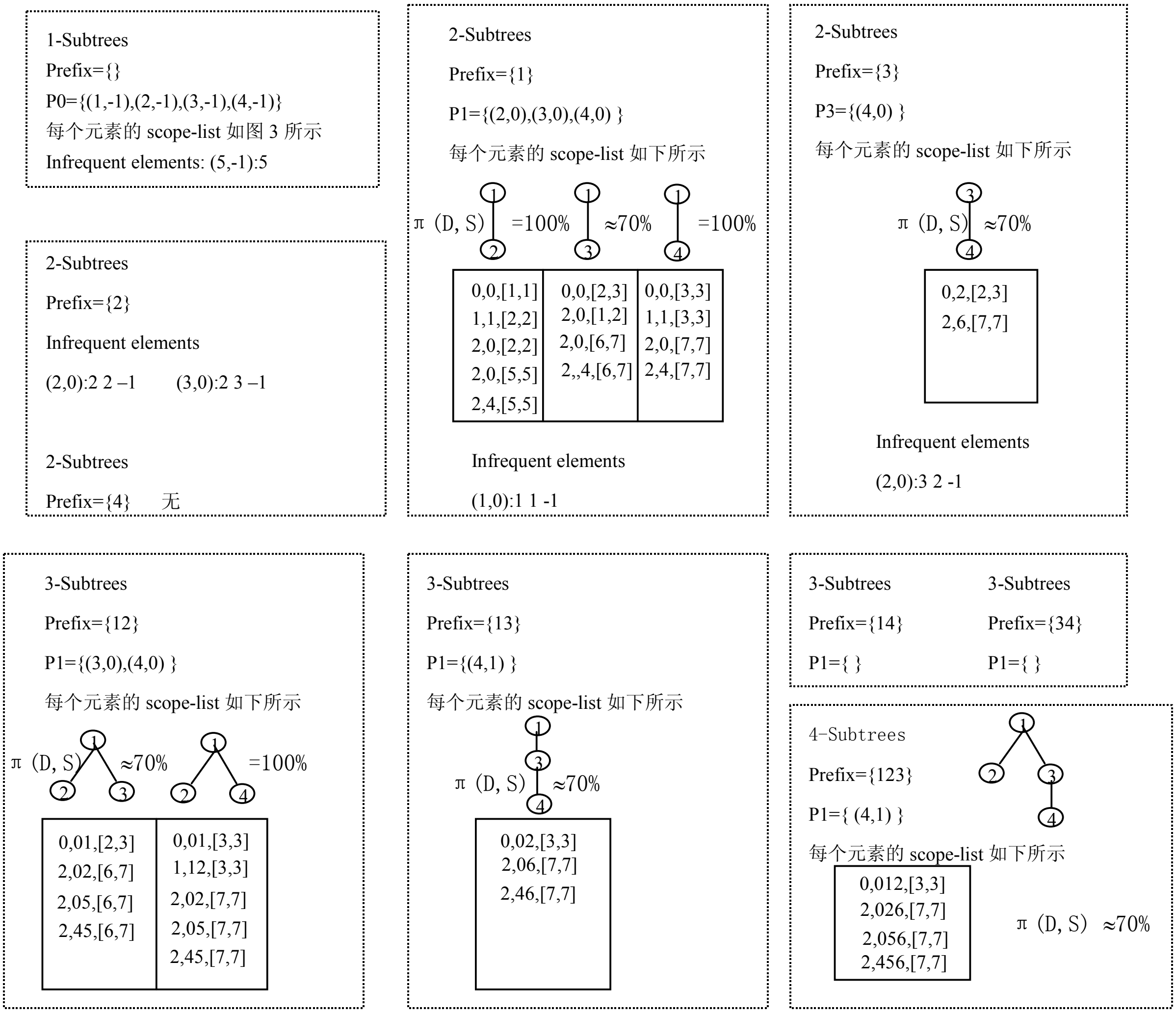


Figure 4 2-Subtrees, 3-Subtrees and 4-Subtrees building

According to the Eq.（3）, any two document similarity is: sim(t0,t1)= 0.322521, sim(t0,t2)= 0.985432, sim(t1, t2)=0.328011, sim(t,t0)= 0.549918, sim(t,t1)= 0.446359, sim(t,t2)= 0.574932

Table 1 Times of Ffrequent Subtree Occurrences in each Document

frequency	S1	S2	S3	S4	S5	S6	S7	S8
t0	1	1	1	1	1	1	1	1
t1	0	1	0	0	1	1	0	0
t2	3	3	3	2	3	2	3	1
t	0	1	2	0	3	2	2	0
B(si)	3	2	2	2	1	1	1	1
DF(si)	2	3	2	2	3	3	2	2

The result of calculating makes clear that XML document similarity measuring based on frequent structured vector model gets very high calculating precision. The experiment below has proved this by a large of data sets.

5 Experiments

5.1 Datasets and experimental designing

All experiments were performed on a 2.10GHz AMD Athlon 4000^+ PC with 1G memory running Windows 2000 Server. The algorithm is performed by JAVA programming with Java 2 Platform Standard Edition 5.0.

We use the file OrdinaryIssuePage and IndexTermsPageXML of ACMSIGMOD[20] Datasets as experimental datasets, table 3 shows the data sub-sets. The file OrdinaryIssuePage in year 1999 and OrdinaryIssuePage in year 2002 have very similar structure. The first number of amount of document is used to train mining frequent subtrees; the second number is used to test similarity measuring. The similarity threshold of same class between tested document and trained document is 99%. We tested separately on each dataset itself and between datasets; the structure similarity threshold of ACMSIGMOD-1and ACMSIGMOD-2 is 0.85, and the structure similarity threshold of IndexTermsPage and OrdinaryIssuePage is 0.3. If the experimental result is within this scope, then the data is precision, the precision of experiment use Eq.（4）to evaluate.

P=the count of document similarity is precision/ the amount of tested document （4）

Table 2 Vector Expression of Frequent Subtree in mModel FSVM

weight	S1	S2	S3	S4	S5	S6	S7	S8
t0	0.9031	0.3522	0.6021	0.6021	0.1761	0.1761	0.1761	0.1761
t1	0.0000	0.3522	0.0000	0.0000	0.1761	0.1761	0.0000	0.0000
t2	2.7093	1.0565	1.8062	1.2041	0.5283	0.3522	0.5283	0.1761
t	0.0000	0.3522	1.2041	0.0000	0.5283	0.3522	0.3522	0.0000

Table 3 Data Sub-set of Experiment

Datasets	Sources	Total num.
ACMSIGMOD-1	OrdinaryIssuePage(1999)	40+10
ACMSIGMOD-2	OrdinaryIssuePage(2002)	20+10
ACMSIGMOD-3	IndexTermsPage(1999)	40+20

In the process of training, the minimal support (minsup) of frequent subtrees in each datasets is between 0.5 and 1, see figure 5. When three datasets were combined to train, minsup is between 0.7 and 1, and we compared our TreeMiner$^+$ for mining frequent subtrees to algorithmTreeMiner. See figure 6.

5.2 Experimental Results analysis

The experiments make us know that the XML document similarity measuring method based on frequent structure vector model has very high veracity, and the minsup of frequent subtrees is smaller, the amount of frequent structures are greater; the veracity of calculating is higher. In general, the threshold of minsup is 0.7. The veracity of three combined datasets test is smaller, because the structure between the document have great difference. If we reduce the minsup, the veracity can be high. The time cost of algorithm TreeMiner$^+$ and TreeMiner mining is better with increasing minsup. When minsup is bigger than 0.7, the performance of algorithm can be improved three times, especially for large document sets, the effect is prominent. With the increasing minsup, the possibility of embedded subtrees being frequent subtrees become smaller. TreeMiner$^+$ algorithm firstly judge frequent structure, and then carry out class extension, avoiding needless node joins. At the same time, it adds hierarchy information to encoding which make scope-list join fast.

6 Conclusion

This paper studied the vector space model of expressing text content, provided XML document structure similarity measuring method based on frequent structured vector model by improving TreeMiner algorithm, and gave the model denotation and the formula of calculating weight of frequent structure. The experiments showed that TreeMiner+ is an efficient mining algorithm, and this XML document structure similarity measuring has a high precision.

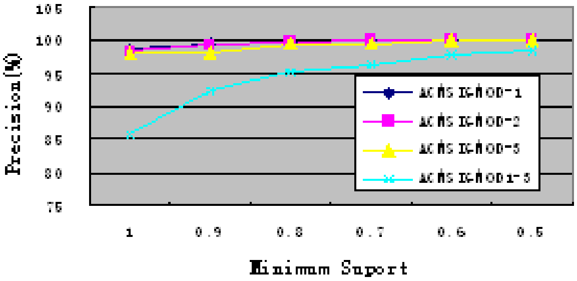


Figure 5 similarity accuracy on different data set

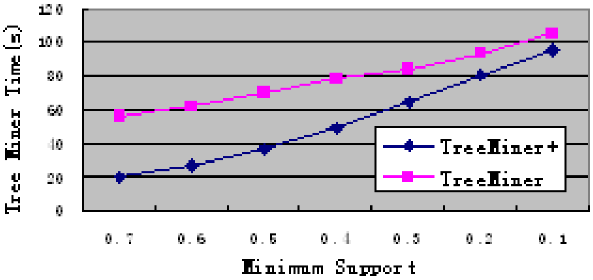


Figure 6 mining time cost of frequent subtree

XML document similarity measuring has been applied to XML document clustering query, net page classification and log analysis and so on; especially for classified query of taking into account structure and content also has good application value which will become our further study domain.

References

[1] Tian Feng,DeWittDJ,et al. The Design and Performance Evaluation of Alternative XML Storage Strategies. SIGMOD Record,March 2002,31（1）

[2] Abiteboul S, Cluet S, et al. Querying and updating the file In:Proc. of 19th International Conference on Very Large Data Bases,Dublin, Ireland 1993

[3] Kanne C,Moerkotte G. Efficient storage of XMI, data In: Proc.of the 16th International Conference on Data Engineering, 28February-3 March, 2000, San Diego, California, USA IEEE Computer Society 2000

[4] Arenas M, Libkin L. A normal form for XML documents. ACM Trans. on Database Systems, 2004,29（1）:195-232

[5] Christophides V, Cluet S, Moerkotte G, Siméon J. On wrapping query languages and efficient XML integration. In: Chen W, Naughton J, Bernstein P, eds. Proc. of the 2000 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2000. 141-152

[6] Ludascher B, Papakonstantinou Y, Velikhov P. Navigation-Driven evaluation of virtual mediated views. In: Zaniolo C, Lockemann P, Scholl M, Grust T, eds. Advances in Database Technology-EDBT 2000, 7th Int'l Conf. on Extending Database Technology. Berlin, Heidelberg: Springer-Verlag, 2000. 150-165

[7] Fankhauser P, *et al.* XQuery 1.0 and XPath 2.0 formal semantics. 2005. http://www.w3.org/TR/query-semantics/

[8] Chamberlin D, *et al.* XQuery 1.0: An XML query language. 2005. http://www.w3.org/TR/xquery/

[9] Buneman P, Fernandez M, Suciu D. UnQL: A query language and algebra for semistructured data based on structural recursion. The VLDB Journal, 2000,9（1）:76−110

[10] COBENA G，ABITEBOUL S，MARIAN A. Detecting changes in XML document[A].In Proc 18th Int Conf on Data Engineering(ICDE'02) [C]，2002

[11] CHAWATHES，RAJARAMANA，GARCIA-MOLINAH，et al. Change detection in hierarchically structured informa tion[A].In Proc ACM SIGMOD Int.Conf. On Management of Data(SIGMOD'96) [C] Montreal,Quebec,June 1996. 493-504

[12] Zhang ZP, Li R, Cao SL, Zhu YY. Similarity metric for XML documents. In:Ralph B, Martin S,eds. Proc.of the 2003 Workshop on Knowledge and Experience Manag ement(FGWN2003).Karlsruhe, 2003.255-261

[13] COSTA G,MANCO G,ORTALE R,et al. A Tree-based Approach to Clustering XML Documents by Structure[Z]. Rappoito Tecnico N.04: RT-ICAR-CS-04-04 April 2004

[14] FLESCA S,MANCO G,MASCIARI E, et al. Detecting structural similarities between XML documents[A].In Proc 5th Int. Workshop on the Web and Databases(WebDB'02) [C]. Madison,Wisconsin, 2002

[15] M. J. Zaki. Effciently Mining Frequent Trees in a Forest. SIGKDD, 2002

[16] M J Zaki, C C Aggarwal. Xrules: An effective structural classification for XML data [C]. Int'l Conf on Knowledge Discovery and Data Mining (SIGKDD'03), Washington, DC, 2003

[17] MA Hai-bing, WANG Lancheng. Efficiently Mining Unordered Frequent Trees. Mini-Macro Systems[J].Vol.27, No.11, Nov 2006. Pp.2104-2108

[18] YANG Jian-wu, CHEN Xiao-ou. Similarity Measures for XML Documents Based on Kernel Matrix Learning. Journal Of Software[J]. Vol.17,No.5, May 2006. Pp.991-1000

[19] DOUCET A ,MYKA HA . Native clustering of a large XML document collection [A]. In Proc 1st Annual Workshop of the Initiative for the Evaluation of XML retrieval(INEX'02) [C]. Schloss Dagstuhl, Germany, 2002

[20] http://www. acm. org/ sigs/ sigmod/ record/ XMLSigmod RecordMarch1999.zip ttp: // www. acm. org/ sigs/ sigmod/ record/ XMLSigmodRecordNov2002.zip

A Novel Text Clustering Method Based on DSOM-FS-FCM

Jinzhu Hu   Chun Fang   Bin He   Cong Zhang   Dongmeng Zhao   Yi Zhang

Department of Soft Engineering and Information System, Center China of Normal University,
Wuhan, Hubei 430079, China

Email: fangchun0178@sina.com

Abstract

Because of some problems existing in text clustering such as the high-dimensional sparse text data, poor efficiency of unsupervised feature selection, and defects existing in classical clustering methods and so on, a novel text clustering flow model (TCFM) and a effective text clustering approach called DSOM-FS-FCM according to TCFM are proposed. DSOM-FS-FCM fully combines the dynamic self-organizing maps (DSOM) neural network, features selection (FS) and fuzzy C-means (FCM) clustering. Experimental results indicate that DSOM-FS-FCM clustering outperforms traditional clustering methods such as K-means, and the precision is better than DSOM-FCM and DSOM clustering.

Keywords： dynamic self-organizing map neural network; text clustering; Fuzzy C-means clustering

1 Introduction

Clustering analysis is an important content in the domain of data mining, and also plays an important role in text mining [1]. The task of text clustering is to group similar documents together according to text content, which is indeed viewed as an unsupervised learning process. With tremendous growth in the volume of text documents available in Internet and digital libraries, the research of fast and high-quality text clustering methods has attracted both domestic and international scholars' wide attentions.

There are some text clustering methods currently ,such as partitioning method, hierarchical clustering method[2], density-based method[3], K-means is a classical and simple partitioning method. However, it is prone to local optimal and obtains unstable results because it needs to specify the parameter K to determine the final number of clusters and its initial cluster center is random. Hierarchical method's defect is that the process will not be revoked once the merger or split was done. Self-organizing maps (SOM) neural network proposed by T Kohonen[4] is more applicable to text clustering [5]. Nevertheless, the structure of SOM neural network is fixed and the number of neurons in competition layer must be given in advance. In view of the deficiencies in SOM, Alahakoon proposed DSOM [6], a variety of SOM that determine the shape and number of neurons in the training process.

Fuzzy C-means (FCM) is capable of text fuzzy clustering. However, FCM clustering algorithm needs to specify the number of clusters, which is an intractable problem in text clustering. Thus, it is necessary to combine DSOM method with FCM clustering. DSOM-based initial clustering determines the number of clusters and cluster centers then FCM-based further clustering improves the quality.

To apply any clustering algorithm, documents must be represented in a suitable form firstly. Vector space model is always used to represent documents. This representation raises a severe problem: the high dimensionality of the feature space and the inherent sparsity. Therefore it is highly desirable to reduce the feature space dimensionality. Feature selection is a good way to deal with this problem, It is a process that chooses a subset from the original feature set according to some criterions. The selected features retain original physical meanings and provide a better understanding for the data and learning process. Depending on whether the class label information is required or not, feature selection can be conducted with either unsupervised or supervised methods. Some supervised feature selection methods have been successfully used in text classification [7], such as Information Gain (IG) andχ2 Statistics (CHI). But due to the lack of class label, supervised feature selection methods cannot be used directly in text clustering. Therefore text clustering normally adopts unsupervised feature selection methods to reduce dimension of feature space. However, unsupervised feature selection methods can't effectively reduce dimension of feature space due to lack of class information [8]. Inspired by the literature [9], the supervised feature selection can be applied to the initial cluster results and then cluster again.

Combining all the above analysis, a novel text clustering flow model (TCFM) is proposed and a new text clustering method based on DSOM and FCM according to TCFM, called DSOM-FS-FCM, is also put forward. Firstly, preprocess the data by unsupervised feature selection. Secondly, implement initial clustering to determine the number of clusters and class information. Thirdly, use supervised feature selection to further reduce the dimension of feature space. Finally, cluster again to optimize the clustering performance. DSOM-FS-FCM clustering algorithm adopts DSOM-based method for initial clustering and FCM-based for further clustering.

2 Text Preprocess and Similarit Calculation

The general idea of text clustering is: given a documents set $D=\{d_1,d_2,...d_n\}$ without any class information, and obtain a cluster set $C=\{c_1,c_2,...c_k\}$, which makes $\forall d_i(d_i\in D),\exists c_j(c_j\in C)d_i\in c_j$ and the cost function f(C) achieve minimum value.

2.1 Text representation and feature selection

Vector space model (VSM) is widely used to represent documents. In VSM, each document is treated as a vector di and each dimension in di stands for a distinct term in the term space of the document collection. We represent any document as a vector $d_i=\{w_{i1},w_{i2},...w_{in}\},$, where w_{ij} is the term weight of term t_j in document di, which represents the significance of this term in a document. The classical weight calculating formula $tf*idf$ is defined as follow:

$$w_{ij}=tf_{ij}*idf_{ij}=tf_{ij}*\log_2(N/n) \qquad （1）$$

Where tf_{ij} is term frequency, idf_{ij} is reverse document frequency indicates the importance of the term among documents, N is the total number of documents in the collection, n is the total number of documents where term i occurs. Formula (1) supposed that the fewer the document frequency of the word is, the more important the term to cluster is, which regardless of some noise words that may appear in several documents have greater weight. In order to overcome the defect in formula (1), it is highly desirable to remove noise words before calculating the weight of terms. There are several effective feature selection methods, including two supervised methods, IG and CHI, and two unsupervised methods, DF and TC. All of these methods assign a score to each individual feature and then select features that are greater than a predefined threshold.

2.2 Similarity

To use a clustering algorithm we need to judge the similarity between two documents. We used the traditional cosine distance, represented as formula (2):

$$\cos(d_1,d_2)=\frac{d_1d_2}{\|d_1\|.\|d_2\|}=\frac{\sum_{t\in\mathrm{T}}(f(t,d_1)*f(t,d_2))}{\sqrt{\sum_{t\in\mathrm{T}}f(t,d_1)^2}.\sqrt{\sum_{t\in\mathrm{T}}f(t,d_2)^2}} \qquad （2）$$

3 A Novel Text Clustering Flow Model TCFM

TCFM takes many factors into account such as high dimensionality and sparsity of text data, poor effectiveness of unsupervised feature selection, the limitation of supervised feature selection (cannot directly apply due to the lack of class label) and advantages or disadvantages of classical text clustering methods. The specific processes of TCFM are described as the following:

1) Text preprocesses: the main works in text preprocess are segmenting words, counting term frequency, removing stop words, calculating term weights and generating vector space model. Of course, the related preprocesses vary in specific circumstances.

2) Unsupervised feature selection: Literature [8~10] made a detailed analysis and comparison on several common unsupervised feature selection methods. Appropriate methods can be chosen according to the requirement. It is worth noting that DF selection should be implemented before the weights calculation, which can't bring additional computation time and overcomes the deficiency that noise words may gain greater weight value according to the formula (1).

3) Initial clustering: the main purpose of initial clustering is to obtain the number of clusters and class information of documents. In general, dynamic clustering algorithms are suitable for initial cluster. For example, the clustering algorithm based on DSOM neural network fits well.

4) Supervised feature selection: Since initial clustering assigns a class label for each document, supervised feature selection methods can be used to further remove useless features. Literature [10] has proven that supervised feature selection methods IG and CHI can make slight increase in accuracy of the classification in the condition of removing as much as 98% features. Lacking of class information, unsupervised feature selection method is incapable to pick out more important features. So, we do the supervised feature selection after initial clustering.

5) Fuzzy clustering: the above section has analyzed that fuzzy clustering is more suitable for practical application. There are only few features that make greater contribution to cluster retained after step 4 and the number of clusters is determined after step 3. In this paper, we take advantage of fuzzy C-means for further clustering.

It is worth mentioning that each step in TCFM can choose different method according to the actual situation. For example, unsupervised feature selection can choose DF or TC or other methods.

4 A Novel Text Clustering Method Based on DSOM and FCM

According to TCFM model, a novel text clustering method DSOM-FS-FCM is proposed, which uses DSOM-based method for initial clustering and uses FCM-based method for fuzzy clustering. We give a brief introduce on DSOM-based clustering algorithm and FCM-based clustering algorithm in this section.

Algorithm 1: DSOM-based text clustering

Step 1: Generate and simplify vector space model (also called input model here). Do pretreatment and unsupervised feature selection for the entire original documents collection. Given the size of input model is N, i.e. the number of documents in dataset is N.

Step 2: Generate first neuron in the competition layer. Given the input sample set $P=(p_1,p_2,...p_n)$, each sample P_i is an m-dimension vector. Supposed the first input sample $P_1=(X_1^1,X_2^1,...X_m^1)$ and $W_1^1=P_1$. That is, $w_{j1}^1=X_j^1, j=1,2,...m$. N_c denotes the number of neurons in competition layer and N_S denotes the number of activated neurons, and K is the number of input samples. The initial values of N_c, N_S and K are 1. Array $CL[N]$ records the class labels of documents in collection and let $CL[1]=1$. Matrix $C[N_c][m]$ preserve center vector of each cluster. Initialize adjustive factor SF and the growth threshold θ according to SF.

Step 3: For the Kth input sample Pk, select the weight vector Wj that has the most similarity with input vector Pk according to formula (3).

$$D(W_j^K,P_K)=\left\|W_j^K-P_K\right\|=\min_{i=1,...N}\left\|W_j^K-P_K\right\| \qquad （3）$$

Where $\left\|W_j^K-P_K\right\|$ represents the similarity between weight vector and Pk according to formula (2).

Step 4: Finding the winner neuron according to formula (4):

$$\begin{cases} D(W_j^K, P_K) > \theta, \text{ then } j \text{ is the winner neuron} \\ D(W_j^K, P_K) \le \theta, \text{ then generate a new neuron} \end{cases} \qquad \text{（4）}$$

If there are several neurons meet threshold θ, then select the neuron with the maximum similarity. If exists neuron j as the winner, then implement operation a, otherwise, implement operation b.

Adjust the weights of the nodes which are within the neighbor of the winner neuron j and j itself, according to formula (5):

$$\begin{cases} W_j^K = W_j^{K-1} + LR(K)(P_K - W_j^K) \\ W_i^K = W_i^{K-1} + 0.5LR(K)(P_K - W_i^K), i \in N_j(d) \\ W_i^K = W_i^{K-1}, i \notin N_j(d) \end{cases} \qquad \text{（5）}$$

Where $LR(K)$ is the learning rate, which decreases with the increase of K, when K→∞, then LR→0; W_j^k and W_j^{k-1} denote the weights of the node j before adjusting and after adjusting respectively. $N_j(d)$ denotes the neighborhood of the winner neuron j in the Kth training, $N_j(d) = \{i, D(W_j^{k-1}, W_i^{k-1}) \le d/t\}$, where t denotes iteration number. Adjust $CL[K] = j, K = K+1$. and $N_s = N_s + 1$.

Generate a new neuron and make the following adjustments:

$N_c = N_c + 1$, $W_N^K = P_K$, $CL[K] = N_c, K = K+1$

If K<N, then go to step 3, otherwise go to step 5.

Step 5: In order to overcome the shortcoming of FCM that generating cluster center at random, we calculate the centers for all clusters. The center of each cluster is a weight vector, which has the greatest similarity with the competition neuron that represents the cluster. All centers are saved in matrix $C[N_c][m]$. In DSOM algorithm, the threshold θ determines whether generate a new node or not. When θ is too small, the target network will be generated soon, but the network consists of few nodes and just achieves coarse clustering. On the contrary, when θ is too big, the objective network will comprise many nodes and can achieve high accuracy clustering, but the speed of generating objective network is too slow. In order to select appropriate θ, we introduce the regulatory factors SF referred to in literature [7] and θ is defined as $\theta = SF^2/(1+1/n(t))$, where $n(t)$ is the number of network nodes in the tth training. SF is a random value in the scope of (0, 1) given by user. To get more satisfactory clustering results, SF should be chosen a lower initial value (range between 0~0.4) in the first clustering and a higher value (range between 0.4~0.7) in the second clustering by experience.

When DSOM network training is finished, N_c and matrix $C[N_c][m]$ are the cluster number and cluster centers respectively, which will be used in fuzzy C-means clustering.

Algorithm 2: Fuzzy C-means clustering

Fuzzy clustering proposed by Dunn and generalized by Bezdek, divides data sets $X = \{x_1, x_2, ... x_n\} \in R^{mn}$ into c categories, where the random sample xi belongs to class i with probability u_{ik}. The classification results are represented by a fuzzy membership matrix $U = \{u_{ik}\} \in R^{cn}$, satisfying the conditions shown in formula (6).

$$\begin{cases} u_{ik} \in [0,1], \forall i, k \\ 0 < \sum_k u_{ik} < n, \forall i \\ \sum_i u_{ik} = 1, \forall k \end{cases} \qquad \text{（6）}$$

Fuzzy C-means clustering is achieved by minimizing the objective function $J_m(U,V)$ that is about fuzzy membership matrix U and cluster center V. Function $J_m(U,V)$ is defined as formula (7):

$$J_m(U,V) = \sum_{i=1}^{n} \sum_{i=1}^{c} (u_{ik})^m d_{ik}^2 (x_k, v_i) \qquad \text{（7）}$$

Where $U = \{u_{ik}\}$ is a fuzzy membership matrix and meets all conditions defined in formula (6). $V = \{v_1, v_2, ... v_c\} \in R^{pc}$ are focal points for clusters collection, $m \in [1, \infty)$ is the weight index. The research results of Nikhil showed that the best value range is 1.5~2.5 and the ideal value for m is usually equal to 2. The distance between the Kth sample and the center of the ith class is defined as formula (8):

$$d_{ik}^2 (x_k, v_i) = \|x_k - v_i\|_A^2 = (x_k - v_i)^T A (x_k - v_i) \qquad \text{（8）}$$

Where A is a $P \times P$ positive definite matrix, when $A = I$, the distance defined by formula (8) is Euclidean distance. FCM algorithm is repeated iteratively to optimize the objective function (7). We detail the FCM algorithm in the following part.

Step 1: Initialize the cluster center $V = \{v_1, v_2, ...v_c\}$. Here, V is initialized according to the matrix obtained in DSOM initial clustering.

Step 2: Calculate the membership matrix U according to formula (9):

$$u_{ik} = \left[\sum_{j=1}^{c}\left[\frac{d_{ik}(x_k, v_i)}{d_{jk}(x_k, v_i)}\right]^{2/(m-1)}\right]^{-1} ,k=1,2,\cdots,n \qquad （9）$$

Step 3: Update the cluster center by formula (10):

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m} \qquad i=1,2,....c \qquad （10）$$

Step 4: Repeat steps (2) (3) until the formula (7) converged.

5 Experiment and Result Analysis

In order to better assess TFCM clustering model and DSOM-FS-FCM clustering method, the DSOM-FS-FCM method is compared with the other three methods which are K-means clustering method, DSOM-based clustering method, and DSOM-FCM method that combine dynamic SOM neural network with fuzzy C-means but not do supervised feature selection according to TCFM.

5.1 Evaluation Criteria

Usually, recall Re and precision Pr are the two indexes to test the cluster effect, the definition is as follows:

$$\Pr_i = \frac{TP_i}{TP_i + FP_i}, \quad \mathrm{Re}_i = \frac{TP_i}{TP_i + FN_i} \qquad （11）$$

Where Rei represents the integrity of the cluster, $\Pr_i$ represents the accuracy of the cluster, And, FP_i is the number of the documents which are assigned to the C_i category by mistake ,but it actually belongs to other kind category, TP_i is number of the documents which are assigned to the C_i category correctly, FN_i is number of the documents which are belongs to the C_i category, but they are assigned to other kind category by mistake.

Moreover, Precision is also an very intuitive evaluation criterion [8].For each cluster, it is likely to comprise documents belong to several different classes. We choose the dominant class shared with most documents in this cluster as the final class label. The precision for each cluster is defined as formula (12). To avoid the possible bias that small cluster has very high precision, we calculate the weighted average of precision shown in formula (13).

$$Precision\ (A) = \frac{1}{|A|}\max(|\{d_i \mid label(d_i) = c_j\}|) \qquad （12）$$

$$\Pr ecision = \sum_{k=1}^{G'} \frac{|A_k|}{N} precision(A_k) \qquad （13）$$

5.2 Experimental result and Analysis

The corpus used for experiment is the “Chinese text classification corpus” given by Ronglu Li published on the website http://www.nlp.org.cn, and contains 4 categories with 75 documents in each. DSOM-FS-FCM method is implemented strictly following the steps of TCFM. Preprocess is firstly done and 4165 distinct words are retained. Filtering features that occurs in less than 3 documents or more than 60 documents and then 735 distinct features are left. Do initial clustering according to DSOM and then different parameters are set for CHI method to select 345, 242, 189 and 150 features respectively, which account for 7%, 5%, 4% and 3% of the number of features. the dataset for the mentioned four methods is the same.

Table 1 when 189 characteristic words are chosen，the experimental results of the four clustering methods

	DSOM-FCM		K-means		DSOM		DSOM-FS-FCM	
	Re%	Pr%	Re%	Pr%	Re%	Pr%	Re%	Pr%
Politics	90	91.2	81	76.4	82	83.6	89	91.5
Transportation	71	89.5	78	68.4	77	72.5	95	94.7
Medicine	82	87.6	66	70.1	69	78.6	91	89.8
Computer	91	84.3	74	79.5	81	80.1	86	91.7
Average	85.5	88.1	74.7	73.6	77.2	78.7	90.2	91.9

When chooses the different number of the characteristic words, the four method's results as shown in Figure.3.

From Table 1 and Fig.3, we can get the the following conclusions:

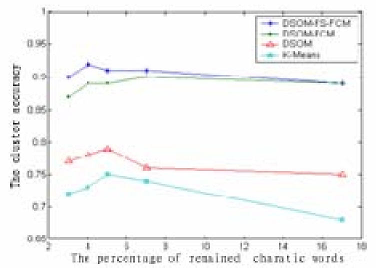


Figure 3 the clustering precision of four clustering methods

(1) DSOM clustering accuracy is better than K-means. Because K-means clustering exists some drawbacks such as random initial cluster centers, unknown cluster number and local optimization.

(2) Fuzzy C-means clustering or K-means clustering is superior to DSOM-based clustering method when their initial cluster centers and the number of clusters are determined.

(3) Supervised feature selection methods play an important role in clustering from the result comparison between DSOM-FCM and DSOM-FS-FCM. Because the unique difference of the two clustering methods is that DSOM-FS-FCM adopted supervised feature selection (χ^2 Statistic (CHI)) while DSOM-FCM didn't do this. The precision of DSOM-FS-FCM is up to 92% in condition of removing 96% features. However, with the percentage of reserved features increasing, the difference of performance between the two methods diminishes and converges at last. This change trend results from mislabeling some documents in DSOM clustering, which declines the effectiveness of feature selection.

6 Conclusions

In this paper, the advantages and disadvantages of current popular clustering methods are analyzed at first, then a novel text clustering flow model is proposed, which fully combines the characteristics of feature selection methods and clustering methods. Based on the proposed TCFM, a novel text clustering method DSOM-FS-FCM is presented. In the same condition, we compare DSOM-FS-FCM clustering method with other three text clustering methods DSOM-FCM, DSOM and K-means. The experimental results show that our proposal DSOM-FS-FCM outperforms other three methods. It also indicates that DSOM-based text clustering method is superior to K-means clustering but inferior to K-means clustering method, which initial centers and the number of cluster are known.

References

[1] Jian-Suo Xu, Li Wang. TCBLHT: A New Method of Hierarchical Text Clustering. Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005. Page(s):2178-2181 Vol.4

[2] L Kaufman, P J Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley & Sons, 1990

[3] M Ankerst, M Breuning, H P Kriegel, et al. OPTICS: Ordering points to identify the clustering structure. In: Proc of the 1999 ACM SIGMOD Int'1 Conf on Management of Data (SIGMOD'99). New York: ACM Press, 1999. 164~169

[4] Kohonen T. The Self-organizing Maps[J]. Proceedings of the IEEE, 1990, 78(9): 1464~1480

[5] Lei Jing-Sheng, Ma Jun, and Jin Ting. A Fuzzy Clustering Technology Based on Hierarchical Neural Networks for Web Document. Journal of Computer Research and Development [J]. 2006, 43(10): 1695~1699

[6] Alahakoon D, Halgamuge S K. Dynamic Self-organizing Maps with Controlled Growth for Knowledge Discovery. IEEE Transactions on Neural Networks, 2000,11(3):601~614

[7] M. Rogati, Y. Yang. High performance feature selection for text categorization. The CIKM-02, Mclean, 2002

[8] Luying Liu, Jianchu Kang; Jing Yu; Zhongliang Wang. A comparative study on unsupervised feature selection methods for text clustering. Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on 30 Oct.1 Nov. 2005 Page(s):597~601

[9] L. Tao, L. Shengping, C. Zheng, et al. An evaluation on feature selection for text clustering. The ICML03, Washington, 2003

[10] Liu Tao, Wu Gong-Yi, Cheng Zheng. An Effective Unsupervised Feature Selection Method for Text Clustering. Journal of Computer Research and Development [J]. 2005, 42(3): 381~386

An Empirical Study on Online Shopping Intention of Chinese City Consumers

Linhai Wu[1] Shijiu Yin[1] Lili Du[2]

1 School of Food Science and Technology, Jiangnan University, Wuxi, Jiangsu , 214122, China

Email: ysjchenmo@126.com

2 School of Business, Rizhao Ploytechnic College, Rizhao, Shandong, 276826, China

Email: laura821207@163.com

Abstract

This paper utilizes 562 sample data collected from four cities---Jinan, Qingtao, Zibo and Rizhao, builds the Logit model, then based on which Eviews3.1 statistical software is employed to process sample data by Logit regression. The result of the study indicates that online shopping intention of city consumers is more affected by factors like consumers' income, network usage capability and the security of online shopping and less affected by such factors as consumers' age, education level and convenience. Finally, suggestions are put forward on promoting the development of online shopping in China on the basis of measurement result.

Keywords：Online shopping; Intention; City consumer

1 Introduction

With special characteristics, Internet surpasses time and space to share and effectively transit information and resources, which makes the society enter into information age. Especially since 1990s, the development of Internet has deeply influenced every aspect of social life and greatly enhanced the economic development of the whole society. With a rapid development of Internet, online shopping, as a new model of individual consumption, invites more and more attention.

Up to June 2007, the number of netizens in China has amounted to 162 million and takes the second place in the world, only next to America which has 211 million netizens. However, compared to the rapid development of Internet, it is obvious that Chinese Internet users are lack of positive perception and trust on online shopping. Only 25.5 percent of Internet users purchase online, which forms a striking contrast against 71 percent that of America [1]. Therefore, it is necessary to do research on consumers' motivation of online shopping and decision-making process to identify the main factors influencing online shopping in order to promote orderly and healthy development of online shopping and upgrade development level of modern service industry.

Scholars from home and abroad have conducted a lot of research on factors influencing online shopping. Han Yan built multi-layer SEM model to analyze consumers' intention to purchase online and pointed out that perception risk was the key factor influencing consumers' online shopping[2]. He Qiguo, Sun Qiang and Zheng Ranran utilized structure function and structural model of consumers' perception of value to analyze consumers' attitude and Internet experience which may affect their online shopping decisions[3][4][5]. Sandra M and Forsythe Bo Shi, Thompson S H Teo and Yon Ding YeOng looked into many factors influencing online shopping respectively[6][7]. Cass A O and Fenech T studied the behavioral mechanism of Net-surfers' online consumption[8]. Lin and Hsiu-Fen used empirical test of competing theories to study consumer intention to online shop[9].

From the current research, it can be seen that meteorological measurement is seldom used to carry out the empirical research on online shopping intention. Therefore, this paper uses empirical research as basic means to build Logit model, utilized meteorological measurement to make empirical analysis on factors influencing city consumers' online shopping and puts forward suggestions on promoting the development of online shopping in China on the basis of measurement result.

2 Data sources and model setting

2.1 Data Sources

Data of this paper come from previous survey, which combines Internet survey and on-site survey together. In November 2007, fifty students were chosen from Economy College and Management College of Qufu Normal University and Department of Business Administration of Rizhao Polytechnic College to implement the survey in four cities--- Jinan, Qingtao, Zibo and Rizhao. The objects of this survey are citizens all with Internet usage experience. This survey adopted the method of sample survey and was carried out by distributing questionnaires in supermarkets. Four hundred questionnaires were distributed and 371 were effectively returned. The recovery rate was 92.75 percent. From January to November in 2007, 956 e-questionnaires were released by means of e-mail and Internet forum, of which 191 were effectively returned after getting rid of those filled out by citizens beyond the survey scope and unqualified questionnaires. Among 191 e-questionnaires returned, 155 objects had done online shopping, taking up the proportion of 81.15 percent. 562 effective questionnaires were returned through two channels. The result of this survey indicates that the survey scope is extensive from the perspective of demographic characteristics such as age and income, which mainly fits the social structure of those surveyed cities.

2.2 Empirical Model

This paper focuses on influencing factors of city consumers' online shopping in China, which indicates if city consumers choose to purchase online. The relationship between online shopping intention and the affectting factors is concluded as the following function: consumers' online shopping intention = F (age, education degree, annual income...) + random disturbing factor. This paper takes if consumer online shopping as a dependent variable, e.g 0-1 type dependent variable (when doing online shopping, y=1;

Table1 Introduction of Model Variables

Variable	Definition of Variable	Mean	Std.Dev.
Explanation of Variable			
X1	Below 30=1; 30～40 =2; 40～50 =3; 50～60 =4; above 60=5	1.769	0.930
X2	Female=1; male=2	1.427	0.496
X3	Below middle school=1; middle school=2; university=3; postgraduate and above=4	2.527	0.692
X4	Unmarried=1; married=2	1.310	0.471
X5	Below 30 thousand=1; 30～70 thousand=2; 70～100 thousand=3; over 100 thousand=4	1.779	0.850
X6	Very inconvenient=1; not very convenient=2; general condition=3; convenient=4; very convenient=5	1.932	0.755
X7	Very secure=1; relatively secure=2; general condition=3; not very secure=4; very unsecure=5	1.644	0.794
X8	Very proficient=1; relatively proficient=2; general condition=3; not proficient=4; knowing nothing=5	2.722	0.757
X9	Less than one hour=1; 1～3 hours=2; 3～7hours=3; 7～14 hours=4; over 14 hours=5	2.107	0.872
X10	Yes=1; no=2	1.377	0.486
X 11	Never=0; 1～5 imes=2; 5～10 times=3; over 10 times=4	2.011	0.646
X12	Surburbs=1; relatively remote area=2; ordinary area=3; prosperous city centre=4	2.779	0.605
Explained variable			
If doing	Yes＝1,no＝0	0.516	0.501

otherwise, y=0) . Assuming the probability of y=1 is P, the function of y is as follows:

$$f(y) = P^y(1-P)^{1-y}; y = 0,1$$

This paper adopts Logit model of binary choice, confines the number of dependent variable within [0-1] scope and utilizes the maximum likelihood estimating method to compute its regression parameter. The basic form of Logit model is as follows:

$$P_i = F(\alpha + \sum^m \beta_j X_{ij} + u) = 1/\left\{1+\exp\left[-(\alpha + \sum^m \beta_j X_{ij} + u)\right]\right\}$$

In this function, "Pi "is the probability of online shopping, "i" Serial number of consumer, "βj" Regression parameter of influencing factors, "j" Serial number of influencing factors, "m" the number of influencing factors, "Xij" independent variable, representing the influencing factor "j"in the sample "i", "α" Intercept, "u" error term.

2.3 Variables Setting

According to the above analysis, when investigating if city consumers conduct online shopping, this paper uses twelve variables, namely consumers' age(X1), gender(X2), education degree(X3), marriage(X4), annual income(X5), convenience(X6), security(X7), Internet usage capability(X8), net-surfing time(X9), possession of Internet banking cards(X10), usage of banking cards(X11), location of residence(X12). Model variables and description of statistics are shown in table1.

3 Result of utilizing model and discussion

3.1 Result of Utilizing Model

This paper uses Eviews3.1 statistical software to process sample data by Logit regression. First of all, the author introduces all variables into regression equation to implement investigation to regression coefficient to result in model one (see Table 2); and then adopts backward screening method to get rid of unobvious variables until all variables are statistically significant (α=0.05), so to result in model two(see Table 3). The connotation of these two models is identical; therefore, in the following discussion, it focuses on Model Two.

3.2 Discussion

Based on the computation result, main factors influencing city consumers' online shopping intention, significance and influencing degree are concluded as follows:

1. Consumers' income, evaluation of online shopping security and Internet usage capability have a significant impact on online shopping. Consumers' income has a positive effect on online shopping. It means the same to traditional shopping, a higher income represents higher purchasing power or a higher income makes consumers pay less attention to online shopping security, which makes them more willing to purchase online. Evaluation of online shopping security has a negative effect on online shopping. Consumers are suspicious of online shopping security, so security is still an important factor restricting the development of online shopping. Internet usage capability has an obvious

Table2 Logit Model Result Considering All Variables

Vari-able	Coefficient	Std. Error	z-Statistic	Prob.
C	2.766631	1.644793	1.682055	0.0926
X1	-0.258874	0.192558	-1.344395	0.1788
X2	0.078828	0.394176	0.199982	0.8415
X3	-0.160625	0.225972	-0.710819	0.4772
X4	0.050757	0.289978	0.175037	0.8611
X5	0.440722	0.199491	2.209226	0.0272
X6	-0.211342	0.239791	-0.881360	0.3781
X7	-0.397603	0.209977	-1.893556	0.0583
X8	-0.433175	0.193666	-2.236716	0.0253
X9	-0.053173	0.175103	-0.303668	0.7614
X10	-0.469271	0.417232	-1.124724	0.2607
X11	0.362984	0.226962	1.599317	0.1098
X12	-0.190240	0.265650	-0.716131	0.4739
LR statistic (12 df)		47.64145		
McFadden R2		0.122390		
Log likelihood		-170.8095		

Table3 Logit Model Result after Phasing Out Unobvious Variables

Vari-able	Coefficient	Std. Error	z-Statistic	Prob.
C	1.397745	0.679587	2.056758	0.0397
X5	0.511104	0.164728	3.102713	0.0019
X7	-0.520037	0.175681	-2.960125	0.0031
X8	-0.507197	0.182952	-2.772299	0.0056
LR statistic (3df)		40.05855		
McFadden R^2		0.102909		
Log likelihood		-174.6009		

impact on online shopping. Online shopping is a newly-rising means of shopping, which requires consumers to possess some relevant Internet knowledge and skills, such as searching information, knowing information about retail websites, using computers and buying processes. With more Internet experience, consumers will acquire more online shopping skills and information sources and then will be more likely to purchase online. Besides, Miyazaki and Fernandez point out that although risk is an important factor blocking consumers from purchasing online, most of the risk perception results from consumers' unfamiliarity to this brand-new means of shopping. Therefore, pure Internet experience and skills can reduce risk perception and thus enhance shopping intention and actual purchasing [10].

2. Consumers' age, gender, marriage status and education degree have no obvious impact on online shopping. Generally speaking, the older the consumers are, the less they are willing to purchase online because youngsters are easier to accept new things and more willing to do online shopping. But our research is not supportive of this point. Consumers' age, gender, marriage status and education degree are only served as a threshold to determine if consumers are net-surfers and can't determine if they purchase online.

3. Evaluation of shopping convenience and location of residence in the city have no obvious effect on online shopping. It means logistics industry has developed rapidly in recent years. Network of logistics distribution has been improved and logistics distribution is no longer a main factor restricting the development of online shopping in China. What is more, factors like possession of Internet banking card, usage of credit card in daily life and weekly average net-surfing time have no obvious impact on online shopping.

4 Connotation of policy

According to the research, city consumers' online shopping intention is mainly affected by income, evaluation to online shopping security and Internet usage capability. In the process of promoting the development of online shopping, different kinds of factors should be taken into consideration. On this basis, this paper puts forward suggestions as follows.

4.1 Improving Internet Knowledge Level

From the model result, it can be seen that Internet usage capability is an important factor influencing online shopping. Therefore, on one hand, the proportion of courses relevant to information and Internet should be increased in academic credentials education to improve students' Internet capability. On the other hand, government should provide policy support and necessary capital guide and carry out training and disseminating of Internet to improve the level of informationization in China.

4.2 Improving Consumers' Degree of Belief in Online Shopping Security

First, main innovators should be encouraged to improve security technology of online shopping and enhance online shopping security. Second, knowledge of Internet security are to be disseminated to make consumers acquire correct and overall recognition to Internet security. Third, it needs to speed up building network and e-commerce regulations, strengthen supervision over online selling and regulating the order of online market to promote healthy development of online shopping.

4.3 Lowering Consumer's Risk Perception

As for shopping websites, they have to lower consumer's risk perception and improve degree of belief in online shopping to transform consumers from a "browser shopper" to a real purchaser. Adopting strategies like price guarantee, free sample, goods returned guarantee, the third party's inspection and trust establishment can lower consumer's risk perception and thus improve consumers' recognition to the website to attract consumers to purchase online. As for the whole B to C network selling industry, it needs to build the third party authentication mechanism, grade the degree of trustworthiness of shopping websites to help consumers perceive risks and improve consumers' degree of belief in online shopping security.

References

[1] CNNIC, "The Twentieth Statistical Report on Chinese Internet Development", 2007

[2] Han Yanmin, "Analysis on Multi-layer SEM Model of Consumers' Online Shopping Intention", Statistics and Decision-making, No.4, 2007, pp.9~11

[3] He Qi-guo, LIin Mei-hua, "An Empirical Research on Factors Affecting the Adoption of Online Shopping", Economy and Management, No.10 , 2006, pp.44~49

[4] Sun Qiang, Si Youhe, "Study on the Composition of Customer Perceived Value on Internet Shopping", Research on Science and Management, No.7, 2007, pp.185~187

[5] Zheng Ranran, Song Ze, "An Empirical Research on Factors Influencing Online Shopping Intention", Business Economy and Management, 7, 2007, pp.55~61

[6] Sandra M,Forsythe Bo Shi, "Consumer Patronage and Risk Perceptions in Internet Shopping", Journal of Business Research, No.5, 2003, pp.867~875

[7] Thompson S H Teo,Yon Ding YeOng, "Assessing the Consumer Decision Process in the Digital Market Place", International Journal of Management Science, No.31, 2003, pp.349~363

[8] Cass A O, Fenech T, "Web Retailing Adoption: Exploring the Nature of Internet Users Web Behavior", Journal of Retailing and Consumer Services, No.10, 2003, pp.81~94

[9] Lin, Hsiu-Fen , "Predicting consumer intentions to shop online: An empirical test of competing theories", Electronic Commerce Research & Applications, Vol.6, No.4, 2007, pp.4337~442

[10] Miyazaki A D, Fernandez A., "Consumer Perception of Privacy and Security Risks for Shopping", The Journal of Consumer Affairs, Vol.35, No.1, 2001, pp.27~44

4WS Vehicle Virtual Prototype Modeling and Dynamics Simulation Based on ADAMS

Ying Yang Ting Li Siping Yan

School of Mechanical Engineering and Automation, Northeast University, Shenyang, Liaoning, 110004, China

Email：yangyingsy@163.com

Abstract

A new style 4WS vehicle virtual prototype was established based on ADAMS. The proportional control relationship of the front and back wheels' angles plus the feedback control relationship of yaw rate were presented. The transient response and the steady-state performances of 4WS and 2WS vehicles were analyzed, and the influence of tire lateral stiffness on steering performance was studied. The simulation results indicate that the 4WS vehicle takes on rather agility at the low speed and excellent stability in at the high speed. The future developing trends with 4WS system control technology was prospected in the end.

Keywords：4WS; Vehicle; Dynamics; ADAMA; Virtual Prototype

1 Introduction

Digital virtual prototype technology is an important way to reduce vehicle development cycle, lower development costs and improve the quality of the product. With the mature of virtual product development and virtual manufacturing technology, the computer simulation technology has been used extensively. Many famous automobile enterprises have established the environment of the design and development of digital virtual prototype; many products have been achieved digital fully [1]. System dynamics simulation is a main factor of virtual prototype technology. Vehicle dynamic characteristic is very important to vehicle. In order to reduce the risk of product development, it is very necessary to make dynamics computer simulation according to digital virtual prototype.

The traditional 2WS vehicles have a lot of shortcomings such as the slow steering response at low speed, not very flexible at steering and poor stability at high speed. In order to improve the performance of steering handling, enhance the driving stability and increase the quality of comfortable and security. With the continuous development of automotive technology, the continuous improvement of speed, the technology of 4WS which can realize the initiative safety of automobile has been applied [2].

Especially in recent years, People not only in the pursuit of vehicle safety and comfort, but also requires better performance and flexibility of manipulation. In this article, 4WS vehicle virtual prototype was established based on ADAMS which is multi-body dynamics simulation software of mechanical system and then make a simulation analysis about the transient response and the steady-state performances and the influence of tire lateral stiffness on steering performance between 4WS and 2WS under the closed-loop control.

2 Establishment of 4WS vehicle virtual prototype

Set up a 4WS vehicle virtual prototype. Including the suspension system of front and rear, steering system of front and rear, body, engine and tires. Both front and rear suspension system have adopted the dual-arm suspension model. Consult the principle of front wheel steering system, establish the 4WS rear wheel steering system and the relationship of four-wheel proportional and the rate of sidelong angle control. It has become realization that the rear-wheel steering system controlled by closed-loop initiatively. After the preliminary completion, compared simulation result and theoretical calculations, modify and adjust the models and algorithms again and again, finally we have been relatively satisfied with the 4WS vehicle virtual prototype model.

Table 1 Main parameters of body frame

Mass(kg)	1300	Distance of front wheel and center of mass(mm)	1185
Wheel space(mm)	2560	Yaw moment of inertia(kg.m^2)	3000
Front tread(mm)	1520	Obliquity of front and rear wheels(°)	-0.5
Rear tread(mm)	1500	front and rear wheels restrict angle (°)	0
Front el lateral stiffness(N/rad)	44400	Rear wheel lateral stiffness(N/rad)	43600

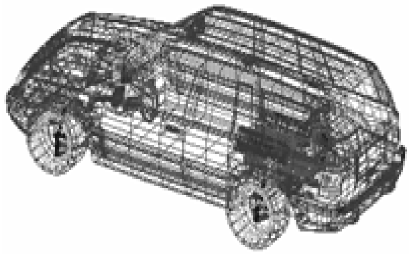


Figure 1 4WS vehicle virtual prototype model

3 Establish active closed-loop control model

There are many ways to control four-wheels and the H2 and H ∞ multi-target integrated with high robust [3]**.** It can improve the manipulate stability while require measure many state variables at the same time and especially cannot measure side-slanting angle by reasonable cost. This reduced the usefulness of the algorithm greatly. Currently, most of the studies we researched are measured by many state variables and adopted complex algorithm to achieve better performance by tracking models. But it is not very clearly that which the best model is with being tracked performance. At the same time, it increase the high cost of hardware by request steering control of front and rear wheels.

At present, the main method used in the vehicle is the control method of the relationship between the front and rear control ratio. This method is not perfect in many ways because it is constitutes of a feedback forward compensation and it does nothing to outside interference. Early research on 4WS system and design of control system are based on the assumptions of linear dynamic equations. But the changes of vehicle dynamics parameters through the actual moving process can not fully meet the actual needs according to this theory. If the lateral acceleration is large when the car turning, tire cornering properties will enter nonlinear region, which can not be researched by linear control theory. This article from virtual prototype perspective, on the basis of taking into account the ratio of front and rear wheels control method .Then integrate yaw rate feedback control further and can be formed the closed-loop control system with the ability of feedback to the outside interference and optimize their performance.

Two freedom mathematical model of 4WS car is:

$$\begin{aligned}\frac{d\beta}{dt} &= \frac{-2(k_1+k_2)}{mu}\beta - \left(\frac{2(ak_1-bk_2)}{m^2u}+1\right)\omega_r \\ &\quad + \frac{2k_1}{mu}\delta_f + \frac{2k_2}{mu}\delta_r \\ \frac{d\omega_r}{dt} &= \frac{2(ak_1-bk_2)}{I_z}\beta - \left(\frac{2(a^2k_1+b^2k_2)}{I_zu}+1\right)\omega_r \\ &\quad + \frac{2ak_1}{I_z}\delta_f + \frac{2bk_2}{I_z}\delta_r\end{aligned} \quad (1)$$

Algorithm can be summarized as formula (2):

$$\delta_r = k_\delta \delta_f + k_\gamma \omega_\gamma = \frac{k_\delta}{n}\delta + k_r \omega_\gamma \quad (2)$$

Let yaw angle equal zero we can draw form formula (1):

$$k_\delta = -\frac{k_1}{k_2} \quad (3)$$

$$k_r = \frac{mu^2 + 2(ak_1 - bk_2)}{2k_2 u} \quad (4)$$

δ_r ,rear rotary angle； δ_f , front rotary angle; δ ,Steering wheel angle ; ω_r , yaw rate; n , Transmission ratio of steering system; m ,Vehicle mass； α ,distance of center of mass front axle； b , distance of center of mass to back axle; u, vehicle speed; β, yaw angle of center of mass; k_1, k_2, lateral stiffness of front and back tires respectively; I_z, Yaw moment of inertia.

4 Performance simulation analysis of 4WS vehicle

4.1 Simulation test of 4WS steady-state performances

The purpose of rotary steady test is to measure the steady raw response when the input of car steering reached the stable drive state. Vehicle steady-state rotary characteristics (US-OS) have an important impact on the manipulation control of vehicles and have a veto to active vehicle safety. So it is always attracted attention by automobile sector. With the development of computer technology, there are three main methods to simulate vehicle steady rotary performance: continuous acceleration simulation by fixed angle of the steering wheel, continuous acceleration Simulation of unchanged turning radius and simulation of the unchanged steering wheel angle. In this article, we use the second method to do Steady-state rotary simulation experiment according to standard ISO4138-82. In this experiment, automobile always turn to the right along a circle radius of 30*m* and adjusting the steering wheel angle constantly, accelerate slow and even, until the vehicle reaches the maximum lateral acceleration of 6.8 m/s^2. The rotary steady test is shown in Figure 2.Subheadings should be in a bold font or underlined lower case with initial capitals. They should start at the left-hand margin on a separate line.

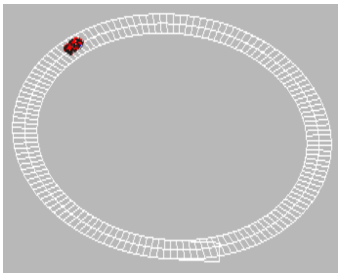


Figure2 Rotary steady simulation tests

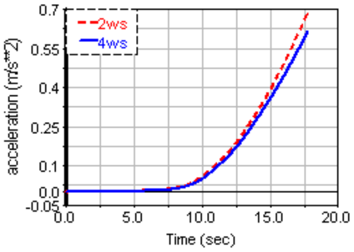


Figure3 Lateral acceleration

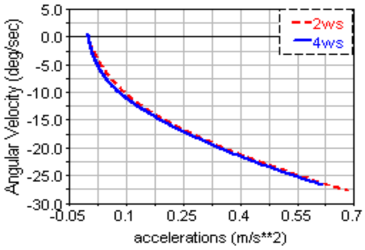


Figure4 Relations between changes of yaw rate and lateral acceleration

From Figure 3, we can see that the lateral acceleration rate of 4 WS change less than 2WS.It shows that the regulating role of rear wheels bring on that the stability and security of 4WS is better than 2WS when turning direction at the same radius and the unchanged acceleration. But with the increase in lateral acceleration, gyration radius of 4WS monotonely increasing faster than 2WS, the inadequate turning trend of 4WS is bigger than 2WS.

From Figure 4, we can know the yaw angle of 4WS is larger ith the same lateral acceleration. That illuminates that the 4WS is easier to adapt the situation of steady-state turning. Meanwhile, with the increase in lateral acceleration, the yaw rate of 4WS and 2WS tend to identical as the regulating role of the rear wheels. Overall speaking, the steady-state of 4WS is better than the steady-state of 2WS.

4.2 Vehicle transient simulation test

Vehicle transient performance is a main factor in evaluating the vehicle handling stability of the railway vehicle. There are many ways to evaluate the transient characteristics such as angle step, angle pulse, power step, power pulse, steering return ability, lane, double lanes, and sinusoidal input and so on. In this article, we adopt the conventional angle step test to compare the transient characteristics of four-wheel steering with two-wheel steering. Angle step test set the speed of 100*km / h*, steering wheel angle of 22d and step time of 0.4 s by GB/T6323.2-94 standards. Compared with 2WS and 4WS in the respects of transient characteristics control of ratio and yaw rate. The 2WS represents of the front wheel steering, 4WS represents the four-wheel steering, and the simulation results are as follows:

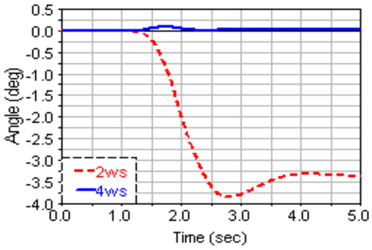


Figure5 Comparison of lateral declination of center of mass by angle step

From Figure 5 and Figure 6 we can see that 2WS car has a high steady-state value of yaw rate and lateral declination of center of mass in high-speed turning and the tracking performance of body is bad. Automobile will induce pendulum of rear easily when turning in high-speed and it increase difficulty in the operation of the driver.

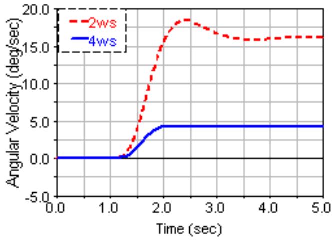


Figure 6 Comparison of yaw rate by angle step

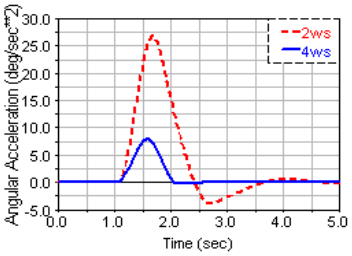


Figure 7 Comparison of yaw acceleration by angle step

The steady-state value of lateral declination of center of mass is near to zero in Figure 5 and the transient response time is short, fast response and so on. Therefore, 4WS car has better stability and security; furthermore road tracking performance is good when turning in high speed. The steady-state value of 2WS is negative, that shows 2WS has dumping trends in a steady state, security and stability is bad relatively.

In Figure 6, the yaw rate overshoot of 4WS car is about 75 percent less than the value of 2WS. It can enhance the stability of the vehicle, and get a short time to achieve steady-state, so that the flexibility of vehicle increased. The driver's subjective evaluation, security and maneuverability are enhanced, and not easy to produce drift phenomenon.

In Figure 7, the yaw angular acceleration peak value of 4WS is less than the value of 2WS. It illuminate that the turning radius increased at same speed, at the same time, it needs the driver turn the steering wheel frequently and leads the fatigue of drive. However, the time to reach steady-state of 4WS is shorter than 2WS, and then the rapid response and road tracking performance are improved.

4.3. Influence of vehicle steering characteristics by tire lateral stiffness

Table 2 Tire lateral stiffness and experimental data

NO.	k_1	k_2	k_δ	k_r
1	44400	43600	-1.0183	1.4892
2	53362	73421	-0.7268	0.8802
3	43260	83421	-0.5186	0.7716

From formula (3), (4), we can see that k_δ , k_r are changed when the tire lateral stiffness changes. Now we study the changes of four-wheel steering performance when pass and change lane under tires of different stiffness. Generally, use the 70 percent (or 100*km / h)* of the maximum speed of the vehicle to do single lane test. Set the cycle is 3.14 and the biggest steering angle is 11.3degree.

From Figure 8 and Figure 9, we can see that the lateral displacement reduced along with the decrease of absolute value k_δ . It improves high-speed stability performance, but it also reduces the yaw rate, increase turning radius and operational flexibility will be reduced. In other words, it do harm to pass and change lane at high-speed. However, from Figure 10, we can see that with the reduction of absolute value k_δ , the steering wheel torque reduced and the driver operation easier.

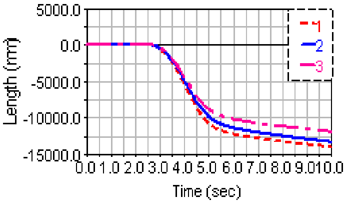


Figure 8 Comparison of 4WS lateral displacement

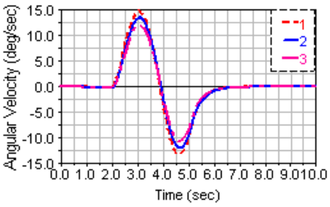


Figure 9 Comparison of 4WS yaw rate

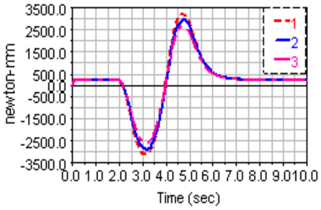


Figure10 Comparison of 4WS steering wheel torque

In actual traveling process, the tire lateral stiffness is changing. For example, the vertical load of wheels will change a lot if the vehicle emergency braking when turn direction. The tire lateral stiffness of front wheels will be increased and the lateral stiffness of rear wheels will be reduced according to the characteristics of tire cornering force. Based on Figure 8 and Figure 9, increase of lateral displacement and yaw rate peak has worsened the vehicle handling stability and caused serious loss control of steering. Therefore, the selection of appropriate tire lateral stiffness of the car is very important to the safety and stability of vehicle operation.

The stability conditions of 4WS:

$$\frac{bk_2-ak_1}{I_z}+\frac{(a+b)k_1k_2k_\gamma}{mI_zu}>0 \qquad (5)$$

Therefore, we must select the appropriate front wheel lateral stiffness in the premise of ensuring stability. From what has been discussed above, we can see that the inadequate steering is increasing along with the increase of k_r .

5 Conclusions

In this article, we make use of mechanical dynamics simulation software to establish the model of four-wheel steering prototype. Meanwhile, the simulation study of vehicle steady state has been done according to the method of active feedback control of four-wheel angle proportional and yaw rate. At the same time, the impact of different tire lateral stiffness to vehicle steering characteristics has been analyzed. Research shows that the virtual prototype technology has an important practical and engineering significance. It can simulate the vehicle handling performance quickly and accurately, thus making prediction and assessment of its performance and provide the basis and reference to design products; it also can reduce development costs, shorten the design cycle and improve the competitiveness of products. Furthermore, we can built a more accurate model and consider more factors to make other handling stable tests. In these ways, we can predict and evaluate the vehicle manipulation stability in the round.

References

[1] Chen Liping, Zhang Yunqing, Ren Weiqun, Tan Gang.Dynamic Analysis of Mechanical Systems and Application Guide of ADAMS [M]. Beijing:published by Qinghua University.2005

[2] Lu Qiang. Active Control of Four-wheel steering vehicle - based on neural network method [D].Jilin University of Technology .1996

[3] Yu Fan, Lin Yi. Vehicle System Dynamics [M]. Beijing: Machinery Industry Press .2005

[4] Ren Weiqun. Virtual prototype of car - road system

published by electronic industry .2005

[5] M Abe .”Vehicle dynamics and control for improving handling and active safety: from four-wheel steering to direct yaw moment control” [DB/OL] . IMechE.1999

[6] Bian Mingyuan. Four-wheel steering technology and its development prospects [J / OL]. Research and Development.1999 (12)

[7] Lin Ming. Automotive Engineering Manual - the first fascicle [M]. Beijing: Machinery Industry Press. 1984

[8] Yang Ying, Sheng Jing and Zhou Wei “The Monitoring Method of Driver’s Fatigue Based on Neural Network”, IEEE ICMA2007, Harebin, 2007, 3555-3559

[9] Ying Yang, Wei Zhou，Guang-yao Zhao “Driver’s Face Image Recognition for Somber Surroundings based on computer vision”, DCABES2007, Hubei, pp1125-1128

[10] Ying Yang, Dongliang Zhu “The Simulation Research of Control Arithmetic for Automobile ABS Based on MATLAB ”International Conference on Mechanical Engineering and Mechanics 2007,Wuxi, 2058-2062

Maintenance Oriented Condition Monitoring and Fault Diagnosis Expert System for Excitation System*

Zeng Hongtao[1] Liu Yajin[1] Xiao Zhihuai[1] Guo Jiang[1] Tian Zhihu[2]

1 Hubei Provincial Key Laboratory of Fluid Machinery and Power Equipment Technology, Wuhan University
Wuhan, 430072, China
Email: zenghongtao@hotmail.com

2 China Three Gorges Project Corporation, Yichang, 443002, China
Email: tian_zhihu@ctgpc.com.cn

Abstract

As the availability and maintenance costs of the equipments directly links the economic benefits and market competitiveness of the enterprises, more and more attention paid on the condition monitoring and diagnosis of the equipment in the hydropower plant. The condition monitoring and diagnostic analysis of the excitation system has been studied. An expert system has been developed. The configuration and the software architectures of such a system are presented. The approaches to the condition monitoring and the acquisition of the expert knowledge as well as their experiences are discussed in detail. The fundamental functions of the diagnosis expert system are descried. This system has been developed and put into practice. The results show it is good.

Keywords: Excitation System; Condition Monitoring; Diagnosis; Maintenance

1 Introduction

There has been more and more attention paid on the condition monitoring and diagnosis of the equipment in the hydropower plant, because the availability and maintenance costs of the equipments directly links the economic benefits and market competitiveness of the enterprises. Excitation system is an important component of the synchronization generator set, which directly impacts the performance characteristics of the generator. Excellent excitation system not only ensures the stable and reliable performance of the generator, but also improves the Technical & Economic Indicators of the generator and power system. As the most important auxiliary equipment of the generator set, the excitation system should change to condition monitoring from preventive scheduled maintenance as the development of the equipment maintenance system. Considering the reliability and availability, we should strengthen the monitoring and fault diagnosis of the equipment, consider the requirement from all aspect of the hydropower plant, to optimize the maintenance.

As the universal application of the digital excitation regulator, the excitation system has achieved essential function in condition monitoring and fault diagnosis, such as over- excitation limit, low- excitation limit, and PT breaking detection. While all these function mainly paying attention on advancing the security of the set, only giving alarm signals after the accident on equipment failure, which have not enough consideration on the requirement of equipment maintenance and can not raise the suggestion of equipment maintenance, isolate the implementation of decision-making and maintenance organizations in maintenance activity with other management in enterprise production process and can not achieve the high quality, high reliability and high efficiency of the

* This work is supported by Fund of Hubei Provincial Key Laboratory of Fluid Machinery and Power Equipment Technology Foundation, 200720801002

enterprise overall.

Therefore, under the frame of Information Share and comprehensive integration of control, maintenance and technology management, developing the system of condition monitoring and diagnosis of the excitation equipment, to accurately estimate the health status, performance degradation and developing trend of the equipment, and assist the expert to confirm the correct maintenance decision, proper time and the exact spot to achieve the maintenance automation gradually, reaching the aim that advancing the quality of maintenance while reducing the cost of maintenance, forming the unified-enterprise intelligent system, striving to obtain the greatest economic and social benefits.

The application object of this system is the static silicon controlled digital excitation system of the #19 set in Gezhouba hydropower plant as in Fig 1.

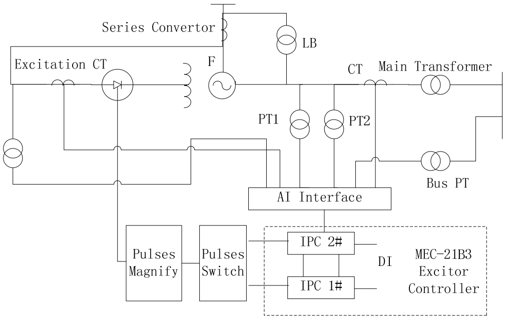


Figure 1　Excitation system of Gezhouba hydropower plant

2　System Architecture

The unit of condition monitoring and diagnosis of the excitation equipment system is a modern test and analysis system with a kernel of industrial computer, which can not only detect the signal, but also analyze and process the gained signal and give the diagnosis result.

Hardware Structure

The unit of condition monitoring and diagnosis of the excitation equipment system is composed by interactive interface between industrial computer and excitation regulator, on-line special monitoring device, pretreatment chip, mobile expert workstation, remote communications interface (Figure 2).

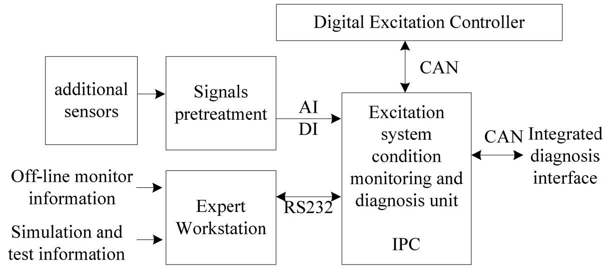

Digital Excitation Controller
CAN
additional sensors
Signals pretreatment
AI
DI
Excitation system condition monitoring and diagnosis unit
IPC
CAN
Integrated diagnosis interface
Off-line monitor information
Simulation and test information
Expert Workstation
RS232

Figure 2　The hardware structure of condition monitoring and diagnosis system for excitation system

The principal part of the condition monitoring and diagnosis of the excitation equipment is industrial computer with high anti-jamming capability and high reliability for data analyzing and processing.

The interactive interface of excitation regulator does the bus communication through the CAN chip to obtain data from the digital excitation regulator.

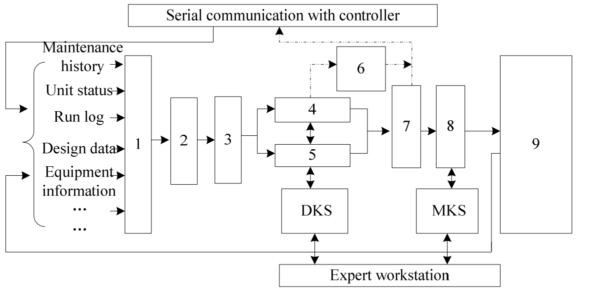
The on-line special monitoring device mainly surveys the equipment status messages which have not been collected by the digital excitation regulator.

The mobile expert workstation, which is a laptop with expert application software, does the message communication between the maintenance experts at scene and the unit of the monitoring and diagnosis of the excitation by the RS232 serial communication interface. One mobile expert workstation can work with all the units of the monitoring and diagnosis of the excitation in the plant, which can save the cost.

The remote communications interface does the interaction between the unit of the condition monitoring and diagnosis of the excitation and the integrated diagnostic server of the upper units, through the Ethernet NICs.

Software Framework

The analyses and process of the signal is the core technology of the condition monitoring and diagnosis of the excitation equipment software. By processing and transforming the monitoring signal, we can pick-up the sensitive fault symptom, combining the expert knowledge from the diagnosis repository, estimate the sort, spot, development trend and harm of the fault, and give the correlative maintenance suggestion (figure 3).



1 Data analysis 2 Information collection 3 On line analyse 4 Fault predict 5 Fault diagnosis
6 Parameters optimization 7 Control decision 8 Maintenance suggestions 9 CAN Communication

Figure 3 The Software function framework of condition monitoring and diagnosis system for excitation system

The information collecting module mainly numerate the data collected, pack it by sort, and write it into the corresponding data buffer.

The data analyzing, adopting the digital filter, spectrum analysis such as Fourier transform, uses different methods of data analyzing aiming at different signals to get their own characteristic vectors.

As the condition changing of the generator and the element aging and replacing of the excitation system, the characteristic of the system will drift in the actual operation, so we have to use the online identification to optimize the parameter.

According to the request of the characteristic of amplitude frequency margin, phase frequency margin and speed-transform rate for the excitation system of national standards, we can use the frequency domain FFT (Fast Fourier Transform) /LSE (least square estimator) or time-domain identification based on EE (equation error) model to determine the optimize control parameter, and to achieve the optimize characteristic of the system. We can also get the omen of the system characteristic degradation by the online identification.

The diagnosis module is consisted with fault prediction and fault diagnosis. During the degradation of the equipment characteristic, the fault prediction module will forecast the development trend of the system status, give the maintenance suggestion in time, and avoid the happing of the fault. The main function of the fault diagnosis module is to decide the exact spot and character of the fault occurred according to the result of the data analyzing. The fourth part of this paper will explain this function in detail.

The control decision-making and maintenance suggestion give the remediation step and maintenance scheme mainly through the diagnosis result. It will only give suggestion on the unit of condition monitoring and diagnosis of the excitation equipment; after integrating the entire sets and plant by the upper diagnostic workstation, it will decide the final maintenance scheme.

The Ethernet communication module pack the current state, equipment fault number, function fault number, and the maintenance scheme number to upload to the integrated diagnostic workstation of the units, and to get some other information useful to the excitation system diagnosis.

The diagnosis repository contains quantitative information of the fault, including the characteristic threshold, occurrence probability, harm intensity, equipment number, and function number; the maintenance database includes maintenance scheme number corresponding with the faulted excitation equipment and history maintenance data.

The experts at scene can not only search the history data of the equipment, but also write the artificial detection and diagnosis information into the repository to update the data, and to perfect the system by the mobile expert workstation interactive module.

The unit of condition monitoring and diagnosis of the excitation also has the function of self-learning and self-improving because it can update the data of the repository itself and perfect the expert knowledge continuously.

3 Condition Monitoring

The base of equipment performance evaluation and fault diagnosis for the expert system is obtaining the equipment condition information in time. Considering the restrictions of the current technical condition, there have three ways been adapted by the digital excitation regulator to improve the economic benefits of the plant, which are real-time getting data from excitation regulator, online special detection, and offline detection and testing.

At the early time of condition maintenance in power plant, we monitor and diagnose the significant equipment of the excitation system following the requirement of the users. The excitation system can be divided into following parts: rotor windings, excitation transformer, excitation convertor, digital regulator, thyristor; pulse transformer, current &voltage transformer, fan, switch, field circuit breaker, de-excitation resistance, carbon brush, slip ring and quick-fuse.

Getting Data from Regulator

The MEC series of digital excitation regulator of the Gezhouba hydropower plant has been integrated part function of condition monitoring. To save fund and guarantee the reliability of the system while not adding any superfluous hardware, the unit of condition monitoring and diagnosis of the excitation system gets the related information from the regulator by the RS485/422 interface for the condition message been collected by the microcomputer governor (Table.1), and write it into the database to make it be shared between the regulators.

Table1 The Main Information Acquired from Digital Regulator

No	Status Variables	Status Variables
1	Stator Voltage	On Line Number
2	Stator Current	Wind Motor Power Signal
3	Excitation Current	6 Phase Pulses Signal
4	Synchronous Voltage	Force Exciting Limit Signal
5	Reduce Magnetic Cmd.	Active Power Overload Signal
6	Start Cmd	Fire Failure Signal
7	Text Height	Fuse Melt Signal
8	Rotate Speed	Field Breaker Status
9	Column Separation	PT Fault Signal
10	Bus Voltage	Excitation Limit Signal
11	Excitation Voltage	V/Hz Limit Signal
12	Reactive Power	Synchronous Fault Signal
13	Frequency	Run-reserve Switch Status
14	Add Magnetic Cmd.	Terminal Breaker Status
15	Fire Cmd	Power Fault Signal
16	Shut down Cmd.	Regulator Parameters

Online Special Detection

If the digital regulator hasn't collected some important condition quantity of the excitation system, while it can be online monitored continuously, we should add relevant detection devices such as sensors and transmitters. We directly read the condition information from the detection device if the equipment has been installed special monitoring device. The electric signals online collected will be adjusted and amplified by the pretreatment board, converted by the AD, and stored into the computer.

For examples: for the rotor circuit, insulation monitoring device has been installed into many hydropower plants, so we can get the insulation resistance from the rotor circuit and calculate the rotor temperature through the voltage, current and temperature coefficient; for the fan, monitor the wind speed and wind pressure; add temperature detection device for the silicon controlled; add temperature online monitoring device for the excitation transformer.

Offline Detection and Testing

By the restriction of technology, we can't and needn't use online monitoring for all the condition quantity. We can use the offline detection for the condition quantity with slower changes, minor auxiliary establishment and some equipment which can't be monitored online technically.

The excitation operation and maintenance staff should do the patrol-inspection, spot-inspection and simulation test periodically or irregularly, including:

- the state of the indicator, such as the signal track record of the power supply, failure alarm and position of the switch;
- the data recorded of the meters indication, such as the reactive power, voltage, current and the frequency of the switch;
- the indication record of the transformer temperature;
- the measure of the direct current flowing through the carbon brushes;
- the insulation resistance measure of the excitation transformer and the excitation series-wound convertor;
- performance testing of the excitation system by the excitation simulation testing device;
- performance parameter measurement of the equipment after repairmen;
- medium ullage factor test of the transformer;
- the factor measure of the rectifiers current

sharing and voltage sharing;

The maintenance staffs write the collected data into the mobile expert workstation, and store the data into the diagnosis repository and maintenance repository of the excitation monitoring and diagnosis unit through the serial communication to update the database.

4 Diagnosis Analyzing

The diagnosis analyzing includes two aspects: fault prediction and diagnosis positioning, which is according to the characteristic parameters from the data analyzing and processing, using related knowledge and experience (including dynamic principle and fault mechanism of the system equipment, and the expert knowledge of the design, installation, operation and maintenance) to identify, diagnose, and forecast the development trend of the excitation system condition, and provide technology support to the further maintenance decision-making.

Table 2 Fundamental Functions and Their Performance

Function	Performance Indexes
Fire	Fire Time
	Overshoot
	Oscillating Number
Force Excitation	Crest Value Multiple
	Forcing Time
Exciting Magnetism	Extinguish Magnetism Time
Voltage Regulator	Regulator Scale
	Terminal Voltage Regulating Windage Rate
	Static Windage Rate
	V/Hz
	Regulate Time
	Rated Respond
	Control Precision
	Overshoot
	Oscillating Number
Signal Measurement	Delay time
	Measure Precision
	Measure Scale

Because the excitation system is consisted with several equipments, with the aging or conking down of the facilities, definitely it will cause the performance falling down or the function breaking down of the system. Therefore, according to the index evaluation of the system performance, we can judge whether the maturity of the function or not and estimate whether or not the damage of the equipment element which is gonging to realizing the very function. According to the guideline of the national standards [3], the function and capability of the excitation system is in the following Table 2.

The diagnosis result can be divided as function fault and facility fault according to the character of the fault.

The facility fault will definitely cause the happening of the related function fault.

The function fault includes excitation failure, inverter failure, loss of field, faulty forced excitation, faulty decreasing excitation, voltage regulation fault; the facility fault includes PT disconnection, rotor disconnection, fan fault, SCR breakdown, synchronous transformer disconnection, fusing of the quick-fuse, pulse transformer insulation breakdown, excitation transformer insulation breakdown.

We can exactly determine the very function condition of the excitation system by monitoring the real time state of the operating system. According to the performance index demand of the very condition, we can judge whether the performance decreasing or the function fault of the system or not.

If the performance is decreasing, we should use the mathematics method to fit the operating characteristic curve, estimate the system state under the condition that the external environment remains unchanged in a certain future, and figure out the time, spot and character of the function fault and facility fault which is likely to happen.

If the function fault is happening, combining the expert knowledge and experience, we will diagnose the very fault equipment by the way of state estimation with model reference, parameter identification and human-computer interaction.

I'll show the diagnosis flow through the following excitation of the set.

The national standards prescribe that: when the synchronous generator promotes the voltage from zero, we should guarantee the over regulation of its terminal voltage not exceeding 15％ of the rating, the time of the voltage swing not exceeding 3, the regulation time not exceeding 10s. The performance index designed for the MEC series of digital regulator is that the over-regulation Mp not exceeding 5％ of the rating, the time of the voltage swing N not more than 3, the regulation time ts less than 5s.

We can easily get the system condition of under constant terminal voltage excitation (or constant terminal current excitation, or tracking the system voltage excitation) through monitoring the switch quantity, terminal voltage and current of the generator, and excitation voltage and current. Then the online identification module does the curve fitting with terminal voltage signal of the generator processed by the data analyzing module, to get a set of terminal voltage data of the generator varying with the time. Finally we can get the related performance index according to the calculation formula including regulation time, over-regulation, and the times of the voltage swing.

If the Mp<5%、N<3 and ts<5s, it suggests that the excitation has succeed. Then the set goes into the next condition favoring.

If 5%<Mp<15% or 5s<ts<10s and N<3, it suggests the performance of the system is falling down. Usually it’s the controlling parameters of the regulator which cause the excitation performance falling down. We should use the parameters identification to optimize the controlling parameters.

If Mp>15% or N>3 or ts>10s, then the function fault that excitation failure will occur in the system. There are a lot of reasons which will cause excitation failure, as following Fig 4.

Once the excitation function fault occurred, according to the fault tree, we can in turn check the characteristic quantity of the equipment which may cause the fault, and diagnose it with the expert knowledge. If we has not gotten enough condition quantity or not known the reason of the fault, we should diagnose with the human-computer interaction, makeup the deficiency of getting omen automatically, and determine whether the fault was caused by the facility fault or the improper operation of the work staffs.

5 Expert Database

The key to diagnose the system fault is analyzing and understanding the mechanism of the facility fault, and making the best use of statistic data and expert knowledge to serve the economic and safe operation of large units. Besides using the existing document and knowing the way of judgment and treatment to the ordinary fault, also we should use the FMEA (Failure Mode and Effect Analysis) and FTA (Fault Tree Analysis) on the facility, collecting the experience from the maintenance staffs at scene and getting the behavior and threshold of the fault through moldering and simulation, to form the inerratic expert knowledge and set up the expert database for diagnoses and maintenance. Here I pay my emphases on the method of collecting the expert experience and getting the analyzing knowledge.

Expert Experience Repositories

Combining with the tree list of the excitation system and investigation table for design, we will translate the expert experience, abstract knowledge, and character description of design, manufacture, installation, operation and maintenance from related field into expert knowledge module which can be easily recognized by computers.

According to the function to the equipment element of the excitation system, list the tree list from system to element to find out the implementing element for the very function. According to the physical structure the excitation system can be divided into six parts, which is regulation counters, rectifier screen, deexcitation screen, excitation transformer, excitation streaming changer, rotor circuit (Table 3).

Combining the tree list of the excitation system to do the FMECA to every single element, list the possible fault mode, reason, consequences of every element, classify and estimate the possible fault mode by the severity and possibility to analyze the corresponding detection method, prevention steps, and compensation control measure(Table 4).

Table 3 The Example for Excitation Equipments Tree

System	Facility	Function Module	Structure Module	Component
Excitation system				
	Rectify cabinet			
		Rectify bridge		
			Controllable silicon	
			Resistance -capacitance loop	
				resistance
				capacitance
			Wave motor	

Table 4 Expert Knowledge and Experiences Investigations

Level	Function	Element	Failure mode	Cause

Effect	Manifestation	Criticality	Examination Approach	Probability

Analysis Repositories

Now the fault simulation and analysis based on model has becoming an important resort to study the fault character [4][5]. Set up the element module model corresponding with the practical physical elements after careful analysis on the structure, function principle, fault mode and consequence of the excitation system equipment. We use the bond graph approach [6] to model and simulate the excitation system by integrative consideration on state quantity of the equipment and the link connection among the energy flow, information flow and object flow of the each module.

Basing on the advancement of modeling and simulation with bond graph approach, we can do such following things:

(1) By simulating the performance and condition of the system or equipment, we'll understand the facilities fully, study different conditions of the equipment under all kinds of operation condition, and find the relevant key position and weakness part to adopt the preventing measure.

(2) Simulating the influence, development trend and consequence of all kinds of fault.

(3) Simulating the implementing effect of all kind of maintenance to cut down the blindness maintenance.

(4) Simulating the reliability and capability of the equipment after maintenance.

(5) Destructive test on the substituted parts to accumulate the fault data of the equipment.

(6) Simulating the reliability of the maintenance on the whole electric power system.

By the simulation test, we'll get the mathematic presentation value of the equipment and element characteristic quantity on the fault and after the very maintenance.

For instance: simulating and modeling on the thyristor using the bond graph approach as in Figure 4.

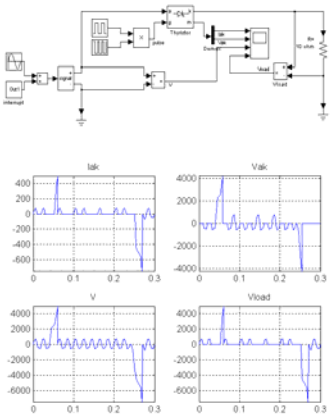


Figure 4 The thyristor model and it's fault simulation

The simulation result is shown above. The anode voltage of the thyristor is frequency sine voltage. From 0.04～0.1s, it controls the trigger pulse to shut down the thyristor, while the by the interfering of the anode voltage, at 0.4s the positive voltage of the thyristor V exceeds the positive turning voltage, the thyristor conducts by accident and spontaneously turns off at the next zero passing point; from 0.14～0.2s the thyristor turns off ,while with the interfering, the increasing rate of the anode voltage of the thyristor exceeds the critical, the thyristor conducts by accident for half cycle;

Also from 0.24～0.3s, it controls the thyristor to turn off, with the interfering signal, the thyristor reversing voltage V exceeds the reversing turning voltage, and the thyristor is reverse breaking down and permanently conducting. We can examine the current Iak passing the thyristor and voltage Vak of the tow ends through the observation port.

6 Conclusion

The unit of condition monitoring and diagnosis of the excitation equipment designed in this paper is an important part of the condition monitoring and diagnosis system, which will improve the operation and maintenance quality of the excitation system to heighten the longevity of equipment and the reliability of the system, and ensure the validity and economy of the excitation system maintenance, thus it will have a extensive application future.

At the same time, this system is designed for the digital excitation device launched, while the research result also can be applied into the new excitation equipment, it will become new equipment which conforms the development trend of the hydropower plant maintenance mechanism if we integrate the function of condition monitoring and diagnosis to the excitation system.

References

[1] C.Allan Morse, et al. Digital excitation enhances performance and improves diagnostics [J]. IEEE Industry Applications Magazine, 2001, 7(2): 28-36

[2] Wang Liang, et a).(A method of time domain identification based on EE model for the excitation system parameters)[J].(Automation of electric power systems), 2002, (8): 25~28, 37

[3] Excitation system for synchronous electrical machines-Technical requirements of excitation system for large and medium synchronous generators. (Technical standard of People's Republic of China), 1997

[4] D.A.Linkens, et al. Fault diagnosis based on a qualitative bond graph model, with emphasis on fault localization [C]. Qualitative and Quantitative Modeling Methods for Fault Diagnosis, IEE Colloquium on, 1995: 1-6

[5] T.Kohda, et al. Identification of system failure causes using bond graph models[C]. Systems, Man and Cybernetics, 1993. 'Systems Engineering in the Service of Humans', Conference Proceedings, International Conference on, 1993, 5: 269 -274

[6] Hao Houtang, Li Zhaohui. Modeling and simulation of power unit for an excitation control system with bond graph approach [J] International journal hydroelectric energy, 2000(1): 32~34

[7] Jiang Guo, "Man-Machine Fusion Techniques based on Virtual Reality and CI-Agent in Maintenance of Power Plants"[D]. Huazhong University of science and Technology, 2003

[8] Jiang Guo, Zhaohui Li, Yitao Chen, "Visualization of a Hydro-electric Generating Unit and Its Applications", 2003 IEEE International Conference on Systems, Man and Cybernetics, vol. 3, pp. 2354-2359, 2003

[9] Jiang Guo, Zhaohui Li, Yitao Chen, Yuewu Wang, Shijie Chen."Virtual environment conception for CBM of hydro-electric generating Units", Proc 2002, International conferencee on power system technology, vol. 3, 1957-1961, 2002

[10] Hongtao ZENG, Zhaohui LI, Yaxiong BI. "Hydroelectric Enterprise Model Based on Community Intelligence", Proceedings of IEEE International Conference on Systems, Man & Cybernetics, Volume: 7,10-13,October, 2004, pp6137-6142

[11] Hongtao ZENG, Zhaohui LI, Yaxiong BI." A New Engineering Management System of the Three Gorges Project Based on Community Intelligence", Proceedings of International Engineering Management Conference, October, 2004, pp297-301

[12] Bellamine, M.; Abe, N.; Tanaka, K.; Taki, H. "Remote machinery maintenance system with the use of virtual reality", 3D Data Processing Visualization and Transmission, 2002. Proceedings First International Symposium on, pp. 38-43, 2002

[13] YU Ren, ZHANG Yong-gang, YE Lu-qing, LI Zhao-hui, "The analysis and design method of maintenance system intelligent control- maintenance-technical management system and its application [J]", Proceedings of the CSEE, 21(4), pp. 60-65, 2001

[14] GUO Jiang.Man-Machine Fusion Techniques Based on Virtual Reality & CI-Agent in Maintenance of Power Plants [D], 2003

[15] JIANG Guo, Zhaohui Li, Yitao Chen. Visualization of a Hydro-electric Generating Unit and Its Applications. 2003 IEEE International Conference on Systems, Man and Cybernetics, 2003(3): 2354~2359

Research on Custom Satisfactory Evaluation in Enterprise Based on BP Neural Network

Xu Jun

School of Economy and Management, Henan Polytechnic University, Jiaozuo, Henan, 454000, China
Email: lmmxj@tom.com

Abstract

The enterprise custom satisfactory is the concentrated body of the relationship between the enterprise and the custom, and it directly determines the enterprise's competitive power. Based on the theories of CS, the custom satisfactory index system of the enterprise is put forward, and the artificial neural network evaluation model of custom satisfactory is established, carrying on the authentic proof operation, and realizing the science evaluation and judge to custom satisfactory.

Keywords: Custom Satisfactory; Index System; BP Neural Network

1 Introduction

At the turn of a century, the intense market competition and People's changing consume views and needs, make the customer become the focus of the enterprise. The process of the enterprise begins with demand of the customer, ends with the satisfaction of the customer, forming a closed, continued improving and creative system; someone who can best understand the customer's expectation, master the customer's satisfied degree in time, and then make fun use of the owned resources to satisfy and surmount the customer' loyalty, and be always in the invincible position. Therefore, the measurement and evaluation of the custom satisfactory become one of the enterprise maker's most important subjects.

2 Custom satisfaction concept and index system

2.1 Outline of Custom Satisfactory

Custom satisfactory is the difference function between the custom's perceived result to the product and service (include the feeling to the paucity, price and value) and the expected result [1-2]. When the perception is lower to the expectation, the customer will be dissatisfied, and even have complaint or appeal, and if take positive measures and proper solution to the complaint, it may change the customer's dissatisfaction into satisfaction, and even make them become loyalty, when the perception is higher to the expectation, the customer will be satisfied, when the percept is far above the expectation, the customer may be changed from satisfied customer into loyalty customer.

2.2 Custom Satisfactory Evaluation Index System

The enterprise's customer includes the exterior customer and internal customer [3-5]. The exterior customer is to aim at the customer; the internal customer includes the enterprise employee. The degree of employee's participation and initiation affect the consumer's satisfied degree in a large extent. If the employee themselves can't be satisfied, it's impossible for the customer to be satisfied. The company of Federal Express even thought that it's beyond imagination that an internal employee who dissatisfied his enterprise can provide satisfactory service to the exterior customer. So, internal customer should be put ahead, and the whole index, encouragement level, education and training and employee's relationship; these five aspects should be taken into consideration.

The custom satisfactory of enterprise's exterior customer can be acquired through product quality, service quality and enterprise image. The product quality can be subdivided into product function, product credibility and product economy. Service quality includes service attitudes and after-sale service. The enterprise image includes enterprise credibility, enterprise concept of value and the employee qualities. The product function is the customer's satisfied degree about the product quality index sign. Product credibility is the credibility of the quality index sign of the product which the enterprise provides. Product economy means the price and usage cost and so on. Product's delivery ability is the enterprise's ability to deliver goods on time according to the contract and service efficiency. Service attitude is the attendance's work attitude in the process of before, on and after the sale. After-sale service means after the product's delivery, the enterprise's treatment situation to the information consultation about the service's quality, quantity, the information feedback and appeal. Enterprise credibility is the degree of an enterprise realizing its commitment to the customer. Enterprise's value concept is the customer's feeling to the directly contacted employee's business level, accomplishment and spiritual appearance [6-8].

3 Method and Steps of Artificial Neural Network in Evaluation the CS of Enterprise

3.1 Principle of Artificial Neural Network

The BP neural network can imitate the human brain's way of thinking to identify and evaluate the complicated system, and it has the characteristic of self-adapt, self-organize, self-study and the good ability to permit error [9-10]. The subjective factors of custom satisfactory evaluation in enterprise as the import value in the neural network, and the degree of satisfaction as the export value; then train this network with enough specimen, and make different amount of import get different amount of export; so the value that the neural network holds are exactly the output of the inner part, which come from the orientation study; once the training of the neural network is completed, it can immediately be taken as the value of the same kind of enterprise's custom satisfactory evaluation; so the comprehensive evaluations of different enterprises can be made through this model.

3.2 Building the Model of the Enterprise's Custom Satisfactory Based on BP Neural Network

According to Table 1 the BP neural network of satisfactory evaluation in enterprise is established as figure 1 shows

1) Input layer. The specific indexes such as the whole index system are regarded as input neuron for the input layer, According to the factor of custom satisfactory evaluation in the enterprise, the number of neuron for the input layer is eight.

2) Hidden layer. Selection of the neuron in the hidden layer affects precise calculation and learning efficiency for the whole BP network, up till now, there is still no unified ways to identify the number of the hidden layer neuron, which is at stage of research and exploration. If the number of the hidden layer neuron is chosen too little, it will reduce fault-tolerant property and

Table 1 Factor of Custom Satisfactory Evaluation in the Enterprise

The factor of custom satisfactory evaluation in enterprise	Employee in the enterprise	The whole index system
		The work index
		The level of encouragement
		Education and training
		The employee relationship
	Customer	The quality of the product
		The quality of the service
		The image of the enterprise

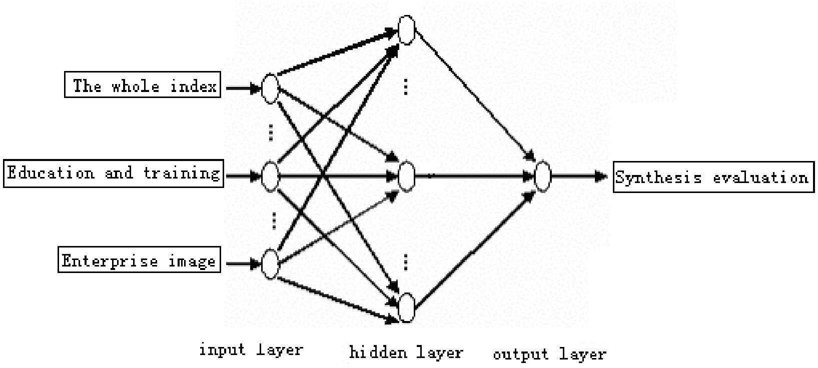


Figure 1 BP Neural Network of Satisfactory Evaluation in High School

self-adaptability of BP neural network, it becomes difficulty for network to deal with more complicated problem, and it leads to the train can′t get ideal result. But if the number of the hidden layer neuron is chosen too more, the time for network training will be greatly increased, and too many process units easily lead to network has too much information processing ability, some information, which has no meanings in training sample data, is remembered, at this time, the true model in data is difficult to be distinguished by network, sometimes, it even doesn't′t converge. Generally, the number of the hidden layer neuron is identified on the basis of the following formula, $P = \sqrt{n+q} + a$, in the formula: n represents the number of the input layer neuron, q represents the number of the output layer neuron, a is a constant between zero and ten. In order to make the number of the hidden node be more appropriate, different numbers of the hidden layer neuron can be chosen to test separately, then, write down the mean-square error between actual output and expected output and the training step of network under every condition, the error and the training step are considered comprehensively, thus the more appropriate number of the hidden layer neuron is fixed. In the paper, the number is fixed for twelve through testing and comparing[11].

3) Output layer. The output layer has only one knot point, namely the synthetic satisfactory of the enterprise automat.

So, BP neural network for custom satisfactory evaluation in enterprise is an 8-12-1 model.

3.3 Calculating Way and Steps of BP Model in Enterprises Customer Satisfactory Evaluation

The eight standardized index values as input signal are passed from input layer to output layer through hidden layer, an output signal is gotten at outlet end (namely actual output value), and this signal is passed forward. If the difference between actual output value and expected output value is bigger than a certain value, the error signal begins to be passed backward, it means error signal revises conjunction weight layer by layer from output layer to input layer through hidden layer, work signal forward passed and error signal backward passed are alternated, learning process doesn't′t end until error signal is less than the certain value. For the question studied by the paper, we use Supervises Learning; the specific learning process is as point:

(1) Variable and parameter are set up.

$X=(x_1,\ x_2,\ ...,x_{21})$,it represents input vector of network, or called training sample;

$W(n)=(w_{ij}(n))$ (i=1,2,…,21;j=1,2,…,12), it represents weight vector between input layer and hidden layer for nth iteration;

$V(n)=(v_j(n))$ (j=1,2,…,12), it represents weight vector between hidden layer and output layer for the nth iteration;

$Z(n)(n$=1,2,…,$N)$, it represents actual output value of network for the nth iteration; d is expected output value; (θ_j) (j=1,2,…,12), it represents output threshold value for every neuron in hidden layer; r is threshold value for neuron in output layer; n is iteration times.

(2) Conjunction weight and threshold value are initialized. Every conjunction weight $(w_{ij}(1))$, $(v_{jt}(1))$ and threshold value (θ_1),$r(1)$ are given a random value in the interval (-1,1).

(3) Compute the implicit layer and output layer's exportation signal.

Sigmoid function $f(x)=1/(1+exp(-x))$ is used as excitation function in the paper. The output value of the jth neuron in hidden layer for the nth iteration is $y_j(n)=f(\sum w_{ij}(n)x_i+ \theta_j(n))$, so, the output vector of the hidden layer for the nth iteration is $Y(n)=(y_1(n), y_2(n) ,\ldots)$; the actual output value of output for the nth iteration is $Z(n)=f(\sum v_{jt}(n)y_j(n)+r)$.

(4) Calculate the margin value d_k between the expected exportation value y_k and exportation value y'_k.

(5) According to adjusted-error margin's anti-spread, you can check every conjunction value and threshold layer by layer.

Conjunction weight $v_j(n+1)$ from the hidden layer to the output layer equals $v_j(n)$ plus $\eta e(n)y_j(n)$, and $e(n)=(d-Z(n))Z(n)(1-Z(n))$; η is learning efficiency.

Threshold value $r(n+1)$ for neuron in the output layer equals $r(n)$ minus $\eta e(n)$.

Conjunction weight $w_{ij}(n+1)$ from the input layer to the hidden layer equals $w_{ij}(n)$ plus $\eta' e'(n)x_i(n)$, and $e'_j(n)=e(n)v_j(n)y_j(n)(1-y_j(n))$; η' is learning efficiency.

Threshold value $\theta_j(n+1)$ for neuron in the hidden layer equals θ_j(n) minus $\eta' e'_j(n)$.

(6) Afresh select a group of training date and return to the step 2, until the margin value BP network's overall square amount is smaller than the pre-established limitative value ε or the number of training times is larger than pre-established number. At this moment, the whole BP neural network's training is over.

Finally, various index specific value that target the enterprise's satisfactory evaluation can be used as the importation of the BP neural network model which has been well trained, and get the synthetic customer satisfactory of the target enterprise.

Through training learning, evaluation network can output appraisal value β, the range of this value is [0,1]. In order to make clear enterprise custom satisfactory level, supposing the state of custom satisfactory are divided into four grades: first grade is higher custom satisfactory, the range of valve is (0.90,1.00]; second grade is high custom satisfactory, the range of value is (0.75,0.90];Third grade is average custom satisfactory, the range of value is (0.60,0.75]; forth grade is low custom satisfactory, the range of value is (0,0.60]. The enterprise custom satisfactory can definitely be received from the value of network output in this way.

During every evaluation process, whether experts approve evaluation result or not, can consider this result the new learning sample, impel the evaluation system of this BPNN to constantly study, continually perfect, drive it to make more accurate evaluation.

4 Authentic Proof Researches

Made use of the model and calculation, based on BP neural network, together with the realization of Matlab software's process, it can be used to evaluate custom satisfactory of the enterprise which belongs to the same organization. Referring to the evaluation index provided by Table 1, seven enterprises are selected, and the related statistic (importation value) and evaluation grades (exportation value) use as training network, and then a well trained neural network is gotten. Finally, the index date of each enterprise in 2004 is input into the well trained BP network to evaluate and exam the custom satisfactory. According to the calculation of Matlab 6.0, the result is gotten, showed in Table 2, and then the comparison of the evaluated result of BP neural network and the result of Delphi method is seen.

The calculated result is tent to be equal to graded result of the experts. What's more, it can be found that the custom satisfactory of enterprise 3 is the best one in these seven enterprises, the custom satisfactory of enterprise 5 is the worst one. So, for enterprise 5, which need take emergency measures to improve its custom satisfactory, or, it will lose market, lose employee, even will be faced bankruptcy.

From Table 2, we also know: the network output values of enterprise 3 is between 0.90 and 1.00, the network output values of enterprise 1, enterprise 6 and enterprise 7 are between 0.75 and 0.90, the network output values of enterprise 2 and enterprise 4 are between 0.60 and 0.75, the network output values of enterprise 5 is between 0 and 0.60, their custom satisfactory levels are showed in Table 3

Table 2 Result of Satisfactory Evaluation

index	enterprise1	enterprise2	enterprise3	enterprise4	enterprise5	enterprise6	enterprise7
The whole index	0.81	0.69	0.93	0.63	0.57	0.82	0.75
Work index	0.73	0.58	0.84	0.65	0.64	0.80	0.84
Encouragement level	0.68	0.79	0.99	0.71	0.74	0.84	0.88
Education and training	0.90	0.75	0.81	0.73	0.72	0.52	0.86
Employee relationship	0.84	0.73	0.86	0.61	0.59	0.80	0.77
Product quality	0.83	0.72	0.95	0.66	0.62	0.85	0.78
Service quality	0.75	0.68	0.92	0.64	0.58	0.79	0.76
Enterprise image	0.86	0.69	0.95	0.68	0.62	0.80	0.80
Synthesis evaluation	0.825	0.705	0.940	0.634	0.583	0.821	0.752
Delphi method	0.819	0.712	0.932	0.641	0.580	0.841	0.762

Table 3 Customs Satisfactory Level of the Eight Enterprises

	Enterprise 1	Enterprise 2	Enterprise 3	Enterprise 4	Enterprise 5	Enterprise 6	Enterprise 7
The value of network output	0.825	0.705	0.940	0.634	0.583	0.821	0.752
The level of sustainable development	Second stage	Third stage	First stage	Third stage	Forth stage	Second stage	Second stage
	High custom satisfactory	Average custom satisfactory	Higher custom satisfactory	Average custom satisfactory	Low custom satisfactory	High custom satisfactory	High custom satisfactory

5 Conclusions

BPNN can realize arbitrary nonlinear reflections between the input and output, which makes it be widely, applied in many areas, such as Function Approach, Model Reorganization, Data Squishing and so on. This paper uses it to evaluate the level of custom satisfactory in enterprises, which is at the base of BP neural network model, is tent to be equal to the result of the Delphi method, the result is quite satisfying. It reveals that the internalized mechanism action between custom satisfactory and its related influencing factors had higher rationality and applicability. The advantage of this method is that it can avoid subjectivity in traditional evaluation methods and that it makes less evaluation time. It is of some scientific value.

References

[1] Zheng X Y, “Index Design for Measurement of Customer Satisfactory Degree of Enterprises and Its Evaluation Method,” J of Tianjin Polytechnic University, 2004.6

[2] Xi J, “Measurement of Customer Satisfactory of the Coal Enterprise,” J of Coal Economic Research, 2004.7

[3] Zhao Q C. Wang B. Liang Z, “Application and Amendment of BP Neural Network Model to Enterprise Integrated Performance Evaluation,” Industrial Engineering Journal, 2005.5

[4] Bart Kosko. Neural Networks and Fuzzy Systems. Prentice hall, 1992(10), pp.37-95

[5] Wu J T, Wang J H. The Application of Neural Network Technology, HarBin Engineering University Press, 1998

[6] XU J, ”Evaluation of sustainable development of coal enterprises based on BP neural network”, Journal of Southwest Jiaotou University, 2005,40(3),pp.375-378

[7] Wang Z C, Evaluation and Measurement of Six Sigma, RenMin University Press, 2003

[8] Li Y ZH, Business Decision Quantity Method, Economic Science Press,2003

[9] Han L Q, Artificial Neural Network Theory, Design and Application, Chemical Industry Press, 2001

[10] Li X CH. Sun Y. Tao X Y, “Application of neural network in evaluating for mining areas”, J of China University of Mining and Technology, 2001 (4), pp.393-395

[11] XU J, Business Competitiveness Improvement Based on Knowledge Generation,Euro-Asia Symposium on Environment and Corporate Social Responsibility (CSR)-From Rhetoric to Practice, 2006.11

Workflow Management Architecture Based on Data Mining*

Lingdong Kong[1, 2] Hong Zhang[3] Lifang Kong[3]

1 School of Environment and Spatial Informatics, China University of Mining and Technology
Xuzhou, Jiangsu 221008, China

2 Department of Software Engineering, Yancheng Institute of Technology
Yancheng, Jiangsu 224003, China

3 School of Computers Science and Technology, China University of Mining and Technology
Xuzhou, Jiangsu 221008, China
Email: njkld@163.com

Abstract

This paper probed into workflow technology applied in Business Process Reengineering and analyzed workflow model, workflow management process and workflow implementing process. Process definition and tool of management monitoring were extended locally in workflow model combining with data mining technology. Static data mining was added in the front-end of workflow process definition to extract rule database; dynamic data mining was added in the back-end of workflow management and monitoring tool for record database to bring strategy data base. Workflow management architecture will be put forward based on data mining. Further key technology has been put forward to actualize that static and dynamic data mining applied in the architecture combining with coal and gas outburst prediction of safety workflow management system.

Keywords: Business Process Reengineering, Workflow, Static Data Mining, Dynamic Data Mining

1 Introduction

Workflow technology originated in the 1970s has become an important enabling technology which may mine business process potency and carry out process reengineering [1]. The most direct use of workflow management system was to manage enterprise business flow as the enterprise-level applications system, which it combined the Business Process Reengineering technology to realize enterprise flow automatization [2]. We may implement process reengineering and transition while propelled workflow instance execution through definition and management of process, and then integrated intrinsic information system and association roles [3]. Data mining is emphasized by discovering and building model to predict the future through reasoning, and extracted the hidden and useful knowledge rules.

Potential knowledge rules can be discovered through deeply reasoning and analyzing business process model integrated workflow and data mining by making the most of enterprise associated data, cases and running record of workflow management system. By finding unreasonable parts of business process we may not only standardized and further optimize and reengineer but also took the building process model as very important rule base and strategy base of enterprise, and run scheduling to guide design and execution of process management automation system. Some studies have been done on workflow in electronic commerce (EC), advanced manufacturing, but they are mostly applied in particular areas. At the same time, data mining have been widely applied in venture prediction, business trade and other fields, integrating workflow and data mining technologies won more and more

* Graduate Innovation Program of Jiangsu Province of China under Grant No.CX07B_139Z ; the Natural Science Foundation of Jiangsu Province of China under Grant No.BK2005021.

attention now.

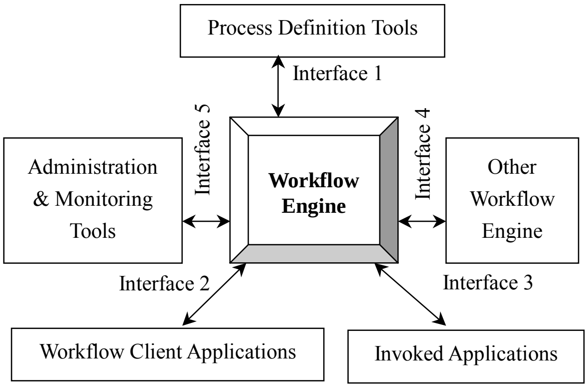


Figure 1 WFMC Reference Model

2 Workflow Applied in Business Process Reengineering

According to Workflow Management Coalition (WFMC) definition, Workflow Management System (WFMS) is a software system, which completes the workflow definition and management, uses software definition to create workflow and manage execution in accordance to preliminary definition workflow logic in the computer. It runs in one or more workflow engines which interpret process propelling and workflow instance execution interact with workflow participant (including person or software), invoke the other IT tools and application according to its demand [2]. It was maintained that WFMS as Business Operating System (BOS), it was finely fulfilled to Enterprise business process operation by integrating concrete application software and operator interface.

2.1 Workflow reference model

Workflow System extracted process management from application software. Through the implementation of workflow management system, we can realize flexibility, system integration, process optimization, organization alteration, enhancing maintainability, iterative development, etc. WFMC gave a general description of the Workflow architecture [2]. Figure 1 illustrates WFMC Reference Model.

The above model shows that the core of Workflow system is the workflow Enactment services by the workflow engine; it promotes the transference of cases in the organization, Enactment services ensure the right people to carry out the right activities according to the correct order. Process definition often facilitates to analysis besides producing process definition and resources classification. This paper combined data mining technology on the basis of the process definition, further extended and realized the extraction of rules and knowledge base.

On the one hand, management and monitoring tools are responsible for the operation and management of workflow. On the other hand they are responsible for record and report of workflow. A lot of information can be recorded and stored during workflow implementation. They are very useful historical data for the management. Related to this are the main performance indicators, the average completion time of cases, the average waiting time and processing time, the percentage of cases completed in a certain period of time, the average level of resource utilization. The recorded data can be stored in database management system. This paper based on the extended recording and reporting tool, dynamic data mining was realized and workflow management strategy base was also built.

2.2 Relationship between workflow and business process reengineering

We can see that the workflow uses different resources from the workflow concept and structure of the reference model; it would facilitate the realization of business process automation through the process definition to manage the case running through the graphical tools of management and monitoring. Business Process Reengineering is based on the Business Process Model. Description of workflow processes for high-level facilitations to reason. The reengineering mapped with original information system. The basic strategy is that these tasks were implemented and monitored in accordance with certain rules and processes through dividing a work into tasks and roles of good definition to improve efficiency, lower production costs, enhance production management level of enterprises and their competitiveness. Using workflow to achieve BPR has become the consensus of many scholars [1].

Workflow management that supports the business process and information reengineering technology includes the following three aspects which was showed in the workflow management process in Figure 2.

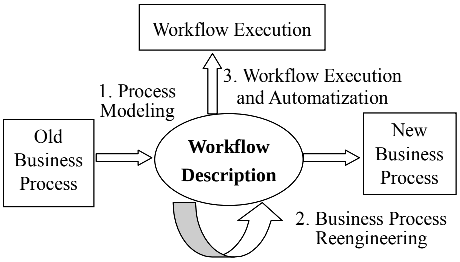


Figure 2 Workflow Management Process

(1) Business process modeling and Workflow description: it is required to describe the workflow using workflow model and method, and access to business processes. Description of the workflow is the abstract of the process, the level of which depends on the purpose of Workflow description.

(2) Business process Reengineering: the need for the method of optimization flow. Process optimization strategy depends on the objective of the reengineering; high-level description of workflow facilities reengineering.

(3) Workflow implementation and automatization: method and technology were demanded to execute and control the description tasks in the workflow definition

It can be seen that business process modeling and workflow description is the core in the implementation of the business process; workflow management process involves what tools to be used directly related to the modeling, what the basis of modeling is as well as how to realize the function and solve problems.

To further clarify the relationship between modeling and business reengineering, as showed in Figure 3 workflow management system implementing process further analyzes the implementing process of a workflow management system; the process reengineering is closely related to the process transformation in this process. It is a constant recycling and improving process according to different changes.

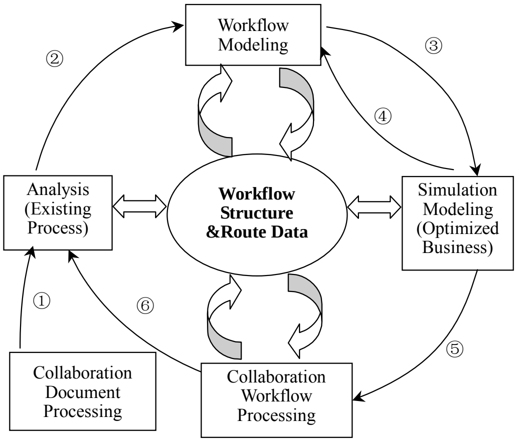

Workflow Modeling
②
③
④
Analysis (Existing Process)
Workflow Structure &Route Data
Simulation Modeling (Optimized Business)
①
⑥
⑤
Collaboration Document Processing
Collaboration Workflow Processing

Figure 3 Workflow management system implementing process

To further clarify the relationship between modeling and business reengineering, as showed in Figure 3 workflow management system implementing process further analyzes the implementing process of a workflow management system; the process reengineering is closely related to the process transformation in this process. It is a constant recycling and improving process according to different changes. ①the model of the existing business process has been gotten through analyzing and summarizing the existing business processes, business rules, regulations and other management;②BPR tool has been used to analyze the existing business process model, to abstract and collate for the existing processes, and therefore the current systematic workflow model has been realized.;③, ④ simulation has been executed for workflow model, questions. Improvement has been done on the model that has been optimized business process workflow model combing enterprise reality; ⑤operating results were received through putting this model into actual operation; ⑥the problems have been revealed to further improve workflow model through analysis of the results of the operation.

Description and modeling of workflow is the core to the workflow implementation through the above workflow process and management system implementation for both the graphic analysis. Business process reengineering means rethinking and reengineering business processes. The objective of BPR is to design a "better" system of workflow，optimization strategy depends on the objective of the reengineering, for example, reducing business costs and providing new products and new services. Process reengineering will reach an initial expectations based on process definition modeling in a relatively stable environment. However, the changes in the course of business with the resources needed to be adjusted at any moment. So we need to be ready to face the issue again in the redesign or optimize business processes. If it is a simple reengineering，it is a huge waste of human and material resources , the original data may be discarded.

In data mining associated rules can be found on the basis of the large amount of data while operating the workflow process involving a lot of original resources and operational data. In a complex and changing environment process, the rule database (DB) can be defined as the guidance of the process definition and process implementation. In the process of implementation dynamic mining can be achieved based on workflow records, and then realized the dynamic management. Thus, we propose workflow management architecture based on data mining.

3 Workflow Management Architecture Based on Data Mining

3.1 Data mining and workflow management

Data acquisition and storage technology advances led huge databases to increase day by day with the business of enterprises constantly updated. The corresponding flow data will be huge with the implementation of workflow case in the workflow management system. Whether extracting valuable information from these data or not is expectable. Data mining is data analysis of the observation (often very large), the aim is to discover the unknown relationship and the data owners can understand and take the novel way to sum up their valuable data [4]. Data mining, also known as "knowledge discovery in databases", is the process of discovering interesting patterns in databases that are useful in decision making. Data mining is discipline of growing interest and importance, and an application area that can provide significant competitive advantage to an organization by exploiting the potential of large data ware houses [5].

Most workflow management systems can record logs in the activities that occur in the executed workflows [6], and even non-workflow-specific computer-based systems can log generally activities that users perform. Thus within and without workflow systems, there is often a rich source of data that can be mined to learn something about the work processes.

Reference model analysis of the workflow and further detailed study show that, on one hand, we have to consider various aspects of the resource data before defining the workflow process what can be anticipated are organizational resources, business organization, management, information resources, which will be a huge database. On the other hand, the workflow in the process of implementation, the record information related to, and related data resources and process characteristics log from the beginning to the end are all destined to be a huge source of data with the passage of time. Obviously, the foundation for a data mining was provided in the process of implementation of our workflow. At the same time, businesses in the reorganization process, we hope to mine unknown and useful strategies on the basis of data in the past. The goal is to capture the structure of the sequence and find disciplinarian which includes execution of exceptions model. The implementation process can be further optimized to achieve the process reengineering. Data Mining is an obvious necessity.

We start from existing process definition and analyze process logs in order to improve and reengineer the process model. To the best of our knowledge, there are no other approaches to process execution analysis and prediction based on data warehousing and data mining techniques for Business Process Intelligence [7]. Analysis of exceptions may help the operator to obtain a possibility of exception and automatically remind staff of doing correctly and understanding what caused these exceptional circumstances may be caused and prevent the occurrence of exceptions by automatic or semi-automatic.

In short, it is necessary to extend workflow model based on the extension of workflow records and process definition to realize process optimization using data mining technology.

3.2 The Extended Workflow Reference Model

To realize the dynamic model management and the reasoning of exception analysis, so as to realize the business process reengineering in the workflow management, workflow reference model was extended on the basis of research of workflow reference model and data mining as following.

In Figure 4 Workflow management architecture based on data mining shows. Static data mining was added for resource database in the front-end of workflow process definition, and on the basis of which foundation rule database can be extracted to guide the workflow process definition; at the same time, dynamic data mining was added in the back-end of workflow management and monitoring tool for record database, and on the basis of strategy data base of workflow process can be extracted for the reengineering of workflow process definition.

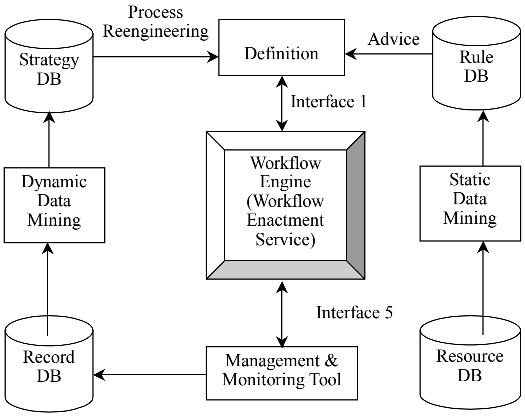


Figure 4 Workflow management architecture based on data mining

In our study, we combined mine gas outburst security prediction issues to probe into key technology implementation of workflow architecture base on data mining. Mainly consider the complex security management environment, frequent data changes and the dynamic monitoring of the workflow characteristics. This architecture can work for other management systems, especially large data information, on state workflow management system with much change. For example, the flow of work for ordinary supermarket business, with the growth over time and the continued operation massive infrastructure information can be used to express the information. At the same time, the cargo loading, inventory control of the flow are very dynamic, we can implement the Dynamic Data Mining-based scheduling.

1) Static Data Mining

The static data mining in this paper is based on the workflow process definition and Enactment, the initial implementation of a data mining can be carried out based on various static database resources analysis of the enterprise.

The enterprise business is a complex process where there are many uncertain factors, it is almost impossible to find an analytic algorithm for the description and analysis of their relative precision. Under such circumstances, we have implemented two measures. First, a variety of integrated data mining algorithms are used in resources database. For example, business intelligence integrated association rules, neural networks, and decision trees of 10 types of data mining algorithms in Microsoft's SQLServer2005. Comparative data mining algorithms for a variety of implementation is economically feasible; we have done certain amount of work in developing gas safety prediction systems. Second, the simulation analysis is a feasible and effective method for workflow model before the definition process. From the perspective of security and the perspective of leading technologies, these two measures could more effectively use of static data mining process to be used to guide the definition of the rule database.

Combining with our studies of coal and gas prediction model, the influence of Methane outburst can be shown by gas pressure and gas quantity. Gas pressure is the necessary condition and the gas quantity is the favorable condition, Gas formation and deposition and preservation of the environment, coal-bearing strata of rock composition and coal, coal rank. Regional tectonics and rock conditions are closely related to the closure. In addition, other factors such as geology and hydrology, erosion of the coal-bearing strata, to a certain extent, affect the content of coal seam gas. Therefore, the forecast for gas quantity should take these factors into consideration [8]. After years of a mine or mine detection and collection of data relevant to the plot based on the direct data gathered will be enormous resources, Meanwhile, we must consider other organizations, businesses in the implementation process of workflow management, initial data is even more huge.

2) Dynamic Data Mining

Business Process Management System is a dynamic workflow, hidden knowledge rules are found by data mining and optimization model and forecast trends may be built through reasoning. In order to enhance the ability of solving problems and take appropriate actions problems should be predicted as early as possible. We put forward the Dynamic Data Mining which was considered to be used in the workflow implementing process. The key problems can be dynamically reasoned and found according to relevant records and different phenomena and results in exception condition to adjust the implement workflow, and further improve the implementation strategies. Drawing on the fuzzy inference technology and the Petri-Net technology closely related to workflow in the implementation of the development, the reasoning mechanism Petri-Net introduced the fuzzy [9][10][11].

In our project, we have to consider at least the following two processes in the workflow management system of Coal mine gas safe prediction. First, the business flow: bed methane extraction process, the process of excavation, tunnel ventilation, etc.; Second, the ground work management process: office of gas outburst prediction -- officials -- Alarm Center -- monitor-- operator. More detailed and professional in the process we do not further introduce here. We only explained that there are a large number of cases recorded in the implementation process. We may implement to manage workflow log，further more, it comes into cooperating with dynamic data mining and static data mining. As showed in Figure 5 Key technology to actualize in the workflow management architecture, we firstly built business modeling for business flow to sent workflow management, and acquired workflow log in the business implement. Furthermore, we started to carry on dynamic data mining through knowledge reasoning based on Fuzzy-Petri net and gained strategy base to send resource database, and while we may use reasoning results to build new business modeling.

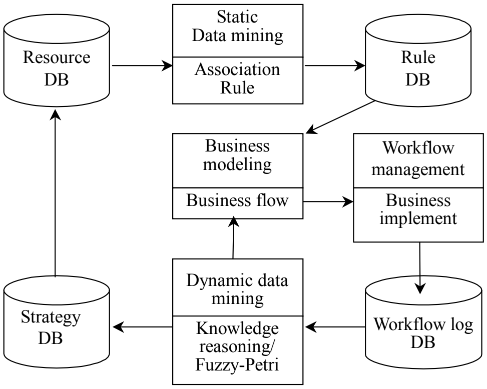


Figure 5　Key technology to actualize in the workflow Management architecture

4　Conclusions

Workflow management system based on Data Mining possesses a natural huge database basis for the implementation of data mining in the process of implementation. The implementation of data mining is necessary in the implementation of business process reengineering and strategy optimization process. In the study of the forecasting of Coal mine Safety prominent data mining process flow, we take into account the flow and standardization of safety prediction and the complexities of the frequent changes in the state during researching workflow data mining in the safety prediction of coal mining gas outburst. Process definition and monitoring management tools in the workflow model were extended based on data mining technology combining workflow with Petri-net technology. The concept of static and dynamic data mining was put forward in the implementation of key technology and workflow management architecture based on data mining was probed into.

Combining the architecture with implementation of key technologies, we will make in-depth research in the following aspects ① filtration and optimization of resources database based on static data mining, this is a prerequisite for the implementation of Data Mining; ② selection of Workflow key recorded and knowledge representation of Petri-Net; ③Petri-net fuzzy reasoning mechanism in the implementation of dynamic mining. ④simulation and prediction of complex environment in the workflow implement.

References

[1] Yushun Fan, Fundamentals of Workflow Management Technology, Beijing: Tsinghua University Press, 2001

[2] WFMC, the Workflow Reference Model, TC00-1003, 1995

[3] Wil van der Aalst & Kees van Hee, Workflow Management Models, Methods and Systems, Beijing: Tsinghua University Press, 2004

[4] David Hand, Heikki Mannila, Padhraic Smyth, Principles of Data Mining, Beijing: China Machine Press, 2003

[5] Indranil Bose, Radha K.Mahapatra, “Business data mining-a machine learning perspective”, Informaition & Management, Vol.39, 2001,pp. 211~223

[6] Jonathan E. Cooka, Zhidian Dua et al, Discovering models of behavior for concurrent workflows.Computers in Industry, Vol.53, 2004, pp. 297~319, 2004

[7] Daniela Grigoria,, Fabio Casatib et al, Business Process Intelligence. Computers in Industry, Vol.53, 2004, pp. 321~343, 2004

[8] Ministry of Coal, Rules of Preventing Coal and Gas Outburst, Beijing: China Coal Industry publishing House, 1995

[9] Amit Konar, Uday K. Chakraborty, Reasoning and unsupervised learning in a fuzzy cognitive map, Information Sciences, Vol.210,2005,pp.419~441

[10] Shlomit S. Pinter, Mati Golani, Discovering workflow models from activities’ lifespans, Computers in Industry, Vol.53, 2004, pp. 283~296

[11] Andras Javor, Petri nets and AI in modeling and simulation, Mathematics and Computers in Simulation, Vol.39, 1995, pp. 477~484

Key Problems of the Data Mining Application in Fiber Network and Implement*

Funing Yang Yongjian Yang Zhizhong Tian

College of Software, Jilin University, Changchun, Jilin, 130012, China
Email: yyj@jlu.edu.cn

Abstract

The telecom industry is a typical data denseness industry. Data Mining technology could offer decision making support for the telecom merchants' next scheme. This paper is about the key problems of the data mining application in the fiber network and the methods of achievement. We will explain it through two examples, the one is to use the algorithm of related regulation excavation to analyze the telecom alarm data. In practical work, it will help the telecom administration department to make prognostic alarm and deal with the obstacles rapidly, the another one is to use the most common used Dijkstra algorithm in shortcut algorithm about the single source point to analyze the problem of the random two nodes' shortest path in fiber network, and draw the flow chart of the algorithm's implement.

Keywords: Data Mining; related regulation algorithm; Dijkstra algorithm; Fiber Network

1 Introduction

The telecom industry is a typical data denseness industry, As the telecommunications reform, and telecommunications companies split and the reorganization, the telecommunications industry’s competition is becoming increasingly fierce. Compared with other industries, the telecommunications industry has more of the user's data. Who can correct analysis of these data, who will be able to provide better services to users, and can find more opportunities to win in the competition. Telecommunications enterprises must retain user data call to mind the fees, surveillance operation of the network status and do a good job network planning, telecom enterprises have to analyse such data to find useful to the law of network optimization. Therefore, data mining in the telecommunications industry has an important value[1].

In the telecommunications industry, they can make use of data mining technology to customers action, customer trust, personalized service, call data to in-depth analysis, to develop decision-making enterprises operators to provide effective support[2].

Below use two algorithms: Mining Association Rules Mining and the shortest path algorithm, described fiber-optic network to the application of data mining technology key issues and solutions.

2 Mining Association Rules

Mining Association Rules (Formatting Instructions) is an important aspect of data mining study [3], Mining association rules can be described as follows: in the affairs of a given database, related function found in the role of database services, the return of the close relationship between the sets. Data Mining researchers in the field of mining association rules done a lot of work, They mainly include: classic prior algorithm and its expansion, generalized association rules, quantitative association rules mining algorithm, as well as incremental Mining Association Rules algorithm[4].

Daily in the telecommunications network will have a large number of warning data, these data hide a lot of

* This work was supported by Research Project for undergraduate students of Jilin University.

valuable information. This information can be used to filter unnecessary alarm, can be used to carry out in the network fault location can be used to predict serious mistake.

However, this information is hidden in the data, data mining can be obtained through [5]. In the alarm data can be excavated all the different types of knowledge, such as neural networks, pattern and risk of discovery rules.

We have chosen a rules-based forms of mining. Generally in the form of the following: "If over a period of time in a certain order, there had been some type of alarm, then another over a period of time will be at a specific order of another particular type of alarm." Choice of this type of knowledge excavation following reasons: (1) understandable: such a rule easy to understand the handling of the alarm sequence operators happy with this and similar rules in the form of knowledge. (2) operable: Such rules can be expressed in the area of this simple small probability. (3) efficiency: the existing algorithm can efficiently excavated above rules.

Description Rules

Formal Description association rules [6]:

Let H = (h1, h2, ⋯, hm) is a collection of projects, the G is set for H affairs. Each branch of the G T by Tid logo, T includes a number of projects hi1, hi2,⋯, hik ∈ H, if the project-XT, X claimed Affairs T support. Association rules is contained in a below -X => Y, which XH, YH, and X ∩ Y = Φ. If in the G have the %s in the affairs of support -X project, said the project has the set X of a support for the size s, remember to sup (X) = s. If sup (X ∪ Y) = s, said the rules X => Y in the transaction database has the size of s for the G in a degree of support. If X support of G in the affairs of the %c Services also support the Y, said the rules X => Y in the G is the size of c confidence, recorded as conf (X => Y) = %c. Meet certain conditions of association rules can be meaningful, needs to designate two thresholds: Minimal Support (MinSup) and the Minimal Confidence (MinConf). If sup (X) ⩾ MinSup, called X is frequent item sets; if conf (X => Y) ⩾ MinConf, then X => Y set up; at the same time meet the threshold of two rules-called rules.

In a Warning in the Database Description Mining Association Rules: The first step of Data Mining is collecting and cleaning data, we assume that the information has been processed, has the higher the quality. Each alarm can be abstracted as a member (com, p, t), the com is a warning components, p is the type of alarm, t is the time for a warning.

We related to the definition of the concept. Alarm predicate is a expression that explain the warning alarm type, or other parameters, For example: "the type of Warning Z is B", "the risk coefficient of warning Z is 2 (can say emergency alarm and there could also have 1: general warning 3: serious warning, etc.). We define an event α = (V, ⩽), which V is a set of predicate warning, "⩽" on the partial sequence V, in the order specified in the warning. Figure 1 below is the icon of the incident of example 1, in the icon ,the node is warning predicate, they expressed alarm type B (or A, C, D, E), V is a collection of nodes, the icon of the four arrows indicate the partial Sequence "⩽."

Example 1: If in 10 seconds, the alarm of type B in the back of the alarm of type A occurred, the alarm of type D in the back of the alarm of type C occurred, then in the next three minutes, the alarm of type E occured by the probability of 0.8.

In classic association rules, the number of projects set for the project definition of the Panel, set the number in the database, warning in the database, we need to define an event in the alarm database (ie alarm sequence S), the frequency of occurrence. because of the member of alarm database have the time attributes, we can be seen alarm data as warning sequence S, First definition the sequence Si(i = 1, 2,⋯) of warning sequence S， given time window width W, the alarm sequence occur in the w which of the sequence Si of the S. In other words, if at the T0 time the first alarm of sequence S occured, then Si contained in the [T0 + iu, T0 + iu + W]occurred within the time interval of warning, u is a window migration. If Si in the incident α meet some alarm in the predicate set V, these warning and the α in order to meet the order of "⩽" says α occurred in the Si, Si can also support α.

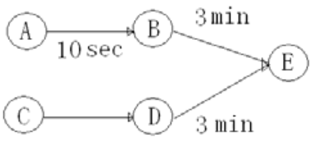


Figure 1 the icon of the case of example 1

Such as Si: X1 (C1, A, t1)，X2 (C2, B, t2)，X3 (C3, C, t3)，X4 (C4, D, t4)，they occurred in the order X1 → X2 → X3, the incident α = (V, ≤) occurred in the Si,

V: (X1 type A, type B X2, X3 type C), "≤": X1 in the X2 after, X3 in the X2 after. The number of Si of α in support of S is the number of α occurred in S. If the frequency of α greater than MinSup, then the α is the frequent incident.

In the alarm database, association rules abstract for R = (α, e, W, W′), α = (V, ≤) is incident, e is a predicate in the V, W, W′ is the width of the window, these variables Value is defined by the user. Rules similar to X=>Y, but the left of the rules is the incident (V′, ≤), V′= V-(e), "≤" is a user-specified order, the right rules is e. Interpretation of the rules are as follows: If the incident time W (V′, ≤) occurred, in the W′ time to meet the predicate e alarm in the "≤" in a location specified.

Than Minimum confidence’s rules make sense, the confidence of the rules R is t’/ t. t is the number of the incident (V ′, ≤) occurred in the alarm sequence S under the conditions of the window of width W. t’ is the number of the incident α= (V, ≤) occurred in the alarm sequence S under the conditions of the window of width W’.

Alarm predicate including typical types and risk factors may also include a warning components. Time window in the general choice between five seconds to half an hour.

alarm data from a rule extraction algorithms

Mining Association Rules algorithm decomposed into two sub-problems: (1) Services database obtained meet the minimum support the frequent item sets. (2) the use of frequent-generation projects to meet the minimum confidence of all association rules[7]. the solving of the child problem 1 is the key part of the mining association rules, That is, how to efficiently obtained frequent item sets. Apriori algorithm was considered the most classic algorithm, the algorithm on the basis of Apriori also have extended algorithm[8]. The solving of the problem two is relatively simple, Frequent item sets for each L, count all of its non-loopholes sets, each nonempty set a , inspection rules, a => (L-a), If the rule's confidence more than the minimum confidence , then export the rules. In the alarm database count data mining algorithm also follow this thinking, the first count frequent incidents, then count the rules of confidence, output greater than the minimum confidence rules.

count frequent incidents: Given S, E, ε, W, minsup, S is a warning sequence, E is set alarm predicate, ε is drawn from E composed of a number of incidents, is the smallest support minsup, W is the width of the window. count the frequency of all the events in the ε , find support for the events more than minsup. In the algorithm, L[K] is the length of the k-frequent events, C[K] is the length of the k-candidate events.

(1) C1={{e}/e∈E}

(2) i: =1

(3) While Ci≠Φ do

(4) Indentification: the reader sequence S, in each of the Ci to verify an event to see the window’s width W whether to meet the conditions of support than minsup, will meet the conditions of frequent incidents into the incident that set Li.

(5) Counting Ci+1, the events include of Ci+2 have the number of i+1 predicate, and the events of the incident in the window W for support under the conditions of more than minsup.

(6) i: = i＋1

(7) End

(8) For all i，output Li;

This iterative algorithm is the first choice for a length of the events set C1 scan length for a generation of frequent events set into L1. Then Step 5 Apriori algorithms used by the L [K] generate candidate incident Set C [K + 1], in the fourth step of warning data to support S sequence of test generation for the length of the frequent incidents K +1 set.

If C [K +1] is the empty set, then the algorithm end, the output of all the frequent incidents Set Li. Following this the efficiency of the algorithm is essential: If the incident is not a frequent event, then it is not a superset of the frequent incidents. Therefore, the events of candidates Ci include only those sets of events is frequently the case.

Mining rules from the incident: Once we know that in a sequence of events like alarm every incident in the frequency, then count the credibility of the rules has become easier. For a rule r (α, e, W, W′), Count the number t of the incident α-W (e) for the time window of W and the number of t' of the incident α for the time window of W', and they are both beyond the minimum support, then we came to the rule's confidence is t'/t, all the rules larger than the smallest confidence is that we have to mining rules.

3 Dijkstra on the Fiber-Optic Network Routing Management

Feasibility Analysis

The paper title (on the first page) should begin 1.38 inches (35 mm) from the top edge of the page, centered, completely capitalized, and in Times 14-point, boldface type. The authors' name(s) and affiliation(s) appear below the title in capital and lower case letters. Papers with multiple authors and affiliations may require two or more lines for this information.

Fiber network topology with some of the features of graph theory, as direction, right and the right of non-negative. For example, Qingdao Road, People's Square and the District Building in some city, which opened three branch, which is different from the level of, is the hub, aggregation, and the general branch, the two nodes A and B of the laying of fiber optic cable, through a branch line C of the physical path is the shortest in the works will save human and material resources. C, but in this routing nodes on the business rarely, although Qingdao Lu branch journey will take much more, but because of the increase in business to make up for the laying of fiber optic cable in the excess consumption. That is to say, cable operators in the works, not considered absolute path, we must also consider factors such as the level of branch, which is similar to the weighted value of the shortest path problem. We can use the Dijkstra algorithm[9][10], as Dijkstra algorithm is applied to the weighted topology of any two nodes in the shortest path.

Dijkstra Algorithm for the Basic Ideas

Assuming each have a grade point (di, pi), which di is the length of the shortest path from the source point i to s (from the vertex to its own road, the shortest path is not arc), and its length is equal to zero; pi is the first point in the shortest path from s to i.the basic process of solving the algorithm of the shortset path from s to i is as follows:

1) Initialization. Source settings:

① ds = 0, ps is empty;

② all other points: di = ∞, pi =?

③ mark source s, k = s , and all other points set not marks.

2) Test the distance from all its points k which has been marked to the other point j which has not been marked, and set: di = min [di, dk + lki],the lki is the weights directly from point k to i.

3) Select a point. From all the nodes which not marking, select the smallest j from di: di = min[di, all unlabeled point i],point j was selected on the shortest path that has been set and marked.

4) Find the point before the point J. select the point i* which directly linking the point j from the points which has been marked, as the before point which is the point from the soures point s to ti, settings: pi=i*.

5) Marker j. If all the points have been marked, the algorithm fully launched, otherwise, k = j, turn to (2) to continue.

Achieve Dijkstra Algorithm

First description of the entire network through Matrix, making a mutual relationship table from the database, in order to storage the nodes values.

Definition the weights of the adjacent nodes (business between the two nodes which is connected to a line) is 1, the weights of the not adjacent nodes is 100. V said that with a collection of all nodes, V arbitrary points in a, b the shortest path through as a point source, use Dijkstra algorithm counting, if b is that point has been marked by a to b has been seeking the shortest path out, stop counting. Algorithm flowchart in Figure 2:

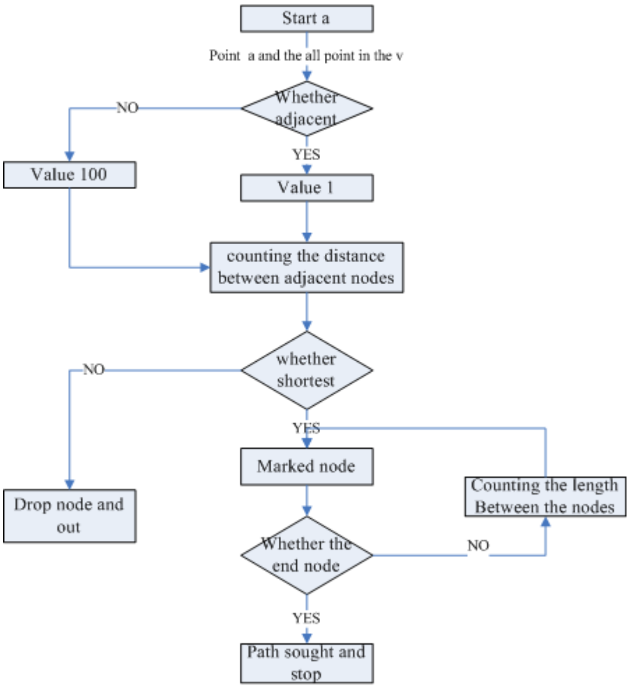


Figure 2 Dijkstra algorithm lowchart

4 Conclusion

Currently association rules mining algorithm and the shortest path algorithm has matured and has been applied to practice. According to the special nature of alarm data, described in the telecommunication network alarm data in the database of mining method and algorithm. Through the application example we illustrate the shortest path algorithm in the telecommunications network construction. Alarm in the telecommunications network in the database to tap the knowledge in the network can be used to carry out fault location, can be used to predict serious mistake, the shortest path telecommunications network operators can help in building a better allocation of resources, both on the management of telecommunications networks, will do great role.

References

[1] Baoming Tan, TMN basic skills, Beijing: Beijing University of Posts and Telecom Press, the first version, 1998.9

[2] Yunfeng Duan，Weining Wu，Jianei Li. Data Warehouse and its Application in the Field of Telecommunication, Beijing: Electronics Industry Press. 2003

[3] Ming Zhu. Data Mining, Beijing: China University of Electronic Science and Technology Press. 2003

[4] Li Liao, Yingze Li. Data Mining and Data Warehouse and its Application in Telecommucation. Chongqing: Journal of Chongqing University of Posts and Telecom. Journal, Vol.27(4), 2000.12, pp389-395

[5] Mingtong Liu, Wei Liu. Research on Space Data Mining Technology and Development Trend. Remote Sensiong. Vol.32(3), 2002.9, pp 257-262

[6] Mingyi Du, Shousong Gu, Dazhi Guo. Data mining Based on Space data warehouse. Computer Engineering and Application. Vol.36(1), 2005.1, pp 25-31

[7] Agrawal R, Imielinski, T, Swami A. Mining Association Rules between Sets of Items in Large Databases. In Proceedings of the 1993 ACM SIGMOD International Conference Management of Data. Washington, DC: ACM Press, 2006, 207-216

[8] Jiawei Han, Micheline Kamber. Data mining Concept and Technology. Beijing: Mechanical Industry Press. 2001.8 the first version

[9] Klemttinen M, Mannila H et al. Finding Interesting Rules from Large Sets of Discovered Association Rules. In Proceedings of the 3rd International Conference on Information and Knowledge Management.Maryland: ACM Press, 1994, 401~407

[10] Shashi Shekhar, SanjayChawla， Kunqing Xie, etc translation. Spatial database[M]. Beijing: Engineering Industry Publishing Company，2004, 167-189

AN Information Filter Based on Reversed Word Segmentation and Complexion Detection

Leiyue Yao[1] Jianying Xiong[2]

1 Software Research Center, Jiangxi Blue Sky University, Nanchang, Jiangxi, China
Email: ylyyly2001@163.com

2 Department of Computer Science and Technology, Jiangxi Blue Sky University, Nanchang, Jiangxi, China
Email: xjy9@21cn.com

Abstract

The kernel of unhealthy network information filtering is to identify the web page properties. To overcome the shortages of traditional algorithm, we proposed a new method based on text analysis and color recognition. A mathematical model was established which combines text and image to judge the properties of web-pages. The new method dramatically improved the recognition rate and reduced the false accept rate.

Keywords: Information filter, Word segmenting, complexion recognition

1 Introduction

Internet is highly anonymous, zighly intimate, highly interactive and non-regional, so it is difficult to effectively manage adult information which may have side effect on teenagers' mental health. In some aspects, adult information is even worse than unhealthy online- game. In this article, we proposed a new information filtering technology. This new algorithm combines keywords training, keywords detecting, skin detecting and can effectively filter unhealthy information. Our experimental results showed that the recognition rate of unhealthy web-page is as high as 98.6%.

2 Text Detecting and Analysis Algorithm

2.1 Construct knowledge base

Knowledge Base has two attributes: keyword and weight. "Keyword" refers to a word that may result in a web-page or an article being classed as unhealthy, such as impolite words, obscenity, alias of sex, etc. "Weight" is a number of the keyword. The higher the number, the more likely it is unhealthy.

"Keyword" and "Weight" can be manually input at the beginning. Both of them can be trained to build an excelsior knowledge base. The training strategy is to add 1 on weight once the keyword is found in an unhealthy web-page or an article. Table 1 shows the detail statistics.

Table 1 Detect Result (The letters in the first column refer to keywords)

Keyword	Web-pages	Unhealthy Web-pages	Contains/ Percentage	Times
AAA	1032	328	348/33.72%	4823
BBB	1032	328	1032/100%	28465
CCC	1032	328	64/6.20%	180
...	...	...	...	...
AAA	328	328	259/78.96%	3808
BBB	328	328	328/100%	8793
CCC	328	328	13/3.96%	34
...	...	...	...	...

Obviously, there are three keywords in the knowledge base. "AAA" occurs in a high rate in unhealthy web-pages, and it is also shows up in common-pages. "BBB" is found in a high rate in both unhealthy web-pages and common pages. "CCC" occurs in a low rate in unhealthy web-pages.

According to the statistics of testing data, a mathematic model can be written below:

$$V_K = \frac{B_r}{A_r} * \frac{T_B / N_B}{T_A / N_A} * k$$

Where, B_r is the rate of a keyword that exits in web-pages and articles which only contains unhealthy text. A_r is the rate of a keyword that exits in web-pages and articles which contains both common and unhealthy text. T_B is the times of a keyword that is found in web-pages and articles which only contains unhealthy text. T_A is the times of a keyword that shows in web-pages and articles which contains both common and unhealthy text. N_B is the total number of unhealthy web-pages or articles. N_A is the total number of both common and unhealthy web-pages or articles. k is coefficient.

According to the above model, set k =10, the “Weight” of keyword “AAA” can be calculated as:

$$V_{AAA} = \frac{B_r}{A_r} * \frac{T_B / N_B}{T_A / N_A} * k$$

$$= \frac{78.96\%}{33.72\%} * \frac{3808/328}{4823/1032} * 10 = 18.48 \approx 18$$

Using the same formula, the knowledge base can be created as:

AAA	18
BBB	7
CCC	4
…… …	

2.2 Text analysis

After obtaining the numbers and times of all keywords that show in the web-page or article, the result can be written as:

$$V_T = \sum_{i=1}^{n} (V_i * t_i)$$

Where, V_T is the test result which will be transferred to procedural module as the judging parameter.

2.3 Keyword matching algorithm

Different from English, there are more difficulties in Chinese phrase matching. Different combinations of Chinese characters always bring different meanings. In this algorithm, we use reversed-word-segmentation method to detect keyword. But traditional method cannot work efficiently, since there are too many Chinese phrases in the knowledge base. Two steps are taken to optimize the old algorithm. First, reduce I/O actions as many as we can. Read the whole Chinese phrases into memory when our application runs firstly. Although it takes a long time (about 1,300 ms) when it starts, it cost little from than on. Secondly, sort all the keywords from small to large, so that binary search can be used to match Chinese phrase in the most efficiently way.

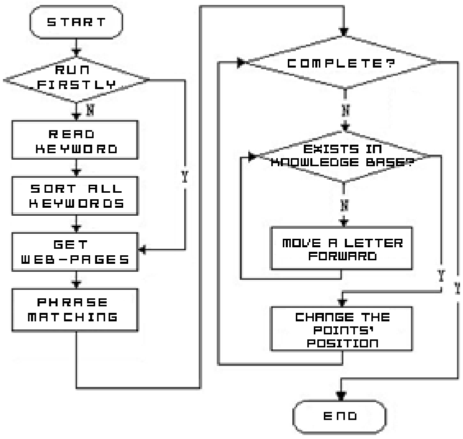


Figure 1 The Algorithm Flow Chat

2.4 Test result

Table 2 shows the test result based on the above Algorithm

Table 2 Test Result

Total web-pages	Unhealthy web-pages	Detected web-pages	Wrongly detected web-pages	Undetected web-pages
1032	328	317	33	11
328	328	317	0	11
634	216	208	18	8
412	100	97	15	3
		96.64%	3.45%	3.36%

It can be seen from the above table that the detecting rate is high; however, the un-detecting rate is also high. In real situation, common sex education may also be identified as unhealthy information. Therefore, an assistant method is needed and necessary to improve the detecting rate and reduce un-detecting rate. In the following part, skin detecting method is described.

3 Picture Detecting Algorithm

Picture is another important element in the web-pages, especially in the unhealthy adult web-pages. Therefore, filter algorithm will be more accurate and effective if picture could be detected inerrably.

3.1 Algorithm model

The main purpose of picture detecting is to find out unhealthy photos in web-pages. Unlike text analysis, no certain criterion can be used to make qualitative photograph analysis. Therefore, the major characteristic of unhealthy picture is the only aspect that can be used to program. In our algorithm, "contains large area of naked skin" is used to distinguish healthy and unhealthy pictures.

According to Anil's complexion model:

$$r = \frac{R}{R+G+B}, g = \frac{G}{R+G+B}$$

(10.333< r <0.664; 20.246< g <0.398; 3 r > g; 4g> = 0.5-0.5r)

Traversal all pixels in the picture, if the pixel (RGB) is a result of above expressions, it will be judged as skin. If the number of skin pixels exceeds a certain scalar, the photo is judged as unhealthy picture, and the web-page is judged as unhealthy web-page.

3.2 Ttest result

Table 3 shows the test result according to above Algorithm.

Although the rate of picture detecting algorithm is not as high as text analysis algorithm, it is useful and effective when it works as an assistant method.

Table 3 Test Result

pictures	Unhealthy pictures	Detected pictures	Wrongly detected pictures	Undetected pictures
138	46	31	19	15
200	46	31	22	25
78	30	22	10	8
46	28	20	5	8
		69.89%	20.89%	30.11%

4 Judging Algorithm

Judging algorithm is a linear return based on the result of text detecting algorithm and picture detecting algorithm. Assume that the result of text detecting algorithm is V_T, the result of picture detecting algorithm is an integer number (the bigger the number, the more likely the picture is unhealthy), a mathematic model can be constructed:

$$\begin{cases} V_I = \dfrac{\sum_{i=1}^{n}(V_i * S_i)}{\sum_{i=1}^{n} S_i} \\ V_{total} = V_T + V_I + k \end{cases}$$

Where, V_i is an analysis result of any picture exists in the web-page; S_i is the area of the picture; V_I is the average value of all detecting pictures. k is a coefficient for adjusting; V_{total} is the final analysis result of the web-page.

In our testing environment, we suppose "200" is the line between common web-page and unhealthy web-page, step level is 4, k=0, than choose 120 common web-pages, 40 sexology web-pages which may easily be judged as unhealthy web-page and 70 unhealthy web- pages which contains sex, forcible and dirty keywords as our test samples. The test result can be seen in follow pictures.

Figure 2 indicates that when the value of C is between 110 and 120, the detecting result is the best, and the rate of wrong detecting is less than 5%. Therefore, the line between healthy web-pages and common web-pages can be defined as 115.

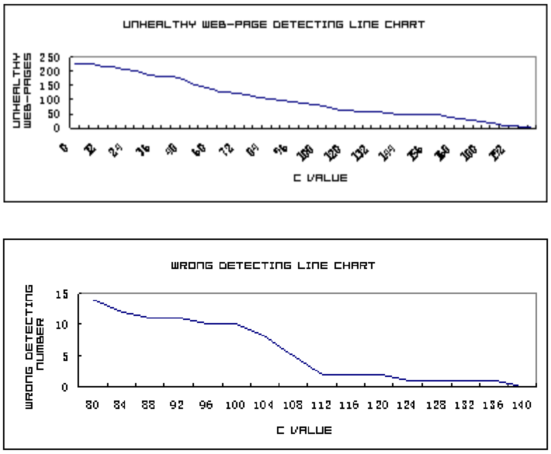


Figure 2 Line Chart of Detecting Rate (C's value plays an important role in the accuracy of detecting)

5 Summarize

Table 4 shows the final test.

Table 4 Final Test Result

Total number	Unhealthy web-pages	Detecting number	Un-detecting number	Wrong detecting number
260	104	102	2	2
180	36	36	0	0
		98.57%	1.43%	1.43%

In the new algorithm, the rate of detected web-page improves 2% while the rate of un-detected web-page and wrongly detected web-page decreased dramatically. Therefore, the new algorithm is more effective.

Multi-media are the tendency of Internet. Picture, audio, and video will become the major information-carriers on the Internet. In order to identify and detect the unhealthy information that hide in these files, chaotic feature extraction, genetic neural network and convex rough set are three major fields that should be imported to improve the algorithm.

References

[1] Anil.K.Jain and R.L.Hsu, A.M.Mohamed,Face Detection in Color Image,ICIP'2001

[2] Margaret Fleck, David Forsyth, and Chris Bregler (1996). Finding Naked People 1996 European Conference on Computer Vision , Volume II

[3] Chang, C.H. and Chen C.D. 1993. A study on integrating Chinese word segmentation and part-of-speech tagging. Communications of COLIPS 3.2.69-77

[4] Lai,B.Y.,Sun M.S.,et al.1992.Tagging- based first order Markov model approach to Chinese word identification, Proceedings of 1992 International Conference on Computer Processing of Chinese and Oriental Languages, Florida

[5] ASHISH A , KNOBLOCK C. Wrapper generation for semi - structured internet sources J . SIGMOD Record, 1997,26 4 :8- 15

[6] LU TC, LEE CC, HISA WY. Supporting large-scale distributed simulation using HLA[J].ACM Transactions on Modeling and Computer Simulation, 2000,10(3):268-294

[7] Zhang Li, Zhou Weida,Jiao Licheng A Kernel Clustering algorithm [J. Chinese Journal of Computer2,2002,25(6):587~590

[8] Aoe,J. An Efficient Digital Search Algorithm by Using a Double-Array Structure. IEEE Transactions on Software Engineering. 1989，(9)

[9] Holland J H. Adaptation in Natural and artificial system[M. Ann Arbor: University of Michigan Press,1975

[10] Kamel SM. New algorithms for solving the fuzzy c-means clustering problem[J. Pattern Recognition, 1994,27:421

The Research and Design of Web Text Mining System Framework

Fanrong Meng[1] Xiaoyun Jiang[1] Lijun Shen[2] Lei Shi[2]

1 School of Computer Science and Technology, China University of Mining and Technology,
Xuzhou, Jiangsu, 221008,China

2 School of Command Automation , PLA University of Science and Technology,
Nanjing , Jiangsu , 210007, China
Email: mengfr@cumt.edu.cn Tel: 13952118025

Abstract

With the flood of the data on the Web, Web data mining has become the focus of the data mining technology .This paper introduce the conception of Web Mining, analysis the difference between Web Mining and Data Mining. On the base of improving Maximum Matching Method, studying Vector Space Model and text classification algorithm, provide a Web text mining system framework, design the module of the framework and validate the system lastly.

Keywords: web data mining; maximum matching method; vector space model; text classification; web text mining system framework

1 Introduction

With the development of Network techniques, informationization is in fast progress and people have greatly improved their ability to search information. People face with the challenge that is “data submerged, but hunger in the knowledge”. “How can we prevent ourselves from being drowned by the ocean of information and promptly discover useful knowledge to improve utilization of information” is an urgent problem to solve .In such circumstances, Data Mining Technology come into being and flourishes, which has shown strong vitality.

Web mining is developed from data mining, but it has some uniqueness compared to traditional data mining methods. Some characteristics decide that it is impossible to apply traditional data mining methods and models to web mining. How to solve the problems of Web data standardization and pretreatment, and integrate the mining system closely with the database, and provide an integrated information processing environment, has become an important prerequisite for Web mining [1, 2].

1.1 Introduction to web mining

Web Mining [2][3] is a comprehensive technology involved to the Web, data mining, computer linguistics, information science, and other fields. Web mining is to extract some interesting potential useful information or schema from Web documents and activities. It can also be defined from a more general perspective: Web mining is to try to find a potential pattern *p* from a large collection of Web documents and materials called *C*. Consider *C* as input, *p* is output, Web mining can be seen as a mapping from input to output: ξ：C→p。

The distinction between Web mining and data mining is shown as follows:

(1) The object of data mining is data in relational table while the object of web mining is a great deal of heterogeneous distributed Web documents.

(2) Data mining discovers knowledge by use of relational table which is a kind of standard storage

* The National Natural Science Foundation of China (50674086)

structure. Web is logically a map made of document nodes and hyperlinks. Because of web documents are semi-structured or unstructured and lack of semantics which can be understood by machine, some data mining techniques are not suitable for Web mining.

The diversity of information on the Web determines the diversity of Web mining tasks. According to the object of web mining, Web mining is divided into three categories: Web Content Mining, Web Structure Mining and Web Usage Mining [4][5]. Web mining category is shown in Figure 1.

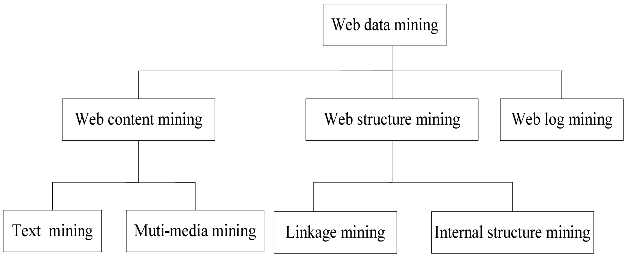


Figure 1 Web mining category

1.2 The summary of web text mining

Web text mining is a branch of data mining, and it analyses the extraction of web text information by the principle of computer linguistics. Web text mining can do lots of jobs on content of Web documents such as summary, classification, clustering, association analysis and trend forecasting.

Among these above, summary is to extract key information from the document and give a short abstract of the document. Classification is to determine the category for each of the documents collection according to the pre-defined categories theme. The difference between classification and clustering lies in that clustering does not depend on pre-defined categories theme and aims at dividing documents into several clusters demanding the similarity of document content in the same cluster as much as possible, and that between different clusters as small as possible.

The difficulties of Web text mining

The Web is a loose distributed information system ,which can expand without limitations and has no centralized control, no unified structure, no integrity constraints, no affairs management, no standard data model and query language. Web mining has to deal with the data on Web, of which the main characteristic is semi-structured. The main problem of web mining is shown as follows:

(1) Heterogeneous Database Environment: from the view of database research, information on the Web site also can be seen as a bigger, more complex database. Every Web site is a source of data. Each data source is heterogeneous, so the information and organization of each site is not the same, which constitutes a huge heterogeneous database environment.

(2) Semi-structured data structure: Data on the Web is different from data in the traditional databases. Traditional databases are based on specific data models, so one can describe specific data according its models. On contrast, data on the Web is very complex and lack of specific model description. Each web site is designed independently and data has dynamic variability. Therefore, data on the Web is not a kind of complete structured data, but usually called semi-structured data.

(3) How to solve the problem of semi-structured data source: First of all Web mining has to solve the problems of semi-structured data source and enquiries and integration of semi-structured data model. So we have to find a semi-structured data model which can describe the data on the Web clearly. Besides, we also need a extraction technique for semi-structured model by which we can extract semi-structured models from existing data.

2 The Web Text Mining System and the Framework and Algorithm Research

2.1 The web text mining system and the framework

Based on the research of Web text mining and XML, we explore a Web text mining system and the framework as shown in Figure 2.The basic idea of the algorithm is as follows:

(1) First, the mining parameters are read and sent to search engine through the user interface, and then the

search engine completes the task of collecting Web documents. There are XML documents and HTML documents in the collecting data, so we should convert the HTML documents into XML documents and then store them in the Web Warehouse. Because XML can only describe semi-structured data, it is still needed to extract features. In our methods, first we do the segmentation and then describe the features.

(2) The mining synthesizer will choose a specific text classification algorithm to mine the text.

(3) Finally, the results evaluation model will evaluate the mining results. If the Enquiries accuracy value and the Enquiries completeness value both exceed a certain threshold, the result will be shown to the user. Otherwise, it will go back to (2) to choose some other text classification algorithm to mine the text.

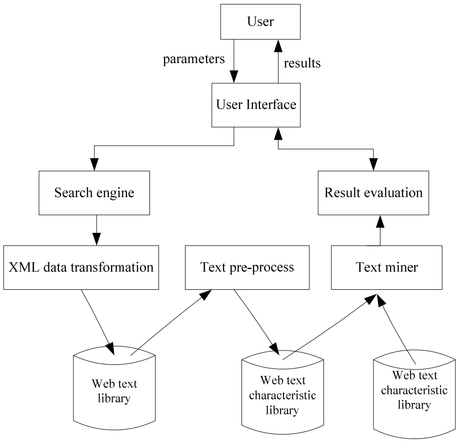


Figure 2 the design of web mining system and framework

2.2 Data conversion by xml

2.2.1 Web data extraction process

Web data extraction process is shown in Figure 3.

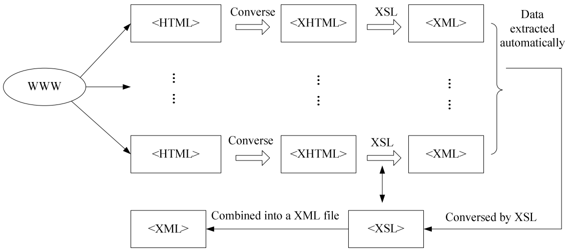


Figure 3 Web data extraction process

(1) Get information of HTML pages. It is only from the most stable and most reliable data sources that the extracted data will be accurate. The system takes web text data as analysis object. After the data source is determined, the first step is to convert HTML into XHTML. There is an algorithm which takes a URL string as input parameter, analyses the information of HTML pages and convert it into an XHTML document.

(2) Find the cited point of data. Most of the information of Web pages and source view of XHTML is not useful for us. What we have to do is find a specific area in the XML tree and extract the data. The web page structure is unchanged, no matter what the data change to be. Therefore, if we find the cited point we will be able to extract data from web pages through the cited point and matched path.

(3) Map data to XML through XSL. XPath is an important branch of XSL, which is specially designed to locate the XML documents or other documents. Hence, it is the core functions of XPath to describe the location of certain resource.

(4) integrate the results and process the data. Real-time data is what we need, so only preserving the individual data is meaningless. It is demanded to extract data again and again to combine different data records to a XML document.

2.3 Text pre-processing

2.3.1 Research on Chinese Automatic word Segmentation

Chinese Automatic word Segmentation is the premise of Text Mining. A Chinese character is the smallest unit of written Chinese, but in Natural Language Processing a word is the smallest unit that makes sense. Chinese Automatic word Segmentation is to convert a series of characters into a series words which can be processed later, that is, to establish the Chinese words border. [6,7,8]

2.3.2 Improvement on the algorithm matched largest

The algorithm Matched Largest is also called MM.

Its main idea is described as follows. If the biggest word record in the segmentation word dictionary or vocabulary includes i characters, the first i characters of the processing object are taken out to match the records in the dictionary. If these i words are found in the dictionary, matching succeeds and the matched characters are segmented as a word. Otherwise, matching fails and the last character will be removed. The remained i -1 characters will be processed continuously till a certain matching succeeds. Then it steps forward, the processing will not stop until the input stream is over. Because it is necessary to search in the word dictionary frequently, this algorithm performs not so well. The algorithm has been improved and its main idea is described as follows.

The first step: Create the index for the dictionary and put the input stream into a String variable fstr.

The second step: Get the processing character from input stream and put the dictionary records with the beginning of this character into a collection of records rs.

The third step: Search the matched results in collection of records rs according to the traditional algorithm and move steps on input stream at the same time. If matching succeeds, the matched characters are segmented. Otherwise, the last character is to be removed.

The fourth step: Check whether the input stream is over. If it is not over, go back to the second step. Otherwise, the algorithm exits.

Assume the count of dictionary records is m, the length of string is l and the counts of records is n(n<<m).The time cost of traditional algorithm is O (m×l)while that of the improved algorithm is (m +l×n). Because of n<<m, the time cost of the latter is smaller than that of the former.

2.3.3 Research on text feature model

The basic idea of VSM [5] is that a text document is composed of a group of characteristics ($T_1, T_2, ..., T_n$,). In accordance with its importance degree, each characteristic is given a weight W$_i$ (usually using TF-IDF algorithm).Thus, the characteristics of a text can be seen as an n-dimensional coordinate, in which $W_1, W_2, ..., W_n$, are the coordinates values. Every document d can be expressed as a standardized feature vector $v(d) = (t_1, w_1(d); t_2, w_2(d); ...; t_n, w_n(d))$, in which t$_i$ is entries, $W_i(d)$ is weighted value for t$_i$ in document d. In order to improve the accuracy of content, all the words or phrases in document d can be t_i

W$_i$(d) is generally defined as a function of *tf$_i$(d)* which is the count of t_i in document d. The function is $W_i(d) = \varphi(tf_i(d))$ .Another function TFIDF is $tf_i(d) \times \lg \frac{N}{n_i}$, in which N is the data of all the documents and n_i is the number of documents including t_i.

Described by VSM model, the documents set becomes a matrix. Each line represents a document, while each column represents a particular characteristic of the document.

2.4 Research on web text classification algorithms

Text classification is to determine the category of the text automatically according to the contents of text. From the mathematical point, text classification is the process of mapping which put the text into a specific given category. The mapping can be 1:1 or 1:n, that is a text can correspond to one or multiple categories. The mapping can be expressed in a mathematical formula f： A→B in which A represents text set to be classify and B represents category set of the category system. The mapping rules of text classification is the summarized classification laws in accordance to the data information of every kinds of samples. When a new text comes out, one can determine its category in accordance to the mapping rules of text classification.

Now there are three common classification algorithms: simple vector distance classification, the Bayesian algorithm, KNN algorithm. All the three algorithms are carefully studied in this section.

2.4.1 Simple vector distance classification algorithm

The classification idea of this algorithm is very simple. First of all, according to the arithmetic average we create a central vector for each category which represents the category information. When a new text is processed, the vector of this new text should be built. Then the metric between this vector and central vector is computed in order to determine the category.

Step one: Compute the central vector for each category of text sets, that is, compute the arithmetic average of all the training text vectors.

Step two: Do segmentation for the new text and represent the text using characteristic vector.

Step three: Compute the similarity between the new text characteristic vector and central vector of each category. The formula is as Eq.(1):

$$Sim(d_i,d_j)=\frac{\sum_{K=1}^{M} w_{ik}\times w_{jk}}{\sqrt{\left(\sum_{K=1}^{M} w_{ik}^2\right)\left(\sum_{K=1}^{M} w_{jk}^2\right)}} \tag{1}$$

In the formula, d_i is the characteristic vector of a new text, d_j is the central vector of category j, M is the dimension of characteristic vector, and Wk is the dimension K of the vector.

Step four: Compare the similarities between new text and each central vector and put the text into the category whose similarity is the greatest.

2.4.2 Bayesian algorithm

The algorithm basic idea is to compute the probability of a text belonging to a certain category. The probability of a text belonging to a certain category equals to the integrated expression of each probability of a word belonging to a certain category. The specific algorithm steps are as follows:

Step one: Compute the vector of probability of characteristics belonging to each category ($W_1, W_2, ..., W_n$), in which there is Eq.(2)

$$w_k = p(w_k \mid C_j) = \frac{1+\sum_{i=1}^{|D|} N\left(w_k, d_i\right)}{|V| + \sum_{s=1}^{|v|}\sum_{i=1}^{|D|} N(w_s, d_i)} \tag{2}$$

Step two: Do segmentation for the new text and compute the probability of text d_i belonging to category C_j according to the formula as Eq.(3):

$$p(C_j \mid d_i;\hat{\theta}) = \frac{p(C_j \mid \hat{\theta})\prod_{k=1}^{n} p\left(w_k \mid C_j;\hat{\theta}\right)^{N(w_k,d_i)}}{\sum_{r=1}^{|C|} p\left(C_r \mid \hat{\theta}\right)\prod_{k=1}^{n} p(w_k \mid C_r;\hat{\theta})^{N(W_k,d_i)}} \tag{3}$$

There is EQ(4)

$$p(c_j \mid \hat{\theta}) = \frac{count\ of\ c_j}{total\ count\ of\ documents} \tag{4}$$

$p(C_r \mid \hat{\theta})$ is the similarity, |C| is the count of categories, N(W$_k$,d$_i$) is the count of W$_k$ in d_i and n is the count of characteristics.

Step three: Compare the probability of a new text belonging to each category and put the text into the category whose probability is the greatest.

2.4.3 Knn (k-nearest neighbor algorithm)

The basic idea of the algorithm is to analyze the K pieces of texts in the training set which is most similar with the new text and determine the category the new text belongs to[9,10,11]. The specific algorithm steps are as follows:

Step one: Describe training text Vector according to characteristics

Step two: Do segmentation for the new text according to characteristics and determine the vector for the new text

Step three: Choose the K texts from the training set which is most similar to the new text. The computing formula is as formula Eq.(1).

Now it lacks of a good method to determine the value of K. Generally K is firstly given an initial value from several hundred to several thousand and then adjusted latter according to the test results.

Step four: Calculate weighted value for each category of the new text's K-neighbor one by one. The calculation formula is as Eq.(5):

$$p\left(\overset{\varpi}{x}, C_j\right) = \sum_{\overset{\varpi}{d}_i \in KNN} Sim\left(\overset{\varpi}{x}, \overset{\omega}{d}_i\right) y\left(\overset{\omega}{d}_i, C_j\right) \tag{5}$$

The $\overset{\varpi}{x}$ is the characteristic vector of the new text, Sim($\overset{\varpi}{x},\overset{\omega}{d_i}$)is the calculation formula for similarity which is the same as the formula of step 3, $y(\overset{\omega}{d_i},C_j)$ is the category attributes function, that is, if $\overset{\omega}{d_i}$ belongs to the category C_j the function result is 1,otherwise the result is 0.

3 The Module Design of Web Text Mining System and Framework

3.1 The figure of system architecture modules

According to the design and analysis, the web text mining system is composed of 4 modules, which are shown in Figure 4.

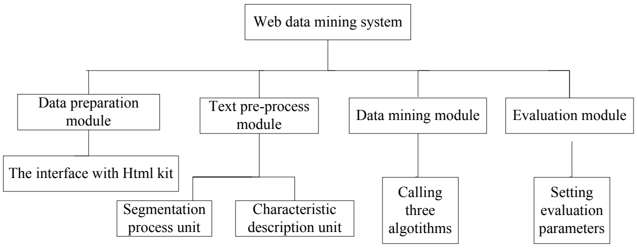


Figure 4 web text mining system modules

3.2 System design

3.2.1 Data preparation module

This module implements the conversion from HTML to XML. For the facilitation of using this software, the interface of the software is implemented. The running process of the module is shown in Figure 5.

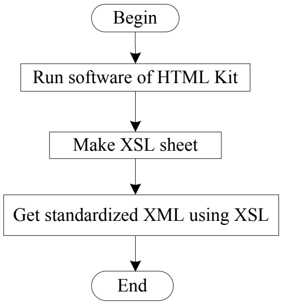


Figure 5 Data preparation program flow

3.2.2 Document pre-processing module

This module is divided into three processing units. The first unit is text segmentation unit. The input of text segmentation unit is the processed document with XML technique. The output of this unit is a document with segmentation marker, which provides the facilitation for feature extraction at next step. The second unit is text feature represent unit, which makes use of VSM to express the relationship between words and documentation as a matrix by means of calculating characteristic weighted value. The third unit is text feature extraction unit. In order to compress characteristic, this unit reduces the dimensions of text characteristic vector by calculating mutual information value between words and categories. The running process of the module is shown in Figure 6.

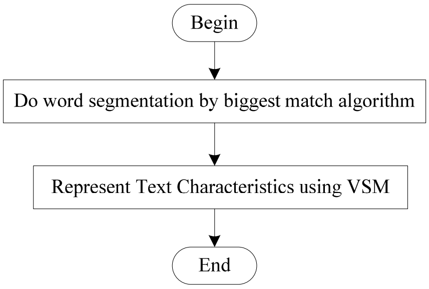

Begin
Do word segmentation by biggest match algorithm
Represent Text Characteristics using VSM
End

Figure 6 document pre-processing program flow

3.2.3 Text mining module

There are three algorithms we analyzed Simple Vector distance classification algorithm, Bayesian algorithm and KNN (K nearest neighbor algorithm). The strategy used in this system is described as follows: the user chooses one from the three algorithms and chooses the radio button to trigger the running process.

The running flow of this module is to run the chosen algorithm and get the category which the document belongs to. As for the document difficult to decide the category, the threshold is used to help with deciding the category. The running process of the module is shown in Figure 7.

3.2.4 Mining results and Evaluation

There are two indicators to evaluate the text classification system, precision and recall.

Precision is count of accurate classified documents divided by the count of all the documents processed. Its mathematical formula is expressed as Eq.(6):

$$precision = \frac{count\ of\ accurate\ classifed\ documents}{count\ of\ all\ the\ documents} \quad (6)$$

Recall is count of accurate classified documents divided by the count of total documents. Its mathematical formula is expressed as Eq.(7):

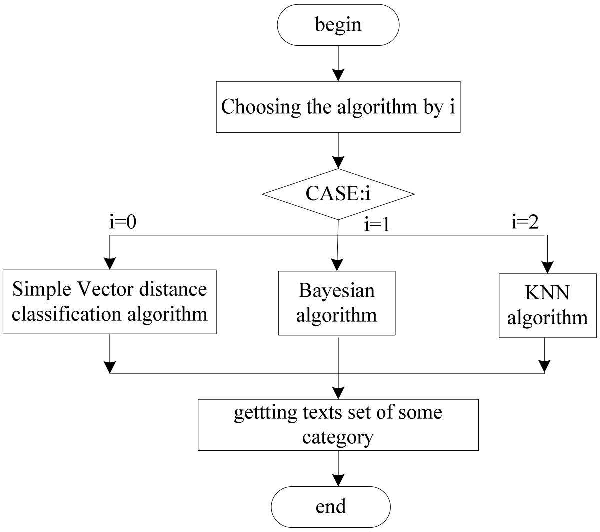

begin
Choosing the algorithm by i
CASE:i
i=0
i=1
i=2
Simple Vector distance classification algorithm
Bayesian algorithm
KNN algorithm
gettting texts set of some category
end

Figure 7 document mining module program flow

$$recall = \frac{count\ of\ accurate\ classifed\ documents}{count\ of\ all\ the\ documents\ processed} \quad (7)$$

Precision and recall reveal two different aspects of classification, which should be combined. Thus, a new evaluation standard is proposed named as F1 test value. Its mathematical formula is expressed as Eq.(8):

$$F1 = \frac{precision \times recall \times 2}{precision + recall} \quad (8)$$

According to the algorithm used in the text, we have the three words software, literature; basketball tested in 148, 65, and 80 documents respectively, the test results are in table 1 as follows.

From the test results, the three indicators precision, recall and F1 test value are all above 80%, which is satisfying.

Table 1 Evaluation of test results

category	software	literature	basketball
total texts count	148	65	80
classified texts count	126	60	71
accurate texts count	114	52	67
recall	85.1%	92.3%	88.8%
precision	90.5%	86.7%	94.4%
F1 test value	87.7%	89.4%	91.5%

4 Summaries

On the basis of web text mining analysis, a new web text mining system and framework is proposed, which can overcome the problems of semi-structured data source and large quantity of data. Our main researches include:

(1) Analyze the difference between data mining and web mining, and find the difficulties in web mining.

(2) Analyze and optimize the MM algorithm to improve the matching performance, analyze VSM to express a document clearly, and analyze three text classification algorithms and their using domain.

(3) Propose a new web text mining system and framework and design the running flow of the modules. Finally the mining results are evaluated through some evaluation indicators, and it verifies that this system could get satisfying results.

References

[1] Han Jia-wei, Meng Xiao-feng.Web mining Research[J]. Journal of Computer Research and Development, 2001,38(41):pp405-410

[2] WangShi,GaoWen. Web data mining [J]. Journal of Computer science, 2000,27(4):pp28-31

[3] R Kosala, H Blockeel.Web Mining Research: A Survey [J].SIGKDD Exploration, 2000, 2(1):pp1-15

[4] ChenLi, Jiao Li-Cheng. Internet / Web data mining esearch and the latest progress[J]. Xi Dian university.(Science), 2001,28(1):pp114-119

[5] Liu Zhuo. Automatic Chinese text classification algorithm based on KNN[D].Jilin university, 2004.pp20 25

[6] D. Mladenic and M. Grobelnik. Efficient text ategorization[C]. Presented at Proc. Text Mining Workshop 10th European Conf. Machine Learning ECML98

[7] Alex Markov, Mark Last. Model-Based Classification of Web Documents Represented by Graphs[C]. In Proc. of WebKDD 2006, 2006, Philadelphia, PA

[8] P. Kingsbury and M. Palmer. Propbank: the next level of treebank[C]. In Proceedings of Treebanks and Lexical Theories,2003

[9] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques[C]. In Knowledge Discovery and Data Mining (KDD) Workshop on TextMining, August 2000

[10] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering[C]. In Proceedings of 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI), pages 58–64,2000

[11] D. Gildea and D. Jurafsky. Automatic labeling of semantic roles[C]. Computational Linguistics, 28(3): 245–288, 2002

Fan-rong Meng is an Assistant Professor and a head of digital mine Lab of School of Computer Science and Technology, China University of Mining and Technology. She graduated from China University of Mining and Technology in 1984. She has published two books, over 20 Journal papers. Her research interests are in distributed parallel processing, database theory, data mining, data fuse.

Self-design of Virtual Measuring System for Bionic Plough

Shucai Xu[1] Jinhuan Zhang*[1] Jianqiao Li[2] Rui Zhang[2] Lianghua Zhu[3]

1 State Key Laboratory of Automotive Safety and Energy, Tsinghua University, Beijing, 100084, China
Email:xushc@tsinghua.edu.cn

2 Key Laboratory of Terrain-Machine Bionics Engineering (Jilin University),
Ministry of Education Changchun, 130025, China

3 Shanghai Ouji Metallurgy Machine Manufacture Co., Ltd Shanghai, 201907, China

Abstract

Bionic plough is the actual application of the bionics on the agricultural engineering .Based on the law of action and reaction, the working capability of the bionic plough is always indirectly analyzed by measuring and studying the forces F acting on the bionic plough by soil currently, so a stress-strain virtual measuring system for bionic plough, which was designed by the graphical programming language DASYLab, was used to measure the horizontal force F. In addition, during the course of design, the experimental complexities and the interferential factors influencing on signal logging were analyzed when bionic plough worked, so the anti-jamming methods of hardware and software technology were adopted correlatively. In the end, the self-design virtual measuring system indicates more merits than the traditional measuring system on many aspects including data real-time display, data real-time storage, data memory format, data processing and replay, and operation convenience. It shows much higher ratio of performance to price.

Keywords: Virtual Measuring System, Traditional Instrument, Anti-jamming, Bionic Plough, Rational of Performance to Price

1 Introduction

The constituent of soil is very complex, and the force acting on the surface of tool by the disturbed soil is unstable, so it is difficult to get the same result of filed experiment under the approximate condition. Therefore, how to design a suitable measuring system is important [1,2].Today, measure instruments and electronic technology developed quickly, however, as for traditional measure instruments, even if digital instruments and intelligent instruments which heighten the accuracy and function of the traditional measurement and analysis system, they still can not change their some shortcomings, such as single function, handle manipulating, large and ponderous volume, fixed channel number and high cost to develop functions etc. Aimed at the above demerits, a kind of virtual instrument that is developed with PC and correlative software emerged. Based on hardware, the key measure functions of virtual instrument depend on software, so the ant-jamming methods of hardware and software can be used to eliminate disturbance including static, magnetic filed and random error. Because of above merits, a stress-strain virtual measuring system for bionic DASYLab[3-5].

2 Hardware and Software Design of the Stress-Strain Virtual Measuring System

According to the flow of signal, both the measure theory of the stress-strain virtual measuring system for

[*] Corresponding author

bionic plough and the different anti-jamming measures were shown in Figure 1. In order to improve measuring accuracy, three demands were put forward as follows. Firstly, the anti-jamming technologies of hardware and software were used to reduce interferential factors correctly; Secondly, sample size must be same in every experiment, in this way, the acquired data can be analyzed with statistic theory; Thirdly, the interference of random noise must be minimized.

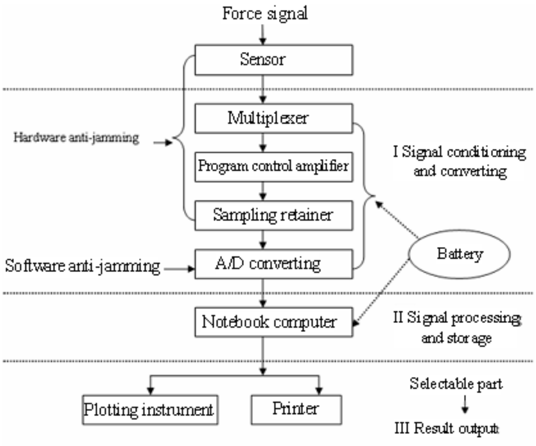


Figure 1 Diagram of the data acquisition principle and the anti-jamming method

Based on above demands, the virtual measuring system for bionic plough was designed with the graphical programming language DASYLab, which was substituted for the traditional stress-strain measuring system, data logging system, and data analysis system on the test of the horizontal force acting on the bionic ploughs. From the aspect of hardware, it was constituted of computer, DAQP-12 and QTC-300 data logging sets, as shown in Figure 2[6]. plough was designed by graphical programming language

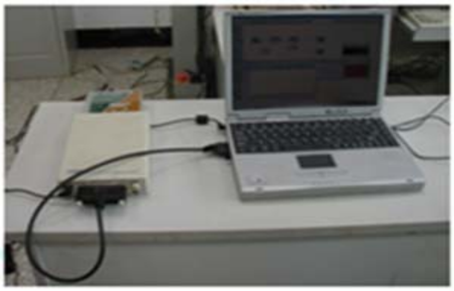


Figure 2 Hardware component of the virtual measuring system

Seen from the Figure 2, a signal condition module QTC-300 was used to condition the analog signal from the force sensor (Type: BLR-1). In order to convert the sensor output (volt) into load (kN), the force sensor must be validated before experiment correctly. The QTC-300 is a four channel strain gage signal conditioning adapter, which was shown in Figure 3. Each channel is equipped with separate instrumentation amplifier, multiple excitation current source options and a fourth order low pass filter. Each channel has selectable gain of 1, 10,100 or 500. In order to measure strain accurately, the QTC-300 must be calibrated using the DaqCal Utility and configured correctly when a new sensor is added to the board, or a previously used sensor is reconnected to the board every time.

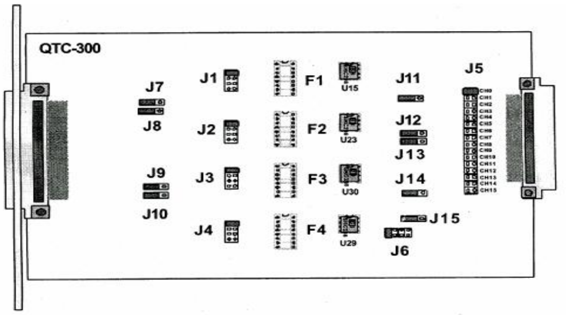


Figure 3 QTC-300 jumper and filter block locations

Figure 3 shows the location of the main A/D channel and board selection jumper blocks (J5/J6), SSH option jumper block (J11), A/D gain selection jumper blocks (J1 through J4 for input channels 0 through 4 respectively), filter selection jumper blocks (J12 through J15), filter blocks (F1 through F4) and the AC/DC coupling selection jumpers (J7 through J10). The settings for J5 and J6 are “CH0” and “BD0”. The option for J11 is “SSH disabled”. Because the 0V value on data acquisition board corresponds to 0 mV at the pre-amplifier, the configuration for J1 is “1000”. Using the A/D gain selection option provided on the QTC-300 interface board would result in less signal noise than using the A/D gain selection options available on the data acquisition adapter in the host computer. In order to filter the high frequency disturbing signal, J12 and F1 are set “LP” and “80Hz”. As shown in Figure 1, DAQP-12 is a 12-bit analog input PCMCIA card. The parameters of daqp-12 were set using DAQDRIVE ConFigure Utility .The base address and IRQ level of DAQP-12 are set “10A0” and “15”. The channels, input mode and signal type of A/D Converter are configured “15”, “single-ended” and “bipolar” respectively. The select driver of DAQP-12 is “daqlab32.dll”. In addition, the signal conditioning adapter above mentioned was a series. Therefore, the virtual measuring system for stress-strain can realize many other measure functions, such as measuring temperature, velocity, and acceleration, via substituting corresponding signal conditioning adapters for QTC-300. And what’s more, the channels of the virtual measuring system can be expended multiply by adding the amount of the same signal conditioning adapter[7]. However, the traditional stress-strain system changes the function or adds the channels through adding the amount of the strain gauge NEC16 made in Japan, as shown in Figure 4. The price of NEC16 is 300 thousand RMB, so the total price of the traditional stress-strain measuring system and the data logger is 380 thousand RMB. However, the total price of the stress-stain virtual measuring system is 40 thousand RMB. Therefore, compared with the former, the latter indicates more superiority on the extensibility and price.

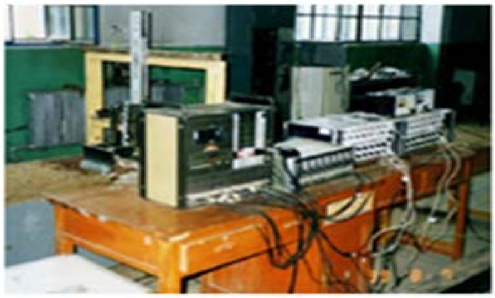


Figure 4 Hardware composing of the traditional measuring system for bionic plough

The virtual measuring system for bionic plough, which was used to acquire signal of the horizontal force F acting on the bionic ploughs by disturbed soil, was designed by the graphical programming language DasyLab. Considered the signal features of soil experiment, the stress-strain virtual measuring system was designed by choosing function modules including A/D, FFT, Y/t, Filter, Switch, Action, Bar Graph, Dig.meter, Scaling, Arithmetic, Ref.Curve, Write, and Status lamp modules from the module bar of DASYLab to realize the measure functions in Figure 5.

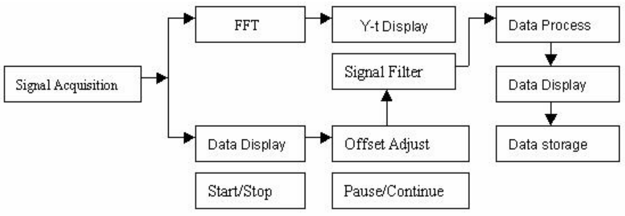


Figure 5 The work flow of software

Additionally, during the course of software design, the experimental complexities and the interferential factors influencing on signal logging were analyzed when bionic plough worked, so the anti-jamming methods of software technology were adopted. For example, the best method that dispels the random noise was real-time fitting disposal, so the module, Regression, was set to least squares method. After the above modules were set correctly, the stress-strain virtual measuring system had the following functions, such as data real-time acquisition, data real-time display and storage, and data real-time analysis and control. It can also realize the data replay after experiment[8-10]. Figure 6 shows its virtual operation panel.

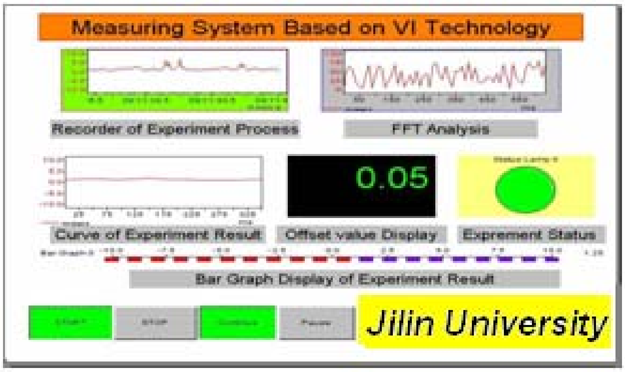

Measuring System Based on VI Technology
Recorder of Experiment Process
FFT Analysis
0.05
Curve of Experiment Result
Offset value Display
Exprement Status
Bar Graph Display of Experiment Result
Jilin University

Figure 6 The operation panel of the self-design virtual measuring system for bionic plough

3 Results of Application and Discussion

The stress-strain virtual measuring system for bionic plough was substituted for the traditional measuring system, data logging system, and data analysis system on the test of the horizontal dynamic forces F acting on the bionic ploughs by disturbed soil in the field.

Figure 7 shown the horizontal forces acting on the bionic plough at two kinds of working speed (e.g. 0.71m/s and 0.83m/s), and the experimental forces were recorded five times at the approximate speed. Seen from Figure 7, it can be deduced that the resistant forces for bionic plough increased with increase of speed from 0.70 m/s to 0.93 m/s. and the resistant force F acting on the bionic ploughs changed greatly at different working speeds, and the forces were unequal in spite of the approximately equal working speeds every time. On the one hand, the reason was that the factors including water content of soil, flat of ground and stability of working speed etc, had not constant value in the field. On the one hand, this also agreed with Wolde (1997) finding that increased speed might result in more rapid acceleration of soil mass which eventually increased the normal load on the soil engaging surface due to frictional force and the kinetic energy transmitted to the soil[11] Meanwhile, the errors of experiment result were very small, which was satisfied with the demand of the field experiment, so the virtual measuring system for bionic plough based on virtual instrument technology was credible. In addition, the measure results of field experiments have bad reproducibility, which was testified again. From above all, the results of application indicated that the virtual measuring system for bionic plough based on virtual instrument technology not only had measuring functions which the traditional measuring system had, but it realized data real-time processing and analyzing successfully. And furthermore, it was substituted the traditional data processing instrument for the data computation, analysis, process and storage by fully using powerful functions of computer completely.

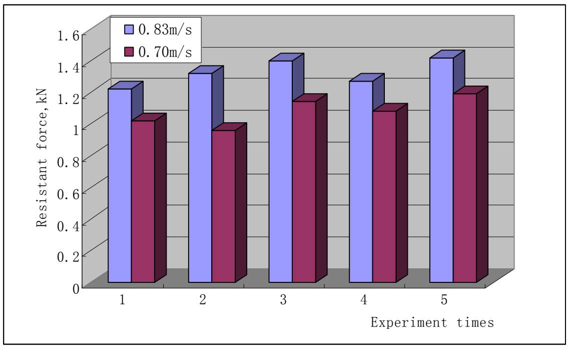


Figure 7 The resistant forces of the bionic plough at two working speeds

4 Conclusions

The qualitative and the quantificational analyses of the interactions between the bionic plough and the soil under different conditions indicate that the stress-strain measuring system for bionic plough was substituted for the traditional measuring system, data logging system, and data analysis system on the test of the horizontal dynamic forces acting on the surface of a bionic plough by soil successfully. And what's more, the stress-strain virtual measuring system also indicated more merits than the traditional measuring system on many aspects. Firstly, its functions, including instrument self-checking, data real-time acquire, data real-time storage, data processing and replay etc, were all completed by the computer, so it had the very high automaticity. Secondly, the data of experiment was analyzed with error theory, the result shown that the stress-strain virtual measuring system had good dependability and high precision. The error of the whole measuring system is under 5 percent, which is satisfied with the field experiment. Thirdly, it was one tenth of the traditional stress-strain measuring system in volume, and it need not network power supply, so it is very suitable for the field experiments. In the end, it has much higher ratio of performance to price and made the basis on the further research on the working capability of the bionic plough.

References

[1] Zeng De-chao. Dynamics of mechanical soil. Beijing Science and Technology Press, 1995

[2] Zhang Rui, Li Jian-qiao, Cui Zhan-rong. Application of piezoelectric three-axis force sensor on mechanical analysis of disturbed soil. Proceedings of the 6th international symposium on test and measurement (ISTM/2005), Dalian, china 1-4, June 2005

[3] Salvatore Nuccio, Ciro Spataro. Assessment of virtual instruments measurement uncertainty [J]. Computer Standards & Interfaces, 2001, (23): 39–46

[4] Adrian Mihalcioiu, Lucian Dascalescu, Subhankar Das, Karim Medles, Radu Munteanu. Virtual instrument for statistic control of powder tribo-charging processes [J].

Journal of Electrostatics, 2005 (63): 565–570

[5] F.J. Jime´nez, J. De Frutos. Virtual instrument for measurement, processing data, and visualization of vibration patterns of piezoelectric devices [J]. Computer Standards & Interfaces, 2005 (27): 653–663

[6] Quan Li, Chen Zhao-zhang, Tang Ping, Cheng Li, Su Qing-zu. Design of hardware signal processing virtual instrument. Chinese Journal of Scientific Instrument, Vol 25 No 3 Jun.2004, 413-417

[7] SignalPro Series Signal Conditioning Adapters Users Manual Quatech Inc. USA 2000

[8] JinYan, Zheng Jian-rong, Wu Qing. Comparison of the realization of signal processing based on virtual instrument technology [J]. Machinery Electronics， 2003, (2): 55-58

[9] SignalPro Series Signal Conditioning Adapters Users Manual. Quatech Inc. USA 2000

[10] Zhang Zhi-jie. Research on error analysis in dynamic measurement [J]. Journal of test and measurement technology, 2004,18(2):139-143

[11] Ballel Z M, Biomimetic Design and Experimental Research of Resistance-Reducing Surfaces of Soil-Tool Systems [D]. Changchun Jilin University, 2005

Research and Application of Fuzzy Multi Attribute Decision Making Based on Information Entropy and Grey Relevance Degree

Dongjing Pan

Department of Computer Science and Technology, Dezhou University, Dezhou, Shandong 253023, China
Email: pdj1970@163.com

Abstract

With the development of uncertain theory, fuzzy multi attribute decision making has been applicated in more and more areas. Considering of previous fuzzy multi attribute decision methods, some were based on information entropy, some were based on grey relevance degree, etc. Information entropy is a kind of quantity of the system uncertainty and grey relevance degree can indicate the correlation strength between different things. In this paper, a new fuzzy multi attribute decision method is presented which combines information entropy with grey relevance degree, this method well overcomes the incompleteness and indeterminacy of the information, and provides a feasible way for fuzzy multi attribute decision making. Finally, the validity of this method is verified by an example.

Keywords: Fuzzy multi attribute decision making, Information entropy, Grey relevance degree, Attribute weight, Matrix standardization

1 Introduction

Multi attribute decision making is one important part of modern decision science, its theories and methods have the extensive application in many domains, such as engineering design, economy, management and military, etc. Because of the complexity and uncertainty of objective things and the fuzziness of man's thinking, people pay more and more attention in the research of multi attribute decision methods in uncertain environment, and has proposed some fuzzy multi attribute decision methods, for example, some are based on OWA operator, some are based on OWGA operator, some are based on information entropy, some are based on grey relevance degree, etc. In this paper, a new method of fuzzy multi attribute decision making is presented which combines information entropy with grey relevance degree, this method is applied in the decision of choosing the optimum fight aircraft, and its validity is verified.

2 Problem Description and Mathematical Model

For some one multi attribute decision problem, supposing $X=\{X_1,X_2,\cdots,X_n\}$ is a plan set, where X_i is one plan, $N=\{1,2,\cdots,n\}$, $U=\{u_1,u_2,\cdots,u_m\}$ is attribute set, $M=\{1,2,\cdots,m\}$, the weight of each attribute is unknown, and each attribute value is real number, measure the attribute u_j of plan X_i, we will obtain attribute value a_{ij} of u_j for X_i, so we can constitute the attribute property matrix A , as shown in Table 1.

Table 1 Attribute property matrix A

	u_1	u_2	$\cdots$	u_m
X_1	a_{11}	a_{12}	$\cdots$	a_{1m}
X_2	a_{21}	a_{22}	$\cdots$	a_{2m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_n	a_{n1}	a_{n2}	$\cdots$	a_{nm}

Attribute has many types, such as efficiency type, cost type, fixed type, deviate type, zone type, deviate zone type, etc. Efficiency type indicates that the larger the attribute value is, the better it is. Cost type indicates that the smaller the attribute value is, the better it is. Fixed type indicates that the attribute value is closer to one fixed value α_i, the better it is. Deviate type indicates that the attribute value deviates one fixed value β_j more, the better it is. Zone type indicates that the attribute value is closer to one fixed zone $[q_1^j, q_2^j]$, the better it is. Deviate zone type indicates that the attribute value deviates one fixed zone $[q_1^j, q_2^j]$ more, the better it is.

2.1 Attribute property matrix standardization

In order to increase the comparability and eliminate the influence of attribute’s different physical quantity to the sorting result, generally the standardization is requested first. Supposing I_i (i=1，2，⋯，6) represents the elememt subscript of attribute set which has the elements of efficiency type, cost type, fixed type, deviate type, zone type, deviate zone type. We can standardize the matrix according to the following Eq.(1) Eq.(6).

$$r_{ij} = \frac{a_{ij}}{\max\limits_i(a_{ij})}, \quad i \in N,\ j \in I_1; \tag{1}$$

$$r_{ij} = \frac{\min\limits_i(a_{ij})}{a_{ij}}, \quad i \in N,\ j \in I_2; \tag{2}$$

$$r_{ij} = 1 - \frac{a_{ij} - \alpha_j}{\max\limits_i |a_{ij} - \alpha_j|}, \quad i \in N,\ j \in I_3; \tag{3}$$

$$r_{ij} = |a_{ij} - \beta_j| - \frac{\min\limits_i |a_{ij} - \beta_j|}{\max\limits_i |a_{ij} - \beta_j| - \min\limits_i |a_{ij} - \beta_j|}, \tag{4}$$

$i \in N,\ j \in I_4$;

$$r_{ij} = \begin{cases} 1 - \dfrac{\max(q_1^j - a_{ij}, a_{ij} - q_2^j)}{\max[q_1^j - \min\limits_i(a_{ij}), \max\limits_i(a_{ij}) - q_2^j]}, a_{ij} \notin [q_1^j, q_2^j], \\ 1, \quad a_{ij} \in [q_1^j, q_2^j] \end{cases}$$

$$i \in N, j \in I_5; \tag{5}$$

$$r_{ij} = \begin{cases} \dfrac{\max(q_1^j - a_{ij}, a_{ij} - q_2^j)}{\max[q_1^j - \min\limits_i(a_{ij}), \max\limits_i(a_{ij}) - q_2^j]}, a_{ij} \notin [q_1^j, q_2^j], \\ 0, \quad a_{ij} \in [q_1^j, q_2^j] \end{cases}$$

$$i \in N, j \in I_6; \tag{6}$$

After standardizing the matrix A, we obtain the standardized matrix $R = (r_{ij})_{n \times m}$

2.2 Defining attribute weight using information entropy

The concept of entropy comes from thermodynamics at the earliest, later Shannon extended entropy to the information science. Information entropy can be used to express the uncertainty of things, it is the uncertainty measure, the larger the uncertainty is, the bigger the information entropy is. When the system has n different states: $S_1, S_2, \cdots, S_n$, the probability of each state is $P_1, P_2, \cdots, P_n$, respectively, then the system's information entropy is defined as follows:

$$E = E(P_1, P_2, \cdots, P_n) = -(1/\ln n)\sum_{i=1}^{n} P_i \ln P_i$$

, where $\sum_{i=1}^{n} P_i = 1$. The main mathematical properties of information entropy are as follows:

1) Nonnegativity: $E(P_1, P_2, \cdots, P_n) \geq 0$

2) Certainty: When someone $P_i = 1$, we will get $E(P_1, P_2, \cdots, P_n) = 0$, it shows that the state of system has been determined.

3) Extremum property: When the probability of each state is equal, that is $P_1 = P_2 = \cdots = P_n = 1/n$, the system's information entropy has the max value.

According to the information entropy theory, attribute weight is definited by the amount of information which corresponding attribute transmits to the valuator. Different attribute weight can reflect the different effect in the potency appraisal. The amount of information that attribute provides is more, the effect that corresponding attribute plays in the quality synthetic evaluation is bigger, the attribute weight definited by information entropy is bigger, it shows that this attribute is more important in synthetic evaluation. Otherwise, the amount of information that attribute provides is few, the attribute weight definited by information entropy is smaller, it shows that this attribute is more unimportant in synthetic evaluation. Attribute weight decides the precision and the reliability of synthetic evaluation.

Attribute weight can depends upon the Delphi and AHP method, but this appraisal process has some insufficiencies: The attribute weight mainly depends on expert's subjective judgment, the objectivity of weight could not be guaranteed well. Now we used the information entropy to define the attribute weight.

The information entropy of attribute u_j is:

$$E_j = -\frac{1}{\ln n}\sum_{i=1}^{n} \dot{r}_{ij} \ln \dot{r}_{ij}, \quad j \in M \tag{7}$$

Where $\dot{r}_{ij} = \frac{r_{ij}}{\sum_{i=1}^{n} r_{ij}}, \quad i \in N, j \in M.$ (8)

When $\dot{r}_{ij} = 0$, defining $\dot{r}_{ij} \ln \dot{r}_{ij} = 0$

Attribute weight vector is $\omega = (\omega_1, \omega_2, \cdots, \omega_m)$.

Where $\omega_j = \frac{1 - E_j}{\sum_{k=1}^{m}(1 - E_k)}$ (9)

2.3 Establishment of grey relevance model

The grey system is the system which the information is not known completely, that is to say, in the system, the partial information is known and the partial information is unknown. Grey relevance degree analysis is an important component of grey system theory, it is one kind of method which can analyze connection degree between various factors in system, the grey relevance degree assessment method carries on the superiority analysis, obtains the assessment results through computing the grey relevance degree between various connection factors.

According to grey relevance decision theory, the grey relevance degree between each plan's attribute vector and the relative best plan's attribute vector is to be taken as the standard to evaluate the quality of the plan. Supposing the relative best plan is $X_0 = (x_{01}, x_{02}, \cdots, x_{0n})$, then the grey relevance degree between the plan X_i 's attribute value x_{ij} and the relative best plan X_0 's attribute value x_{0j} is defined as follows:

$$\xi_{ij} = \frac{\min_i \min_j \Delta x_{ij} + \rho \max_i \max_j \Delta x_{ij}}{\Delta x_{ij} + \rho \max_i \max_j \Delta x_{ij}} \tag{10}$$

Where $\Delta x_{ij} = | x_{0j} - x_{ij} |$, ρ is distinguish coefficient, $0 \leqslant \rho \leqslant 1$, in this paper, $\rho = 0.5$.

1) Establishment of grey relevance degree matrix.

2) Computing each plan's synthetical relevance degree $z_i(\omega)$:

$$z_i = \sum_{j=1}^{m} \xi_{ij}\omega_j \qquad i = 1, 2, \cdots, n \tag{11}$$

3) According to the value z_i to sort the order of each plan, the plan X_i which $\max_{1 \le i \le n}\{z_i\}$ corresponds to is the best plan.

3 Example analysis

Considering to purchase fight aircraft, there are several kinds of fight aircrafts, The policy-maker has considered 6 evaluating factors according to the fight aircraft's performance and expense, they are u_1 — maximum flying speed(Ma), u_2 —flying scope(10^3km), u_3 —maximum load(10^4lb), u_4 —purchasing expense (\$$10^6$), u_5 —reliability(10 points), u_6 —sensitivity (10 points). Now, supposing there are 4 kinds of fight aircrafts, the attribute value of each fight aircraft is expressed in table 2.

Table 2 Attribute value of each fight aircraft

	u_1	u_2	u_3	u_4	u_5	u_6
X_1	2.2	1.8	2.1	5.3	6	9
X_2	2.5	2.7	1.9	6.3	5	6
X_3	2.0	2.1	2.2	4.3	8	8
X_4	1.8	1.9	2.0	5.0	5	5

Among each attribute of fight aircraft, except that the attribute of u_4 —purchasing expense is cost type, others are efficiency type. After standardizing the table 2 using Eq.(1) and Eq(2), we can get the standardized matrix, as shown in table 3.

Table 3 Standardized attribute value of each fight aircraft

	u_1	u_2	u_3	u_4	u_5	u_6
X_1	0.88	0.667	0.9545	0.8113	0.75	1
X_2	1	1	0.8636	0.6825	0.625	0.6667
X_3	0.8	0.778	1	1	1	0.8889
X_4	0.72	0.704	0.9091	0.86	0.625	0.5556

Computing the information entropy of each attribute u_j using Eq.(7) and Eq.(8), we can get information entropy as follows:

$E_1 = 0.994689$, $E_2 = 0.990637$,

$E_3 = 0.998926$, $E_4 = 0.993363$,

$E_5 = 0.985626$, $E_6 = 0.981379$.

Computing the weight of each attribute using Eq.(9), we can get the attribute weight vector:

ω=(0.095903，0.169065，0.019387，0.119841，0.259554，0.33625)

From table 3, we know the relative best plan is $X_0 = (1,1,1,1,1,1)$, so using Eq.(10), we can compute the grey relevance degree between each plan's attribute and the relative best plan's attribute, the grey relevance degree matrix is shown as table 4

Table 4 Grey relevance degree matrix

	u_1	u_2	u_3	u_4	u_5	u_6
X_1	0.649	0.4	0.8302	0.5408	0.471	1
X_2	1	1	0.6197	0.4118	0.372	0.4
X_3	0.526	0.5	1	1	1	0.667
X_4	0.443	0.429	0.7097	0.6135	0.372	0.333

Using $z_i = \sum_{j=1}^{m} \xi_{ij}\omega_j \quad i = 1,2,\cdots,n$, we can compute the synthetical relevance degree $z_i(\omega)$ of plan X_i , the result is as follow:

$z_1(\omega) = 0.669201$, $z_2(\omega) = 0.557407$, $z_3(\omega) = 0.757957$, $z_4(\omega) = 0.410834$.

According to $z_i(\omega)$, sort the order of plans, the result is $X_3 > X_1 > X_2 > X_4$, so the best fight aircraft is X_3 , the worst fight aircraft is X_4 , the result is in accordance with the the objective reality .

From above analysis, we can see that the quality synthetic evaluation method of fight aircraft based on information entropy and grey relevance degree uses the amount of information which the attribute provides to determine the weight of attribute. In this method, we consider six evaluating factors of fight aircraft and analyze the synthetical relevance degree between every attribute and the relative best fight aircraft. Compared with the traditional evaluation method, this evaluation method is more objective and reliable.

4 Conclusion

Fuzzy multi attribute decision making has the wide application prospect. In this paper, A new fuzzy mult attribute decision method is proposed which combines information entropy with grey relevance, and its validity is verified by an example. Fuzzy mult attribute decision model based on information entropy and grey relevance degree can determine the high credible optimal plan without expert giving attribute weight, this model selects relative best attribute from the assessed plans as appraisal standard, which can remove the grey ingredient well. The example shows that this method's sorting result is accurate, and it has feasibility. This method provides a feasible way for fuzzy multi attribute decision making.

References

[1] Xu Ze-shui, Uncertain Multiple Attribute Decision Making:Method and Applications, Tsinghua University Press, Beijing, 2005.11

[2] Li Yong-ping,Chen Min-ye, Liu Ming, "Estimation Method for Aircraft Similarity Based on Fuzzy Theory and Grey Incidence Analysis", Transactions of Nanjing University of Aeronautics &Astronautics, 2007,24(3), pp.194-198

[3] Xie Jian-xi, Song Bi-feng, Liu Dong-xia, "The Decision-model of Grey Association Degree Theory Based Forscheme Selection in Aircraft Top-hierarchy Design",.Mathematics in Practice and Theory, 2005, 35(4), pp.94-100

[4] Dong Jian-kang, Geng Hong, "Optimizing Aircraft Fault Isolation Methods Based on Grey Relation Vague Clustering Algorithm", Journal of Nanjing University of Aeronautics & Astronautics, 2004, 36(3), pp.313-316

[5] Yan Bao-wen, Fany Li, Li Jing, Xu Qin, "Study of Agro-environmental Geology System and Its State Evaluation Based on Information Entropy Theory", Journal of Northwest A & F University(Nat. Sci. Ed.), 2007,35(11), pp.223-229

[6] Bubnicki Z, "Uncertain Variables and Their Application to Decision Making Problems", IEEE Transactions on System, Man, and Cybernetics Part A, 2001,31, pp.587-596

[7] Wang Mei-yi, Zhang Feng-ming, Liu Zhi, "Evaluation

Method of the Multi-attribute Scheme Based on Entropy Weight of Fuzzy Information" Systems Engineering and Electronics, 2006, 28(10), pp.1523-1525

[8] Hu Fang, Huang Jian-guo, Zhang Qun-fei, "Improving Effectiveness Evaluation of Underwater Vehicle System Using Information Entropy Theory", Journal of Northwestern Polytechnical University, 2007,25(4), pp.547-550

[9] Qu Chang-wen, He You, Ma Qiang, "Threat Assessment Method Using Multi-attribute Decision-making",. Systems Engineering and Electronics, 2002,22(5), pp. 26-29

[10] Xie Yan-qing, Zhang Jiang, Guo Qiang, "Fuzzy Gray Matter Element Space Theory and Practical Application and Development—The Policy D Engineering Science", Engineering Science, 2002,4(11):, pp.57-66

[11] Xu Ze-shui., "New Method for Uncertain Multi-attribute Decision Making Problems", Journal of Systems Engineering, 2002, 17(2), pp.177-181

[12] Zhou Gang, Cheng Wei-min, "Fuzzy Grey Correlation Analysis to Appraisal of Factors Influence Personal Thermal Comfort Degree", Journal of Safety and Environment, 2005,5(4), pp.90-93

[13] Lu Da-gang, Wang Li, Zhang Peng, Wang Guang-yuan, "Grey Relation Degree Approach to Fuzzy Multiple Attribute Decision-making for Structural Scheme Design", Journal of Harbin Institute of Technology, 2007, 39(6), pp.841-844

[14] Yang Lun-biao, Gao Ying-yi, Fuzzy Mathematics, South China University of Technology Press, Guangzhou,2003

[15] Serafim O, Gwo-Hshiung T, "Defuzzification within a Multicriteria Decision Model", World Scientific,2003, 11(5), pp. 635-652

On-line Semantic Retrieval on the Multi Heterogeneous Product Information of Virtual Organization*

Meiyu Zhang[1] Chengfeng Jian[2]

Software College at Zhejiang University of Technology, Hangzhou, 310023, China
Email:1 meiyu@zjut.edu.cn; 2 jiancf@zjut.edu.cn

Abstract

The method of the identification and description of multi heterogeneous product information is put forward at first. And then the pattern mapping between STEP and OWL is presented. At last a search system is presented which uses the method of the multi path-layers faced process and can resolve the identification and representation of production information.

Keywords: Virtual organization, Information Retrieval, Semantic association, OWL, Multi Path-Layers

1 Introduction

Virtual Organization (VO) are short-term consortia or alliances of companies formed to address fast-changing opportunities. Members of a Virtual Organization carry out their tasks as if they all belonged to the same organization, under one roof, using a very powerful system to access and manage all heterogeneous information needed to support the product cycle such as STEP, SGML etc. In the Virtual Organization, it is necessary to realize the multi heterogeneous information interchange among Virtual Organization Units (VOUs). However, current web information retrieval can't meet the need to enable the integration and interoperation among VOUs. The general Web application realize information interchange is only in the local site not in multi sites; General search engine can realize in multi sites but too simple function of information interchange [1][2][3][4]. Especially both of them can’t identify the multi heterogeneous information at the same time. On the other hand, because the information among Virtual Organization Unit is dispersive and out-of-order, it is necessary to realize the function faced the process for the information retrieval.

In this paper, at first in order to realize the identification and description of multi heterogeneous information such as STEP, SGML, multimedia, a general syntax description based on the XML[5] is represented. And with XML based syntax standards, we can realize the semantic description with OWL[6][7] for the interoperability and integration of Web services by means of the mapping STEP SchemaGraph and OWL SchemaGraph. We build the semantic relationship by the two description layers: the XML template for syntax layer and the OWL template for semantic layer. And then provides the search method of multi path-layers faced the process on the basis of the uniform template of product information which is used for building the semantic association. At last search system architecture and the corresponding example are presented.

2 The Identification and Description of Multi Heterogeneous Product Information

The representation of product information has many ways. Generally it exists the two kinds standard representation for the product information. One is STEP for product data and another is SGML for product document. And besides, the Non-text mode such as

* Supported by the National Natural Science Foundation of China(No. 60603087), the Project of the Science and Technology Department of Zhejiang Province(No. 2007C31006, and No. X104631).

multimedia contains much related product information. So we should resolve the identification so as to realize the capability of searching the multi heterogeneous product information.

Figure 1 shows the process. It mainly concludes five parts:

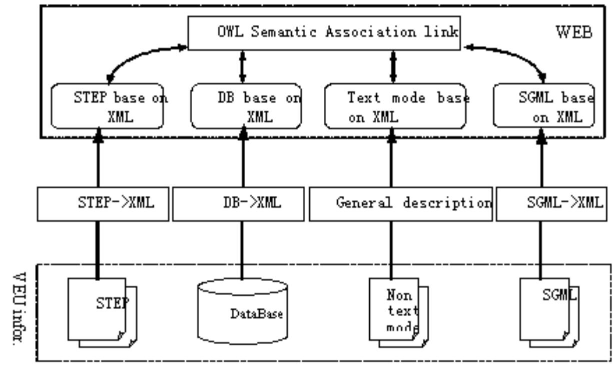


Figure 1 multi heterogeneous product information over the web

1) the identification and description of product data: STEP->XML.

As the representation of the product data, the description language of STEP standard is the EXPRESS language. So the key technology is the pattern matching between EXPRESS and XML. Because the XML has much defect in the representation of the data structures and constraints, we can formulate the rules and definitions by means of the XML extensibility. According to the rules and definitions we can realize the pattern matching[8].

2) the identification and description of database information: Database->XML.

Generally it is the standardization information in the database ,so we can extract its structure and convert into XML DTD. The question lies in the status of DTD.

3) the identification and description of non text mode information: non text mode ->text mode base on XML.

The non text mode information has much ways such as image, video and so on. It can’t be represented by the XML directly. So we should formulate the general description method by means of DTD and ensure the structure is the same with other representation of XML

4) the identification and description of product document: SGML/HTML->XML.

The SGML standard defines the formal data model of documents and other data types and data abstractions. The XML is a subclass of the SGML. So we can easily convert the SGML into XML by means of the syntax parsing.

5) Dynamic building the OWL semantic association link among the XML based heterogeneous product information.

On the above basis we can use the semantic association link to build the semantic relation among the heterogeneous product information over the web.

3 Mapping Step Schemagraph and OWL Schemagraph

Every concept from STEP SchemaGraph[9] is compared against concepts from the OWL SchemaGraph. The function listed in Table 1 calculates the match score (Cos) between a STEP SchemaGraph concept and OWL SchemaGraph.

Table 1 Mapping

FUNCTION	Mapping
INPUTS	sc, oc € W
OUTPUT	mi = (sci, ocj, Cos) where, Cos is the Match degree calculated for the mapping sci and ocj (Cos € [0,1])

The Cos(sc,oc) is composed of two different measures Element Level Match (ElementMatch) and Semantic level match (OWLMatch). ElememntMatch provides the linguistic similarity of two concepts whereas OWLMatch takes care of semantic structural similarity. The Cos(sc,oc) is calculated as the weighted average of ElementMatch and OWLMatch as shown in Equation 1.

$$\mathrm{Cos(sc,oc)}=\frac{w_{sc} * ElementMatch + w_{oc} * OWLMatch}{w_{sc} + w_{oc}}$$

$$where, (0 \le w_{sc} \le 1)(0 \le w_{oc} \le 1)$$

Equation 1. pattern mapping between STEP and OWL Weights w_{sc} and w_{oc} indicate the contribution of Element level match and Semantic level match respectively in the total match score.

4 The Search Method of Multi Path-Layers Faced the Process

4.1 The product information classification for multi VOUs

In the Virtual Organization, product is the core of multi VOUs, it is possible to building uniform contact by means of the uniform product information classification.

Product type, Production factory, Location and Product itself are the basic ways of the classification. Product type is divided into the hierarchy of product type, the son of product type and product leaf type. We use eight characters as the expression of product type according to the standard classification of national industry. It shows in Figure 2. Production factory is divided into the hierarchy of VO, VOU, the son of VO, department up to workshop. Location is divided into the hierarchy of country, zone, province and city/town. It records the distribution and operation of VOU. So far as product itself, it is divided into the hierarchy of product, components and parts. It concerned the process information of the design and manufacture.

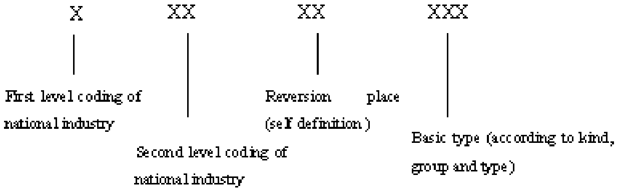

X
XX
XX
XXX
First level coding of national industry
Second level coding of national industry
Reversion place (self definition)
Basic type (according to kind, group and type)

Figure 2 the product information classification

4.2 The XML template definition of product information

In order to build the semantic relationship, we can define the uniform semantic template according to the product information classification for multi VOUs. It includes the two kinds of the template: one is the XML DTD/XML Schema template for syntax layer, the other is OWL template for semantic layer.

Each product information record is divided into two parts: product_description and product_associations. Product_description contains the inner information and product_association contains the outer related information. The ways of each product information record is divided into four parts: Product element, Product type element, Production factory element, Location element. Product element mainly contains the detail information about product such as product code, product name, technical parameter, condition, price and so on. Product type element mainly contains the recursive structure of product type such as every son of product type , the son’s flag of product type ,the code of product type and so on. Production factory element mainly contains the detail information about factory such as the factory code, the main products about the factory, the flag of VO and so on. Location element mainly contains the recursive structure of the location about the product or factory.

4.3 Multi path-layers information retrieval with multi VOUs based OWL-XML

Information extraction and combination by means of the DB-XML register repository accommodates the possibility of multi path-layers information retrieval, meantime, the uniform classification with product information accommodates to the way of multi path-layers information retrieval. Information extraction and combination takes place during the process of the information retrieval[10], the result of information retrieval is showed as the specified style. XML is the bridge with both of them.

On the basic ways of Product type, Production factory, Location and Product itself, by means of the identification of OWL, we can realize the intercross semantic search, tend to exactness step by step and recursion process in information search process. Figure 3 describes the search process.

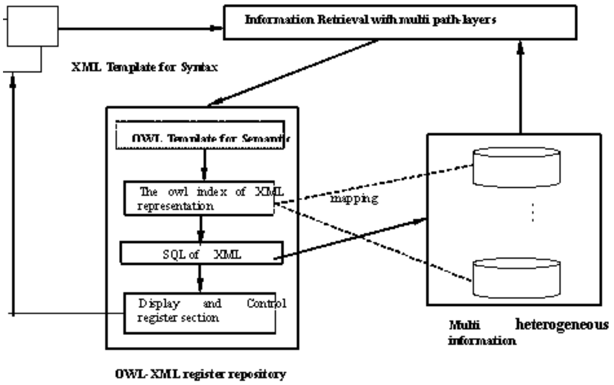

Information Retrieval with multi path-layers
XML Template for Syntax
OWL Template for Semantic
The owl index of XML representation
SQL of XML
Display and Control register section
mapping
Multi heterogeneous information
OWL-XML register repository

Figure 3 Multi path-layers search

5 The Search System Architecture

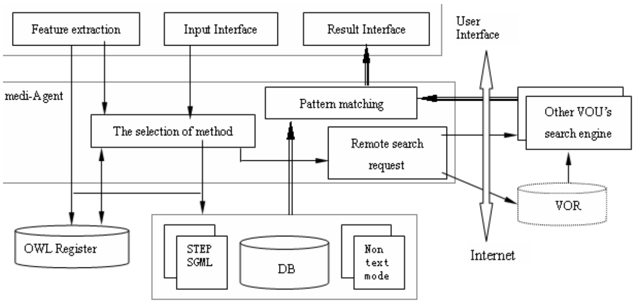


Figure 4 search system architecture

6 Conclusions

In the Virtual Organization, Product is the core of each Virtual Organization Unit. We resolve the identification and description of multi heterogeneous information, realize the semantic description with OWL for the interoperability and integration of Web services by means of the mapping STEP SchemaGraph and OWL SchemaGraph. And then build the semantic relationship by the two description layers, combine and extract the information at each Virtual Organization Unit by means of OWL-XML register repository because of its ability of data identification and description, and then realize the information retrieval by the method of the multi path-layers faced process.

References

[1] S. Agarwal, S. Handschuh, and S. Staab, "Surfing the Service Web", in Proceedings of the 2nd International Semantic Web Conference, 2003, pp. 221-226

[2] Advanced Technology Program NIST, http://www.atp.nist.gov/, 2005

[3] Eurostep-Information Solutions for a Global Age, http://62.181.193.208/, 2005

[4] WEST M, An overview of the modularization, SC4 data architecture Projects [EB/OL], http://www.nist.gov/sc4/wg-qc/wg10/current/n291/wg10n291.pdf , 2004

[5] W3C, Extensible Markup Language (XML) 1.0, http://www.w3.org/TR/1998/REC-xml-19980210, 1998

[6] W3C, OWL Web Ontology Language. http://www.w3.org/TR/2004/REC-owl-features-20040210/, 2004

[7] A. Ankolekar, M. Burstein, J. Hobbs, O. Lassila, D. Martin, D. McDermott, S. McIlraith, S. Narayanan, M. Paolucci, T. Payne, and K. Sycara, "DAML-S: Web service Description for the Semantic Web," in Proceedings of the 1st International Semantic Web Conference (ISWC 2002)

[8] Chengfeng Jian, and Jianrong Tan, description and identification of XML-based STEP production data, Journal of Computer-Aided Design & Computer Graphics, vol.13(11),2001,pp.983-990

[9] Felix Metzger, the challenge of capturing the semantics of STEP data models precisely[EB/OL], http://www.ladseb.pd.cnr.it/infor/Ontology/Baselpapers/Metzger.pdf, 2004

[10] Chengfeng Jian,Jianrong Tan, "the Uniform Information Representation For On-line Reconstruction Of Virtual Enterprise Base On XML", Proceedings of the Third International Asia-Pacific Web Conference, Xi'an, China, October 2000,pp.292-293

Research and Implementation of Data Center with High Reliability Based on VRRP

Yingguo Zhu

School of Information Engineering, Jiangnan University, Wuxi, Jiangsu 214122, China
Information Center, Wuxi Institute of Technology, Wuxi, Jiangsu 214121, China
Email: zhuyg@wxit.edu.cn

Abstract

This paper first analyzes the defects of the data center on traditional storage models, and presents a concept of data center with high reliability based on VRRP. Then, taking the data center of colleges and universities as an example, we design and reason a data center with load balancing and high reliability based on VRRP. And discuss its characteristics from the following three areas: load balancing, port redundancy, and chains redundancy. At last，we draw a conclusion that the high reliability of the data center can be affected effectively by VRRP technology, ports redundancy and chains redundancy technology.

Keywords: VRRP, Data Center, High Reliability, Fault-Tolerant

1 The Concept of Data Center

Modern information technology is the key to improve the competitiveness and to promote economic growth. It plays an important role in the improvement of labor efficiency, and improves the quality of economic operation, and the changes and development in the industrial structure. At the same time, the servers of network center and the volumes of data rapidly increase, which results in the growing complexity in the management.

In these applications, it is the data running in the background database server that plays a crucial and central role. How to manage these centralized data, provide a stable and reliable access services, and make a real-time backup in order to improve the reliability of data resources effectively becomes particularly important. Therefore, the data center comes into being.

Data center usually refers to a system which can realize the centralized information processing, storage management, exchange and transmission in a limited physical space of the data, thus facilitating all kinds of centralized data management. The data center system can provide a better network and operation environment and assume all the tasks for the operation and maintenance of a variety of applications systems, such as enrollment and employment of students, financial system, e-card, e-library, teaching, research and management. Almost all the critical data of these applications are stored in the data center. At the same time, there is a frequent exchange of network data between the data center and the various business units [1].

2 Demands on the Reliability of the Data Center and the Defect of Traditional Storage Mode

Construction of the data center is focused on ensuring the high reliability of business-critical data and critical applications. It includes the system's high reliability technology and the high reliability data. The system reliability is divided into two types of hardware system and software system. There are various redundancy technology, online backup technologies, cluster technology, etc. The data reliability includes RAID, real-time data backup, etc.

Under the traditional model, all types of servers are independent of each other. All data is stored in the server's local hard drives, and distributed in various departments. Their network topology structure is shown in Figure 1.

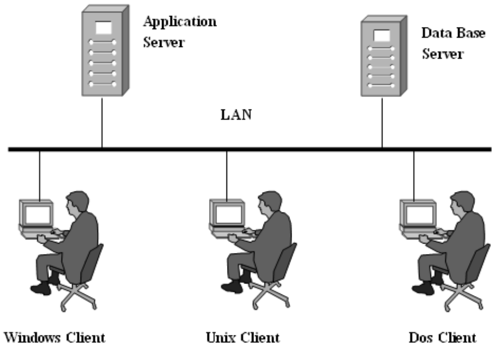


Figure 1 Traditional topological structure

Usually there are two ways to visit inter-subnet server:

The first method is the use of static routing protocol by default gateway. The drawback of this method is that the data cannot be centrally managed. This brings a threat to the data security and reliability. It will result in interrupt transmission if a server had a single point of failure. Therefore the stability cannot be guaranteed. It is not in a position to play the best performance if the data reading and writing speed became slower etc.

The second method is to use dynamic routing protocols such as RIP, OSPF, etc. This protocol has its own shortcomings in many places. For example, RIP protocol cannot be used for large-scale networks, and it converges slowly, and it cannot choose the best path. So, it cannot be used for such VLSM. OSPF is a kind of link state routing protocol. And it can be used in large-scale network. It provides good support for VLSM, with a faster convergence rate, as well as reduces routing information quickly. But OSPF protocol has a routing table calculation by the network topology database. This occupies a large storage space of the router and spends a lot of CPU resources at the same time [2].

For high reliability data center, the above two methods are not competent obviously.

3 VRRP Technology

VRRP technology overview

Virtual router redundancy protocol, or VRRP, RFC2338, is a protocol which focuses on the static allocation of the default gateway on the third layer switches or routers, provides faster and more effective redundant fault-tolerant capabilities for clients that rely on the default gateway for cross-network access. It effectively solves the inadequacy of static and dynamic routing protocol under traditional model and makes it possible for the high-reliability redundant network design [3].

Principle of VRRP

The principle of VRRP is shown in Figure 2. The VRRP group is composed of two Layer 3 Switches or routers. This virtual router has its own IP address, 210.28.149.100, and MAC address. At the same time physical router also has its own IP address 210.28.149.1 and 210.28.149.2. The default gateway address is set to be 210.28.149.100. In the client, the terminal console only knows the virtual router's IP address, without having to know the true router A and router B IP address. In the working group, clear division of work between the routers is available. One of them is regarded as the main router (MASTER), which is responsible for the router forwarding data packets, and another as backup router (BACKUP) [4].

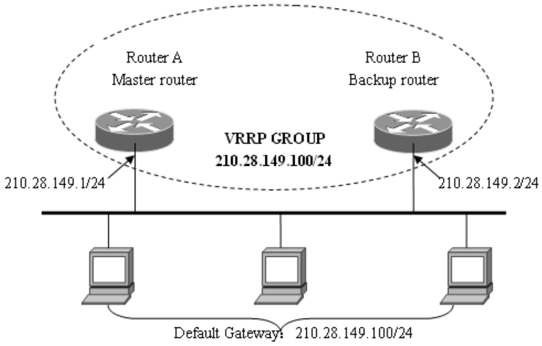


Figure 2 VRRP Principle

Under normal circumstances, router A and router B regularly exchange hello message through VRRP to determine their identity. Once the Main router downtime occurs, the backup router can not receive HELLO packet from the main router within the required time frame, the backup router will automatically change to take over the master router’s job. In this way, the dynamic redundant backup to the core router is realized [5]. Thereby it effectively improves the network reliability.

4 Data Center Design Example

The characteristics of the data center of college and university

The colleges and universities is the forefront of teaching and scientific research, where profound changes have also taken place. For example, the rapid increase in network center servers and volume of data, which results in the growing complexity in the management. Applications, such as distant learning network, intake and employment of students, e-card, applications of multimedia, data trace and search, virtual labs, have become an important link to improve the efficiency of teaching and scientific research work. Integrating with teaching, service and research, it has brought them far-reaching impact.

Principles of design

According to the fact of data center at college and university, it requires the following characteristics: a large-capacity storage space, good expandability, faster access speed, easy manageability, low cost for maintenance, etc. The following basic principles mast be complied with when we design the data center:

Practicability: Data center should be designed to meet the college and university's demand to computer network applications, at the same time it can fully realized the requirements of information and networking during the daily management, the teaching and the scientific research work. So that the overall performance of the data center can be given fully play.

Reliability: Because the network's structure of data center is very complex and the data center has a high technical, so it must be ensured that the system has a high MTBF (mean time between failures) and a low MTTF (mean time to failure) to ensure the security and stability operation with high fault-tolerant performance and no single point of failure for a long time.

Security: Meeting the reliability, data center can withstand the attacks from internal and external. The security measures adopted should be effective and credible, and be able to achieve the target of control safely from multi-levels and multi means.

Forward-looking： Data center of the college and university not only can meet the current demand for network, but also can be easily expanded in the future when necessary. So the investment in the current is well protected. At the same time the configuration should be designed for flexibility in order to meet the other requirements of colleges and universities.

Topology design

Taking advantage of the high reliability of VRRP, and according to the reliability and scalability principle, we can design a data center topology as shown in Figure 3[6]. The convergence layer switches are connected to the core router R1 and R2 and provide a server with the only default gateway through running VRRP agreement on R1 and R2 routers. When any router fails, through VRRP agreement, another router immediately takes over all the work. At the same time, this router also updates the routing table, and notifies other routers through dynamic routing protocol to update their corresponding routing table.

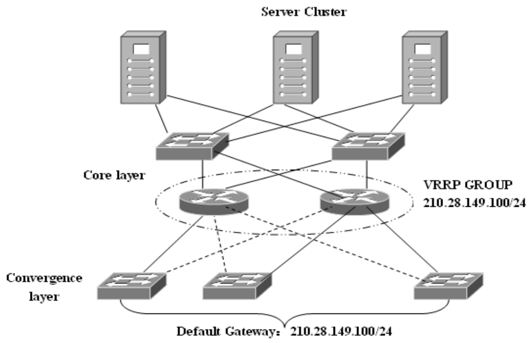


Figure 3 Data Center Topology

Design notes

During the design process, we take full account of reliability and economy, and make it successful in the existing mainstream equipment configuration [7]. This program has the following characteristics.

Load Balancing: R1 and R2 can be formed as a VRRP working group through mutual backup, setting up R1 to be R2's backup router, R2 to be R1's backup router. And this may serve the purpose of load balancing which is conducive to the stability of the core layer equipment.

Ports redundancy: Convergence layer switches are connected to the two core routers through two different links. This ensures the port redundancy. And it can protect the normal data communications by opening spanning tree protocol in the center switches and access switch [8]. At the same time, each server is connected to different switches with two NICs. When one switch fails, the server will automatically switch to the standby card, thereby it is connected to another switch [9].

Thus, port redundancy, from the gathering level switchboard to the core switchboard and the server, has been realized. And high reliability is ensured at ports [9].

Chains redundancy: It includes the following three aspects.

1. Routing chains between the two center routers

As the two center routers support port aggregation function, therefore, a number of routers link can be established between the two routers [10]. Many ports are polymerized to be a logical port by link aggregation. For these ports support port traffic automatic balanced protection, so all the physical channels flow balance.

When one or some polymerized ports of the physical link fails, the flow can be automatically transferred to other chains boost. After the port is resumed, the flow can automatically be redistributed.

2. Chains from the convergence layer switches to the central router

According to the VRRP hypothesized router condition, when the links from the convergence layer switches to the center routers break down, it is possible that the router will transform from one role to another role. In order to enhance the reliability of the link, port polymerization can be used to connect the convergence layer switches and the center routers [9]. Therefore, there is redundant reliability in the links between convergence switches and central routers.

3. Chains from the server to switch

Each server is connected to different switches with dual NIC. Thus it was ensured that dual-link existed between the servers and switches. The server will start backup link immediately and transfer the traffic to the backup NIC when the master NIC or the link from master to switch fails. Thus the chain from server to switch is high reliability.

5 Conclusion

Data center is the core of the whole network. And its reliability decides the entire network reliability. The high reliability of the data center can be affected by VRRP technology, port redundancy and link redundancy technology.

References

[1] Haiyan Wu, Li Qi, Liqiang Shen, and Dongxing Jiang， “Construction and Research of Operating Environment for Campus Data Center with High Availability”， Education Information of China, May 2007, pp 16-18

[2] Rita Puzmanova, Routing and Switching, BeiJing, Posts & Telecom Press, April 2004

[3] www.ietf.org, RFC2338, VRRP, Virtual Router Redundancy Protocol

[4] Shuyou Zhang, “the Principle of VRRP Protocol and the Application on Routers”, Huawei - 3Com Soliciting Articles, Financial Computer of China, No. 10, 2006, pp 28-30

[5] Shouyu Zhang, and Yongqun Wang, “Simple Discussion about the Cluster Technology of Virtual Routing on Gigabit Switching Router.”, http://networ k.ccidnet.com/art/216/20020723/19973_1.html, 2002.07.23

[6] Weiqiang Huang, and Kexun Meng, “Application of VRRP Routing Protocol”, Journal of South China Normal University (Natural Science Edition), No. 4, 2004, Nov. 2004, pp 53-58

[7] Yingguo Zhu, “Construction of Schoolyard Backbone Network for Tolerance”, Journal of Wuxi Institute of Technology, Editorial Dept of WXIT, Vol.2, No.4, Dec, 2003, pp.11-12

[8] H3C Technologies Co., Ltd. H3C S9500 Series Routing Switch Manual, Release1631 V1.21

[9] Juan Jia, Binqiang Wang, and Shuai Yang, “Research and Implementation of Mathod about a kind of Core Router with High Availability Based on VRRP”, Communication and Network, Application of Electronic Technique, No. 2, 2007, pp 110-112

[10] J.L. Ge， J.Z. Wu, and D. Jeff， TCP/IP Routing Technology, BeiJing, Posts & Telecom Press, Oct. 2003

Adding Synonym Query for Chinese Language to Lucene Search Engine

Xu Zhao Wenbo Xu Zhilei Chai

College of Information, Jiangnan University, Wuxi, 214122

Email: aslanzala@126.com

Abstract

While the Internet is developing rapidly, there are more and more information on it. Many of these information have the same meaning. When People Looking for something using search engine, at most time, they do not know there are other meanings of the keywords which they input. To cause the information which may be actually they need, can not be found. For example, “雷克萨斯” has other names “Lexus” or “凌志”. We always hope more useful result, but input as few words as possible. This puts forward claim on search engine to support synonym query. In this paper we make a deep study of the words segmentation in Lucene, put forward the principle of indexing Chinese Synonym, implement a kind of forwards maximum match algorithm. And on this basis, I implement a Lucene analyzer which can deal with both Chinese and English Language.

Keywords：Lucene, Search engine, Words segmentation, Synonyms, Position increment, Token

1 Introduction

Lucene is an open source software develop class library granted by Apache Software Foundation. And it is one of the most popular API for developing search engine. Lucene do not support synonym query of its own, many scholar and developer in the world have been actively seeking the relationships among synonyms. And essay to add synonym query to Lucene. The Lucene extension tool WordNet is a successful solution. But we regret to say there is nothing WordNet can do to deal with Chinese language, though it done well with English words. Because of the difference in structure between Chinese and English words, for example, English words can split by WhiteSpace, but Chinese sentence consists of a sequence of single character where exist fixed relationship semantically with no separator, and there is also difference in specification for encoding between Chinese and English, search engine need a distinct algorithm to analyse Chinese language. But up to now, the analyzer in Lucene can only divide sentence into single Chinese character or double characters. These algorithms are not satisfactory. That is the reason why there is no synonym query for Chinese in Lucene. We analyse the mechanism of words segmentation and indexing in Lucene, then provide an Chinese words segmentation algorithm based on Chinese synonym dictionary, and implement the solution of Chinese synonym words segmentation.

2 Analyzing the Lucene

2.1 Architecture

Lucene is not a completed search engine application, but it supply many API and smart storage organization. These API are wrapped in 7 namespaces:

（1）Lucene.Net.search: wrap behaviors of searching, supply interface of constructing IndexSearcher and Query object.

（2）Lucene.Net.index: interface of indexing input text to file system.

（3）Lucene.Net.queryParser: be used chiefly in analyzing query expression, and transform the expression to Query object.

（4）Lucene.Net.analysis: supply several inline analyzer to deal with words segmentation problem.

（5）Lucene.Net.document: supply the structure of input text in memory which are going to be index to file system.

（6）Lucene.Net.store: supply API of I/O operation in base layer.

（7）Lucene.Net.util: some public data structure and tool.

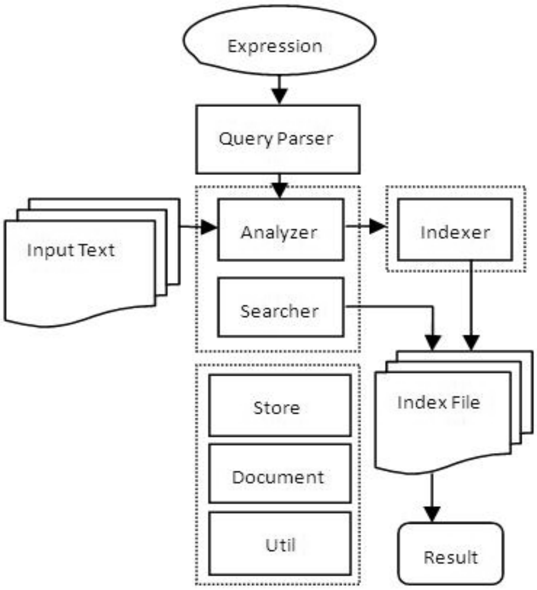


Figure 1 Index and Search Process in Lucerne

2.2 Words Segmentation Mechanism

We can find that analyzer is an important component which is used in both indexing process and searching process in Figure 1. It receive an input stream which is going to be indexed as a parameter, divide it into words(words segmentation) and filtrate the stop words, then output a TokenStream type. TokenStream is a Lucene type that consist of a stream of Token. A Token carries with it a text value(the word itself) as well as some meta-data: the start and end offsets in the original text, a token type, and a position increment.

Figure 2 shows the details of analyzing the text “What is your name”. Tokenizer divide the whole text into single words. TokenFilter filtrate the stop word “is” and transform all characters to lowercase. As text is tokenized, the position relative to the previous Token is recorded as the position increment value. All the built-in tokenizers leave the position increment at a default value 1, indicating that all tokens are in successive positions, one after another. The start offset is the character position in the original text where the Token text begins, and the end offset is the position just after the last character of the token text.

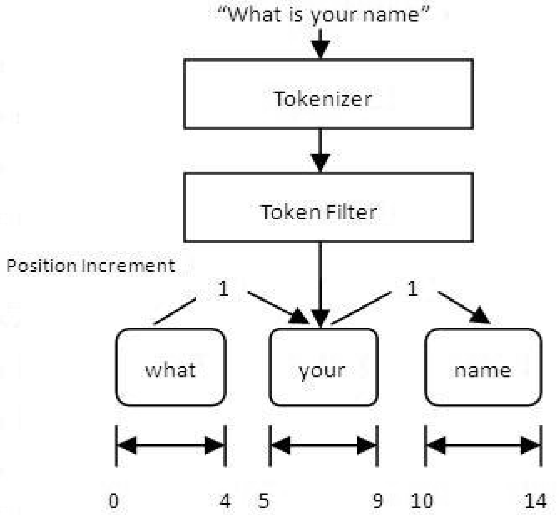


Figure 2 Analysis Process in Lucene

The indexer reorganize the stream of tokens based on inverse indexing algorithm, transform them to Terms and write to file system. That is what we called Lucene storage structure. We must use the same analyzer during searching process as indexing process so that the query expression will match to the terms in index files.

2.3 Indexing Synonyms Principle

Start and end offset as well as token type are discarded when each token is posted to the index as a Term. As the position increment is the only additional meta-data associated with the token carried through to the index, we should attach great importance to this variable. It relates current token to the previous one. A token with a zero position increment places the token in the same position as the previous one. If we add a new token with the text value “familyname” to the stream just after the word “name” and set position increment value 0 in Fig 2. Both “name” and “familyname” will be indexed to the same position just after the word “your”.

Searcher will not distinguish “name” and “familyname” semantically, but conclude that they are the same term by zero position increment. Whether we input a expression with key words “your name” or “your familyname”, the searcher will hit the target of “What is your name” absolutely.

But Lucene is quite limited in analyzing Chinese language that it only divide a sentence to stream of single character or double characters mechanically. For example, we will divide “我们是中国人” to “我”—“们”—“是”—“中”—“国”—“人” using Standard Analyzer in Lucene. In these circumstances, it does not make sense adding synonym for each character. So we need a more powerful analyzer.

3 Forwards Maximum Match Algorithm and Chinese Synonym Dictionary

Words segmentation algorithm based on Dictionary need a words table which used to match the candidate text. So called “forwards maximum match” means pick the word in table as long as possible to match the candidate sentence from left to right. I describe an implement of the algorithm like that:

“S” represent the sentence we operate. “S[n]” represent every character in “S”.

（1）Define a “buffer” to store the string that is going to be separate away from “S” and an “offset” to indicate the current position apart from the starting and a variable “length” to indicate string length in buffer.

（2）Traverse “S” from left to right, read a character and put it into buffer combing with previous string.

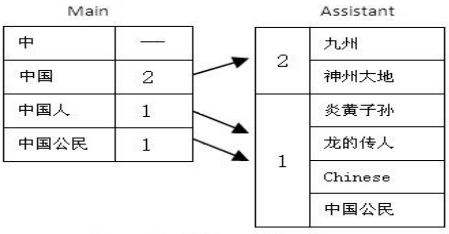
（3）Then matching the string with the words in dictionary, if hit a target, offset++ and length++. Turn back to（2）.

（4）Else, judge whether there is only one character in buffer. If the answer is yes, means the dictionary do not contain this character. Then we segment the character out of buffer. Clear the buffer, set length value 0 and add offset 1. Turn back to（2）.

（5）If the buffer contains more than one character, means “buffer[0..length]” is a maximum matched word (Notice that there are “length+1” characters in buffer, we slough the last one). Then we segment “buffer[0..length]” out of buffer. Clear the buffer, set length value 0 but do not change offset. Turn back to（2）.

（6）The work is done when gone over “S”.

The structure of dictionary make a great impact on system consumption and matching efficiency when load it to memory. I designed the dictionary with two tables. Both of them are HashTable. The main table contains all the words used in matching process with a Key field storing words themselves and a Value field storing an needle to an unit in assistant table. The assistant table contains synonym with a Key field storing an id of this (Key-Vale)unit and a Value field storing the synonyms.



Main

中	—
中国	2
中国人	1
中国公民	1

Assistant

2	九州
	神州大地
1	炎黄子孙
	龙的传人
	Chinese
	中国公民

Figure 3 Dictionary Structure

4 Implementation of Chinese Synonyms Analyzer

Based on above research, I defined four primary Class to implement the Chinese synonyms analysis. They are “MyCjkSynonymAnalyzer”, “MyCjkSynonymTokenFilter”, “MyCjkSynonymTokenizer” and “SynonymsDic”.

4.1 “MyCjkSynonymAnalyzer” Class

Every user-defined Analyzer must inherit the interface “Analyzer” in Lucene.Net.analysis namespace and implement its TokenStream() method at least. This method deal with the input text and output a Token stream segmented.

```
public class MyCjkSynonymAnalyzer:Analyzer
{
public override TokenStream TokenStream(System.String fieldName, System.IO.TextReader reader)
```

```
    {
    TokenStream result = new MyCjkSynonymFilter
(new MyCjkSynonymTokenizer(reader));
    return result;
        }
    }
```

TokenStream() method call a “MyCjkSynonym Tokenizer” type object to do words segmentation process and a “MyCJKSynonymTokenFilter” type object to select every synonym and add them to output Token stream.

4.2 “MyCjkSynonymTokenFilter” Class

Every user-defined TokenFilter must inherit the interface “TokenFilter” and implement its Next() method. It return a Token with a word apart from input text when called once. In order to select synonyms and warp them in a Token output by Next() method, I use stack structure.

```
    public class MyCjkSynonymFilter:TokenFilter
    {
    public static string TOKEN_TYPE =
"MyCjkSynonym";
    private Stack<Token> Stack;
    public MyCjkSynonymFilter(TokenStream instream):
base(instream)
    {
            Stack = new Stack<Token>();
        }
    public override Token Next()
    {
    if (Stack.Count > 0)
            {
                return (Token)Stack.Pop();
            }
    Token token = input.Next();
                if (token == null)
            {
                return null;
            }
              if (token.Type() == ChineseType)
            {
                PushSynonymWordsToStack(token);
                }
            return token;
        }
        private void PushSynonymWordsToStack (Token
token)
        {
    string[] synonyms = getCurrentSynonymWords(token.
TermText());
                if (synonyms == null) return;
            for (int i = 0; i < synonyms.Length; i++)
            {
    Token syntoken = new Token(synonyms[i], token.
StartOffset(), token.EndOffset(),TOKEN_TYPE);

    syntoken.SetPositionIncrement(0);
                synonymStack.Push(syntoken);
            }
        }
    private string[] getCurrentSynonymWords(string
word)
        {
    int id = SynonymsDic.MainWordsTable[word];
    string[] synonyms = SynonymsDic. Synonyms
Table[id];
        }
    }
```

Description of my algorithm:

（1）If the stack is not empty, pop the top Token and return it.

（2）Else read next Token from input stream.

（3）If the Token is null, return null.

（4）If the Token type is Chinese Type, select all the synonyms of this Token`s text value and push them to stack one by one.

（5）Return this Token.

The detail of step（4）is:

（1）Get all the synonyms of current word from assistant table I mentioned above.(“SynonymsDic” class implement the structure and supply all operations on dictionary.)

（2）Wrap them in Tokens, use the same start offset and end offset as current word then set zero position increment.

（3）Push the Tokens to stack.

4.3 “MyCjkSynonymTokenizer” Class

“MyCjkSynonymTokenizer” distinguish Chinese character and English character automatically. This Tokenizer segment English words by user-defined separator(Default separator is WhiteSpace). To segment Chinese words, it use the algorithm I present in section 3.

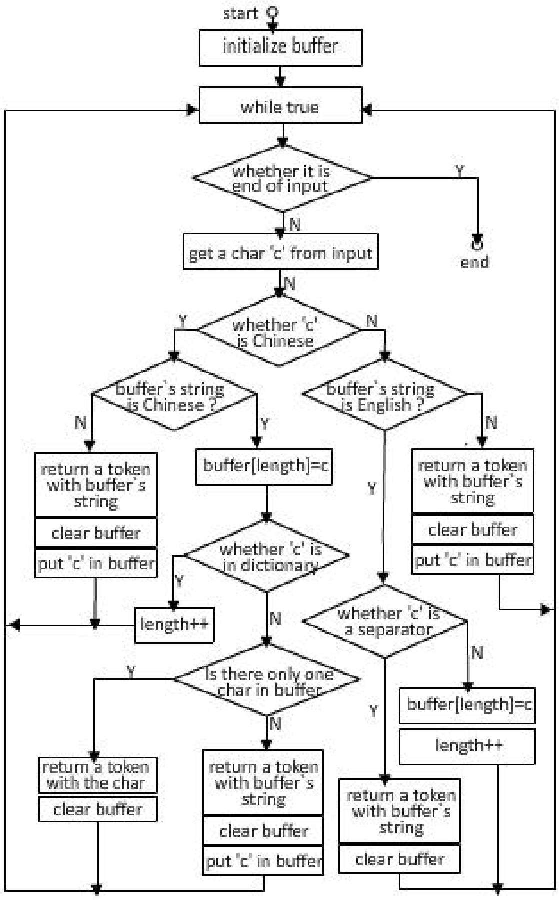
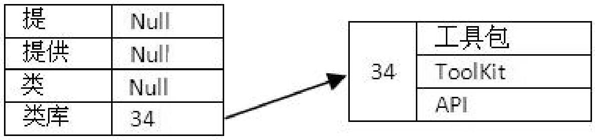

start
initialize buffer
while true
whether it is end of input
Y
N
end
get a char 'c' from input
N
whether 'c' is Chinese
Y
N
buffer's string is Chinese ?
buffer's string is English ?
N
Y
N
return a token with buffer's string
clear buffer
put 'c' in buffer
buffer[length]=c
Y
return a token with buffer's string
clear buffer
put 'c' in buffer
whether 'c' is in dictionary
Y
length++
N
whether 'c' is a separator
N
Y
Is there only one char in buffer
N
buffer[length]=c
length++
Y
return a token with the char
clear buffer
return a token with buffer's string
clear buffer
put 'c' in buffer
return a token with buffer's string
clear buffer

Figure 4 Words Segmentation Process Flow Chart

```
public override Token Next()
{
......
    int start_offset = offset;
    Token result = null;
    while (true)
    {
        if (IsEnd(instream)) break;
        char currentchar = getNextChar();
        offset++;
            ......
    // Chinese segmentation process
            if (IsChineseType(currentchar))
            {
                if (/*characters in buffer in not Chinese*/)
                {
    result=new Token(new String(buffer), start_offset, start_offset + length, "EnglishType");
                    Clear(buffer);
                     start_offset += length;
                     length = 0;
                      buffer[length] = currentchar;
                     length++;
                      return result;
                    }
    buffer[length] = currentchar; if(SynonymsDic.Main WordsTable.ContainsKey(new String(buffer)))
                     {
                     length++;
                     }
                     else
                     {
                     ......
                        if (length == 0)
                        {
    result = new Token(new String(buffer), start_offset, start_offset + 1, "ChineseType");
                            Clear(buffer);
                            start_offset += 1;
                            return result;
                        }
                        else     //length>0
    {
    result = new Token(new String(buffer,0,length), start_offset, start_offset + length, "ChineseType");
                         Clear(buffer);
                         start_offset += length;
                         length = 0;
                         buffer[length] = currentchar;
                            length++;
                        return result;
    }
    }
    }
            else    // English segmentation process
            {
                if (/*characters in buffer is Chinese*/)
                {
    result=new Token(new String(buffer), start_offset, start_offset + length, "ChineseType");
                        Clear(buffer);
                    start_offset += length;
                        length = 0;
                        buffer[length] = currentchar;
                        length++;
                        return result;
                }
                    if (IsSplitChar(currentchar))
                    {
    result = new Token(new String(buffer), start_offset, start_offset + length, "EnglishType");
                        Clear(buffer);
                        start_offset += length;
                        length = 0;
                        return result;
                    }
                    else
                    {
                    buffer[length] = currentchar;
                        length++;
                    }
                    ……
    }
    }
    }
```

5 Experiment

I made a query experiment to test what we done to Lucene. First, I use “MyCjkSynonymAnalyzer” to analyse the text of “Apache Software Foundation 提供的类库”. Then load dictionary with content like this:



提	Null
提供	Null
类	Null
类库	34

34	工具包 ToolKit API

Figure 5 Words in Loaded Dictionary

Construct a Query object use TermQuery like this:

Query q = new TermQuery(new Term("content", "工具包"));

The result shows that it can hit the target of this sentence when search the Query q. Then I Construct a new Query object use PhraseQuery, the searcher can still find the sentence.

PhraseQuery q = new PhraseQuery();

q.Add(new Term("content", "提供"));

q.Add(new Term("content", "的"));

q.Add(new Term("content", "API"));

We can see the details of each token output by “MyCjkSynonymAnalyzer” in Table 1.

Table1 Tokens Information Transact by MyCjkSynonymAnalyzer

text	StartOffset	EndOffset	Position
Apache	0	6	1
Software	7	15	2
Foundation	16	26	3
提供	26	28	4
的	28	29	5
类库	29	31	6
工具包	29	31	6
ToolKit	29	31	6
API	29	31	6

6 Conclusion

We can conclude by the experiment result that this analyzer can support Chinese synonym query and deal with fixed input stream. There is no limitation on length or linguistic. The synonym of Chinese words “工具包” can be English words “Toolkit” and its length is 7. To some degree, it is also a solution of cross languages query in search engine.

References

[1] Li Qinghu, Chen Yujian, Sun Jiaguang, A new dictionary

mechanism for Chinese word segmentation[J], Journal of Chinese Information Processing, 2002,17（4）: 13 - 18

[2] Xiang Hui, Guo Yiping, Wang Liang, Design and implementation of Chinese words dictionary segmentation module based on Lucene[J], New Technology of Library and Information Service, 2006（8）: 46

[3] Chen Yanchun, Li Shuangping, Design and implementation of enterprise search engine based on Lucene[J], New Technology of Library and Information Service, 2007（8）: 63 – 65

[4] Wang Liyun, Wang Hua, Chen Gang, Yao Naiming, Design and implementation of full text search engine based on Lucene [J], Computer Engineering and Design, 2007, 28 (24) : 5959- 5961

[5] Qiu Zhe, Fu Taotao, Develop yurselves search engine, Beijing: Posts & Telecom Press, 2007

[6] Song Jia, Zhu Yunqiang, Liu Runda, Enhanced full text retrieval kit based on Lucene[J], Computer Engineering and Applications, 2008, 44 (4) : 172- 175

[7] JavaCC[EB/OL], https://javacc.dev.java.net/, 2003 - 04

[8] Hatcher E, Gospodnetic O, Lucene in Action [EB/OL], [2005],http: //lucenebook.com/

[9] Lang Xiaowei, Wang Shenkang, Research and development of full text search engine based on Lucene[J], Computer Engineering, 2006, 32（4）: 94 - 96

[10] The Apache Jakarta Project :Lucene[EB/OL], http:// jakarta. apache.org/lucene/ ,2003 - 04

Improved Apriori Algorithm Based on Weighted Mining Association Rules

Yuanyuan Zhao He Jiang Baoyou Sun Xiangjun Dong

School of Information Science and Technology, Shandong Institute of Light Industry, Jinan, 250353, China

Email: xiaofengye1013@163.com, jianghe@sdili.edu.cn, tianyijiujiu@163.com, d-xj@163.com

Abstract

This paper discusses algorithms based on the Algorithm Apriori and discovers their existing questions. An improved weighted association rules (Algorithm W_Apriori) is proposed that makes the variations at the prune step of generating frequent items and generating association rule. As a result, more frequent items are discovered and the time of generating association rule is reduced. Experiments demonstrate the algorithm is feasible and efficient.

Keywords：Association Rule；Weight；Frequent Itemset；Algorithm Apriori； W_Apriori

1 Introduction

Association rules mining are extracting useful and potential rules information hidden in the massive, incomplete, noisy, fuzzy and random database[1]. Association rules mining is firstly proposed by Agrawal et al. It is used to find relationships between commodities that the customs purchase in the supermarkets in database transactions. That can provide much useful information for decision-making.

Traditional Algorithm Apriori treats each item as uniformity. But in the real world, the importance of each item is often different. For example, if some goods have great profits or some goods are on sale-promotion, decision-makers tend to be more concerned about the sales of goods. Traditional algorithm of mining association rules can not distinct from the others in importance and interestingness of these commodities. In order to solve the above problems, the concept of weight is added to association rules. When every item is associated with weight, Apriori closure property is no longer in point, so a new prune method is required. In [3], the concept of weighted boolean association rules is proposed and two algorithms of mining weighted association rules mining are designed. Improved algorithms for mining weighted association rules are proposed in [4][5][6][7][8]. In [4], a algorithm New_Apriori is proposed, that is similar with the traditional algorithm Apriori. It employs an approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets. A set of candidate k-itemsets is generated by joining frequent (k-1)-itemsets with itself. In the Prune step, the items, whose horizontal weighted support is lower than the minimum weighted support, is pruned. Many interesting frequent itemsets are possibly lost if candidate itemset that is infrequent current but can generate the frequent superset are deleted. Algorithm W_Apriori is proposed to solve the corresponding questions in the paper. The algorithm makes some variation in the pruning step to mine more frequent itemsets and does some prunes when association rules are generated to shorten the time of generating the rules.

2 Algorithm Descriptions for Weighted Association Rules

According to [9], Weighted Association Rules is as follows: there are n records and m items in the database

DB. Let $I=\{i_1, i_2,\dots, i_m\}$ be a set of items, each item is associated with weight. Their weights mapped by the function (^{x}) are $\{w_1, w_2,\dots, w_m\}$ ($w_i \in [0,1]$. Suppose the minimum weighted support is *wminsup* and the weighted minimum confidence is *wminconf.*

There are many methods about the support of weighted association rules $X \Rightarrow Y$. For example, [9] proposes an idea about the average weighted support. This idea smoothes the impact of items we concerned about. In [10] the weighted support is calculated by the sum of the weight which makes some values may be greater than 1 and the value is not suited to our logic habit. So the following formulates is selected to calculate in this paper:

$wsup(X\cup Y)= max\{ w_1, w_2,\dots, w_p \} \bullet sup(X\cup Y)$,

where $sup(X\cup Y)$ is the support of $X\Rightarrow Y$, $max\{ ^{U} , ^{U} ,\dots, q_k = \min\{p_{1k}, p_{2k}, p_{3k}, p_{4k}\} \}$ is the max weights of $X\cup Y$ that is called the weights of $X\cup Y$, signed as *weight* ($X\cup Y$) , *p* is the size of ($X\cup Y$) that the number of items itemsets ($X\cup Y$) contains[4].

If the weighted support wsup$(X\cup Y)$ *of the itemset* $X\cup Y$ *is larger than* wminsup*, the itemsets is frequent.*

The weighted confidence of association rule $X\Rightarrow Y$ is

$wconf(X\Rightarrow Y) = wsup(X\Rightarrow Y)/ wsup(X)$

3 The Improved Algorithm for Miningweighted Association Rules Based on Algorithm Apriori

The Problems Existing in The Algorithm

The problems caused by the weights：After the items are associated with weights, the property of traditional algorithm Apriori is no longer in point. So a new prune method is required.

Suppose itemset *X* is the weighed infrequent itemset and itemset *Y* contains itemsets. If *weight*(*Y*)>*wminsup*, then *Y* is potential weighted frequent itemset. For example, if *weight*(I_1)=0.2，*weight* (I_2)=0.2，*weight*(I3)=0.8, *wminsup*=0.05,*sup*（{ I_1，I_2}）=0.1, then the *wsup*({ I_1, I_2}）=0.02 <*wminsup* is weighted infrequent itemsets. Suppose there is an itemset {I_1, I_2, I_3} and its support *sup*（{ I_1, I_2, I_3}）=0.1, then its weight *weight*（{ I_1, I_2, I_3}）=0.8，which is larger than *wminsup*, so the itemset { I_1, I_2, I_3}(^{U} { I_1, I_2}) is the weighted frequent itemset. Based on the above conclusions, in the algorithm of mining weighted association rules, we use the following method to prune:

If the weighted support of *X* is lower than the minimum weighted support *wminsup* and its weight is larger than *wminsup,* then this itemset is pruned, or else is added to $S_{k.}$

The problem existing in generating rules：Generating rules is done by making use of the mined frequent itemset. The most studies of association rules are about improving the effect of frequent itemset being produced and enough many frequent itemset. But it is important for the users to generate rules also. It provides much useful information for decision-making. So it is necessary to improve the efficiency of generate rules.

All frequent itemsets *L* are generated. For each frequent itemset *l,* generate its all association rules. Its process is:（1）generate all the nonempty subsets subl of *l*;（2）for every nonempty subset *s*, if *conf*（$s\Rightarrow l\text{-}s$）is larger than *wmincof*, then output rule "$s\Rightarrow l\text{-}s$", where *wminconf* is the minimum confidence threshold. If so, all the subsets of the frequent itemsets *l* are scanned one by one, then the efficiency of algorithm is reduced. Moreover, Apriori closure property is no longer in point in the weighted association rules, so the subsets are not always frequent.

The prune is done to reduce the unnecessary calculations when the rules are generated. It can not be judged by priori knowledge because every item has a weight. Actually, we can sort by the weighted support of premises in the rules. If the rule whose confidence is lower than the minimum weighted confidence threshold *wminconf* appears, then all rules whose weighted supports are larger than support of the former are unnecessary to be judged. It doesn't need calculating and judging again.

Theorem 1 Frequent itemset *l* is known, for any subset *a*,*b*,if *wsup*(*a*)<*wsup*(*b*) and $a \not\Rightarrow l\text{-}a$, then $b \not\Rightarrow l\text{-}b$.

Proof $wconf\,(a\Rightarrow l\text{-}a)=wsup(l)/wsup(a)$

$wconf(b\Rightarrow l\text{-}b)=wsup(l)/wsup(b)$

$wsup(a)<wsup(b) \quad a \not\Rightarrow l\text{-}a$

$\therefore wconf(b\Rightarrow l\text{-}b)<wconf(a\Rightarrow l\text{-}a) < wminconf$

$\therefore b\not\Rightarrow l\text{-}b$ Q.E.D

The improved algorithm of generating rules is as follows:（1）generate all the nonempty subsets *subl* of *l* and sort them by their ascending weight support;（2）for any subset *s*, if *conf*（*s*⇒*l-s*）>*wminsup*, output “*s*⇒(*l-s*)” according to Theorem 1. When the first subset that can’t meet minimum weighted confidence threshold appears, the process where the rules are generated is over.

Algorithm Formulation

The pseudo-code of algorithms W_Apriori is as follows:

```
C1 . GeneratẹṬ;        //scan database to search all the items and their counts
L1  c∈  C1|  wsup(c)≥wminsup  ;  //produce frequent-1 itemset
S1   L1∪ c∈ C1| wsupç   wminsup.  wsupC1-ç≥ wminsup
k   1;
whilẹ|Sk|   0̣
ḳ;
Ck=Join(Sk-1);
Sk c∈ Ck | wsupç≥ wminsup wsupC1-ç≥ wminsup ;
Lk c∈ Sk wsupç≥ wminsup;
L ∪Lk;
R Generate_rule(L);
```

The *Joiṇ* $S_{k\text{-}1}$ procedure generates the candidate C_k by joining $S_{k\text{-}1}$ with itself. In this algorithm, prune is done for two times. In the first prune, the impossible frequent items whose weights are larger than the minimum weighted support but the weighted support is less than the minimum support are deleted for the join step. The second prune produces the frequent itemsets. The aim to prune for two times is to keep down the potential frequent items for mining the more frequent itemsets because in weighted association rules, the supersets of infrequent itemsets are perhaps frequent.

The Generate_rulẹḶ procedure is to generate association rules, where it needs to check if L is frequent or not and prune to reduce unnecessary judge according to theorem 1. It shortens time of execution.

4 Experiments Analysis

Experiment 1 Production of frequent itemsets

Hypotheses minimum weighted support *wminsup* ＝8％and. minimum weighted confidence *wminconf*＝10％, Table 1 is a transactions database.

Table 1 Transactions Database

TID	List of item_IDs
T1	ẠC̣D
T2	ẠC
T3	ḄD
T4	ẠD
T5	C̣E
T6	ḄḌE
T7	ḄD
T8	ḄE
T9	ḌE
T10	ẠC̣E

The weights are listed according to user’s interesingness:

A: 0.1， *B*: 0.2, *C*: 0.9 *D*: 0.4 *E*: 0.6

Mined frequent itemsets by New_Apriori algorithm are as follows:

$\{C\},\{D\},\{B\},\{E\},\{C,D\},\{C,E\},\{B,D\},\{B,E\},\{D,E\}$

Mined frequent itemsets by W_Apriori algorithm are as follows:

$\{C\},\{D\},\{B\},\{E\},\{A,C\},\{A,D\},\{C,D\},\{C,E\},\{B,D\},\{B,E\},\{D,E\},\{A,C,D\},\{A,C,E\}$

To test the two algorithms again, this experiment was done on the PC of Celeron 1.7 GHz/512M, and the WINDOWS SERVER 2003 system and synthesis database was selected. The database contained 50 transactions and maximum count of the itemsets is 8. Figure 1 showed the result of the experiment:

From the graph, we can see that W_Aprior algorithm mined more frequent itemsets than New_Apriori algorithm. It proved that in the process of pruning, some itemsets are non-frequent now, but its super set may be frequent. More weighted itemsets were mined. These were the itemsets we concerned.

Experiment 2: The improvement of generating rules the efficiency is improved effectively because the frequent itemsets that can't generate rules were pruned according to theorem 1, especially when there are more attributes. The more frequent itemsets were found, the larger transaction database is, as a result the experiment is in embarrassment of generating the frequent itemsets and association rules. Based on the above discussion, choose 50 transactions to test, set same minimum weighted support and confidence, Figure 2 demonstrated the experiment result:

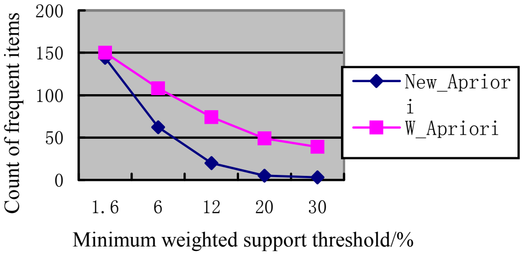


Figure 1 The Comparison of New_Apriori Algorithm and W_Apriori Algorithm in Generating Frequent Itemsets

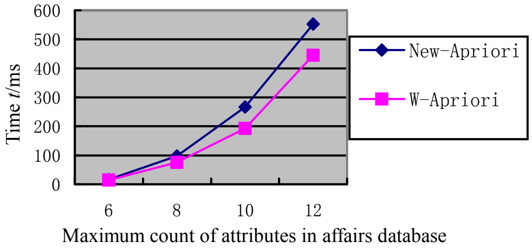


Figure 2 the time comparison of New_Apriori algorithm and W_Apriori algorithm in generating association rules

5 Conclusions

An improved algorithm based on the Apriori is proposed in this paper, on the condition that mining weight association rules is not satisfied with the Apriori property. Make some variations to mine ampler frequent items aiming at the questions that the potential frequent itemsets discard possibly at the prune step. The experiments prove the algorithm is effective and feasible. But the algorithm is at the cost of efficiency, so it needs optimizing.

Acknowledgements

This work was partially supported by Nature Science Foundation of Shandong Province China (Y2007G25) and Young Scientist Foundation of Shandong Province, China (2006BS01017).

References

[1] J.Han et al. Data Mining: Concepts and Techniques, Second Edition. Morgan Kaufmann Publishers,2006

[2] R.Agrawal,T.Imielinski,and A.Swami. Mining Association Rules Between Sets of Items in Large Databases [A].Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data[C].New York: ACM Press,1993:207-216

[3] C. H. Cai, Ada Wai-Chee Fu, C.H. Cheng et al. Mining association rules with weighted items [A].Proceedings of the international Database Engineering and Application Symposium[C]. Cardiff, UK,1998:68-77

[4] Z.J.Zhang, Y.Fang, Y.T.Xu, "Mining Horizontal Weighted Association Rules Based on Algorithm Apriori", *Journal*, Computer Engineering and Application, China, 2003, 14(39),pp.197-199

[5] M,Yang, Z.H.Sun, P.Yang, "Algorithm for Mining Weighted Association Rules Based On Record Partition", Journal, Journal of Chinese Computer System,China, 2003, 24（10）,pp.1779-1782

[6] Agrawal R,Srikant R, "Fast Algorithms for Mining Association Rules in Large Databases", In: Proceedings of the 1994 International Conference on VLDB[C].San Francisco: Morgan Kaufmann Publishers,1994:487-499

[7] W.Wang, J.Yang, P.Yu, "Efficient mining of weighted association rules(WAR)", In: Proc.of the ACM SIGKDD Conf.on Knowledge Discovery and Data Mining, 2000, pp.270-274

[8] M,Yang, Z.H.Sun, P.Yang, "Algorithm for Mining Weighted Association Rules Based On Record Partition", Journal, Journal of Chinese Computer System,China, 2003, 24（10）,pp.1779-1782

[9] F.S. Lu, H.P. He. Ming Weighted Association Rules[J]. Mimi-micro System,2001, 22（3）:347-350

[10] W.M.Ouyang, C. Zhang, Q.S.Cai, Discovery of Weighted Association Rules in Databases [J]. Journal of Softwar,2001,12（4）:612-619

Mining Positive and Negative Association Rules with Weighted Items

He Jiang Yuanyuan Zhao Xiangjun Dong Shiju Shang

School of Information Science and Technology, Shandong Institute of Light Industry, Jinan, 250353, China

Email: jianghe@sdili.edu.cn, xiaofengye1013@163.com, d-xj@163.com, shiju82@163.com

Abstract

Many interesting rules are lost if the minimum support is high, but combinatorial explosion is likely to occur if the minimum support is small. To solve the problem, the traditional association rules are extended by allowing a weight to be associated with each item in a transaction to reflect interest or intensity. The positive association rules can be found in the past weighted association rules, whereas the negative weighted association rules are as important as the positive. Misleading rules maybe occur when simultaneously studying the positive and negative association rules. In this paper, an algorithm for mining the positive and negative weighted association rules (PNWAR) is proposed to solve the corresponding question. The correlation method is applied to the algorithm for mining the weighted association rules. Not only is the difference of the items in the database solved, but also the negative association rules are mined and the contrary rules are eliminated using the algorithm. An experiment is performed and the results of it show that the algorithm is efficient.

Keywords: Association Rule, Algorithm Apriori, PNWAR, Weight, Frequent Itemset, Correlation

1 Introduction

Association rule mining (ARM) is firstly proposed by R.Agrawal, T.Imielinski and A.Swam in 1993. It is applied to find out the interesting characteristics and patterns that are not explicitly presented in the data and to provide the gist for the decision-maker [1]. The fast algorithm called Algorithm Apriori is put forward in 1994[2]. All the items in a database are treated in a uniform in the traditional Algorithm Apriori. However, it is not true in the real world databases, in which different items usually have different importance. The model of weighted association rule and corresponding algorithm are proposed by C. H. Cai et al.[3]. The subsets of the frequent items are not frequent and the supersets of the non-frequent items are not non-frequent possibly in the model. Wei Wang et al. proposed an efficient mining method for Weighted Association Rules (WAR) [4]. If some products whose profit is high or that is on sales promotion is interesting for the decision-maker, then they are set a larger weight to show their importance. For example, [wine salmon, 1%, 80%] may be more important than [bread milk, 3%, 80%] even though the former holds a lower support. This is because those items in the first rule usually come with more profit per unit sale, but the standard ARM simply ignores this difference. In addition, WAR can also be used in many other applications, such as web trace and so on. Previous work can mine the positive association rule with weighted items. Many researchers make a lot of work but mainly to improve the efficiency of algorithm. Improved algorithms for mining weighted association rules are proposed in [5][6].

Negative association rules play an important role in many areas, especially in competitive analysis and investment analysis. But little study is about negative association rules, especially weighted negative association rules. The negative relationships between two frequent items were first mentioned in 1997 [7]. The extended association rules and atom association rules are discussed in [8]. In [9] the authors propose strong negative association rules. Some problems such as self-contradictory rules may occur when studying both the positive and negative association rules simultaneously. In [10] the authors propose correlation method to solve these questions. This method is extended to the weighted association rules in this paper. The items that the user pays attention to are mined more efficiently and the mined misleading rules simultaneously are eliminated. An algorithm for mining the positive and negative weighted association rules (PNWAR) is proposed by the way that the correlation method is combined with the method for mining the weighted association rules. The frequent items are generated by the method of level-wise search. In the prune step k-support bound of itemsets are used to reduce the size of candidate itemsets.

2 Problem Formulation and Related Work

2.1 Formulation of Weighted Association Rules

Weighted Association Rules is as follows: there are n records and m items. Let $I=\{i_1,i_2,....,i_m\}$be a set of items and $W=\{w_1, w_2,\ldots, w_m\}(w_i\in[0,1])$ be the set of non-negative real numbers. A pair(X, $w(X)$) is called a weighted item where $X\in I$ is an item and $w(X)\in W$ is the weight associated with X. A transaction is a set of weighted items, each of which may appear in multiple transactions with different weights. A dataset may therefore be defined as a set D of transactions.

Given a weighted itemset X and a set of transactions, referred to as D, we say X has support in D if $s\%$ of transactions in D support X. The weighted support of an itemset X in a dataset D, denoted as $wsup(X)= countD(X)/|D|*w(X)$, where $countD(X)$ is the number of transactions in D containing X. Note that the support of a weighted itemset X is always less than or equal to the support of any of its generalization.

There are several different formulations for the weighted support of the rule $X\Rightarrow Y$ according to the different demand, where $X\subset I$, $Y\subset I$, and $X\cap Y=\varnothing$. In this paper the following form is used:

$$sup(X\Rightarrow Y)= countD(X\Rightarrow Y)/|D|$$

$$wsup(X\Rightarrow Y) = avg\{ w_1, w_2,\ldots, w_p \} \bullet sup(X\Rightarrow Y)$$

where p is the items' count of the itemset X, $avg\{ w_1, w_2,\ldots, w_p \}$ is the average value of the weight of itemset $X\cup Y$ *and* $sup(X\Rightarrow Y)$ is support of itemset $X\cup Y$ before the item is associated with weight.

An itemset is said to be weighted-frequent (large) if its support is larger than a user-specified value (also called minimum weighted support(*wminsup*)).

The confidence of the WAR denoted as *conf* is the ratio of the weighted support of $X\cup Y$ over the weighted support of X. The weighted support of $X\cup Y$(*wsup*) in the transactions is larger than *wminsup*, furthermore when X appears in a transaction, Y is likely to appear in the same transaction with a probability *conf*. $X\Rightarrow Y$ is a valid rule if its weighted support $wsup(X\Rightarrow Y)$ and weighted confidence $wconf(A\Rightarrow B)$ meet minimum weighted support (*wminsup*) and minimum weighted confidence (*wminconf*).

2.2 Calculate Support and Confidence of Weighted Negative Association Rules

There are very few papers to discuss and discover negative association rules. Brin et.al[7]mentioned for the first time in the literature the notion of negative relationships. In [11], each positive rule $X\Rightarrow Y$ correspond three negative ones, $X\Rightarrow\neg Y$, $\neg X\Rightarrow Y$ and $\neg X\Rightarrow\neg Y$. A transaction t supports $X\Rightarrow\neg Y$ if $X\subseteq t$ and $Y\not\subset t$.

Mining negative association rules, however, raises a number of critical issues. The number of infrequent itemsets increases by exponential, so the search space of the negative association rules is large than the positive. Therefore, the support and confidence of the negative association rules is difficult and disinteresting directly. For instance, if the number of the frequent itemsets is 2^{50}, the number of the infrequent ones is (2^{1000}-2^{50}) approximately 2^{1000}. So support and confidence of the negative ARs can be straightforwardly deduced from the corresponding positive itemset supports. In fact, the support and confidence of negative association rule associated with weight are transformed to the following forms to be calculated [10]:

Calculate weighted support of WARS:

（1） $wsup(\neg A) = 1\text{-}wsup(A)$;

（2） $wsup(A^{x}\neg B) = wsup(A)\text{-}wsup(A^{U} B)$；

（3） $wsup(\neg A^{U} B)$

$= wsup(B)\text{-}\ wsup(A^{q_k = \min\{p_{1k}, p_{2k}, p_{3k}, p_{4k}\}} B)$;

（4） $wsup(\neg A^{U} \neg B) = 1\text{-}wsup(A)\text{-}\ wsup(B) + wsup(A^{F} B)$；

Calculate weighted confidence of WARS:

（5） $wconf(A\Rightarrow\neg B) = \dfrac{w\sup(A) - w\sup(A \cup B)}{w\sup(A)} = 1\text{-}\ wconf\ (A\Rightarrow B)$;

（6） $wconf(\neg A\Rightarrow B) = \dfrac{w\sup(B) -\ w\sup(A \cup B)}{1 - w\sup(A)}$

（7） $wconf\ (\neg A\Rightarrow\neg B) = \dfrac{1 - w\sup(A) - w\sup(B) + w\sup(A \cup B))}{1 - w\sup(A)}$

$=1\text{-}\ wconf\ (\neg A\Rightarrow B)$

3 Efficient Discovery of Both Positive and Negative Assciation Rules

3.1 Correlation

The most common framework in the association rules generation is the “support-confidence” one. Although these two parameters allow the pruning of many associations that are discovered in data, there are cases when many uninteresting even self-contrary rules may be produced especially when both positive and negative association rules are mined simultaneous.

Example 1. Suppose we are interested in analyzing the transactions of apples (denoted by A) and bananas (denoted by B) in a supermarket. It is known to all that $sup(A) = 0.5$, $sup(B) = 0.5$, $sup(A\cup B) = 0.05$, $w(A) = 0.9$, $w(B)=0.9$. Then

wsup (A) = 0.45 wsup(B)=0.45 wsup(A ∪ B)=0.45

$conf\ (A\Rightarrow B) = 0.1 \quad conf(A\Rightarrow\neg B) = 0.9$

$conf\ (\neg A\Rightarrow B) = 0.72 \qquad conf(\neg A\Rightarrow\neg B) = 0.38$

Suppose $wminsup$=0.1 and $wminconf$=0.1, the four rules are valid association rules. Obviously, the rule $A\Rightarrow B$ and the rule $A\Rightarrow\neg B$ are inconsistent. But which one is right or all are right?

The reason of contradiction is the negation among items. In addition, the other possible case is that the items are independent. The example above illustrates some weaknesses in the “support-confidence”. However, the correlation coefficient in the statistics can solve the question efficiently. Therefore, in this paper we consider another framework that adds to the support-confidence some measures based on correlation analysis.

The definition of the correlation is proposed in [11]. The measure correlation is calculated by the following form [10]:

$$corr_{A,B} = \frac{w\sup(A \cup B)}{w\sup(A)w\sup(B)}$$

There are three possible cases for the values of the correlation $corr_{A,B:}$ if $corr_{A,B}=1$, then the two variables are independent. When $corr_{A,B} >1$ the two variables considered are perfectly positive correlated. Similarly, when $corr_{A,B}<1$, the variables considered are perfectly negative correlated.

Moreover, there are such relationships between the four forms: if $corr_{A,B} >1$, then $corr_{A,\neg B} <1$, $corr_{\neg A,B} <1$, $\mathrm{corr}_{\neg A,\neg B}>1$ and vice versa. Therefore, the contrary rules can be avoided only if the correlation between the items be judged. When $corr_{A,B} >1$, the forms $A\Rightarrow B$ and $\neg A\Rightarrow\neg B$ are discovered. When $corr_{A,B} <1$, the forms $\neg A\Rightarrow B$ and $A\Rightarrow\neg B$ are discovered. But when $corr_{A,B} =1$, nothing can be mined.

3.2 Valid Positive and Negative Weighted Association Rules

In this section, we add on top of the support-confidence framework another measure called correlation and an efficient algorithm PNWAR. The rules are mined from the frequent items. Suppose the frequent items are put in the itemset L. The frequent itemset L is generated by the method of level-wise search. In the prune step k-support bound of itemsets are used to reduce the size of candidate itemsets because the weight is added and the Apriori property is no longer in point.

The main procedure of Algorithm PNWAR:

Input L: frequent itemsets, wminsup: minimum weighted support, wminconf: minimum weighted confidence

Output PAR: set of positive associate rule; NAR: set of negative associate rule;

```
（1）PAR =∅；NAR =∅；
（2）//mining PNWARs in L.
for any item set X in L do  {
for any item set $A \cup B = X$ and $A \cap B = \varnothing$ do {
corr=wsup(A∪B)/(wsup(A)wsup(B))；
if corr >1 then {
(2.1)//generate rules of the forms $A \Rightarrow B$ and $\neg A \Rightarrow \neg B$.
if $c(A \Rightarrow B) \geqslant minconf$  then
   $PAR = PAR \cup \{A \Rightarrow B\}$;
if $c(\neg A \Rightarrow \neg B) \geqslant minconf$  then
      $NAR = NAR \cup \{\neg A \Rightarrow \neg B\}$;
 }
if corr <1 then {
(2.2)//generate rules of the forms $A \Rightarrow \neg B$ and $\neg A \Rightarrow B$.
if $c(A \Rightarrow \neg B) \geqslant minconf$  then
  $NAR = NAR \cup \{A \Rightarrow \neg B\}$;
if $c(\neg A \Rightarrow B) \geqslant minconf$  then
  $NAR = NAR \cup \{\neg A \Rightarrow B\}$;
      }
}
 }
（3）return PAR and NAR;
```

In this algorithm, it can produce the positive and negative rules in four forms that meet the support and confidence thresholds and eliminate the self-contrary rules.

4 Experiment Results

We conducted our experiments on a man-made dataset to study the behavior of the algorithms compared. The dataset had 200 transactions, when only the six largest categories were kept. We compare with Algorithm Apriori. Suppose *wminsup*=8%, *wminconf*=10%, each algorithm was run to generate a set of association rules. The result was reported as follows:

Table1 The Number of Rules Generated by two Algorithms

Algorithm	The number of positive rules	The number of negative rules
Apriori	529	
PNWAR	487	520
The number of being pruned	42	

It is clear from these graphs that PNWAR is efficient because the number of positive rules decreases. It demonstrates that the contrary rules are eliminated and a number of negative rules are mined simultaneously.

5 Conclusions and Future Research Directions

In this paper we introduced a new algorithm to generate both positive and negative weighted association rules. Our method adds to the support-confidence framework the correlation to eliminate the self-contrary rules. The experiment results performed in the man-made database prove that our algorithm is efficient.

Acknowledgements

This work was partially supported by Nature Science Foundation of Shandong Province China (Y2007G25) and Young Scientist Foundation of Shandong Province, China (2006BS01017).

References

[1] R.Agrawal,T.Imielinski,and A.Swami, “Mining Association Rules between Sets of Items in Large Databases”, In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data.New York: ACM Press, 1993, pp. 207-216

[2] Agrawal R,Srikant R, “Fast Algorithms for Mining Association Rules in Large Databases”, In: Proceedings of the 1994 International Conference on VLDB[C].San Francisco:Morgan Kaufmann Publishers,1994:487-499

[3] C.H.Cai, Ada Wai-Chee Fu, C.H.Cheng et al, “Mining Association Rules with Weighted Items”, In: Proceedings of the international Database Engineering and Application Symposium, Cardiff, UK,1998, pp. 68-77

[4] W.Wang, J.Yang, P.Yu, “Efficient mining of weighted association rules(WAR)”, In: Proc.of the ACM SIGKDD Conf.on Knowledge Discovery and Data Mining, 2000, pp.270-274

[5] Z.J.Zhang, Y.Fang, Y.T.Xu, “Mining Horizontal Weighted Association Rules Based on Algorithm Apriori”, Journal, Computer Engineering and Application, China, 2003, 14(39),pp.197-199

[6] M,Yang, Z.H.Sun, P.Yang, “Algorithm for Mining Weighted Association Rules Based On Record Partition”, Journal, Journal of Chinese Computer System,China, 2003, 24（10）,pp.1779-1782

[7] S. Brin, R.Motwani, C. Silverstein, “Beyond Market: Generalizing Association Rules to Correlations”, In: Processing of the ACM SIGMOD Conference, 1997, pp. 265-276

[8] X. Li, Y. Liu, J. Peng, “The Extended Association Rules and Atom Association Rules”, Journal, Journal of Computer Research and Application, China , 2002.12, pp.1740-1750

[9] A. Savasere,E. Omiecinski, S. Navathe, “Mining for Strong Negative Associations in a Large Database of Customer Transaction”, In: Proceedings of the 1998 International Conference on Data Engineering, 1998 ,pp. 494-502

[10] X.J. Dong, S.J.Wang, H.T.Song et al, “Study ON Negative Association Rules”, Journal, Journal of Beijing Institute of Technology, China, 2004, pp. 978-981

[11] X.D. Wu, C.Q. Zhang, S.C. Zhang, “Mining Both Positive and Negative Association Rules”, In: Proceedings of the Nineteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2002, PP. 658-665

The Present Situation and Development Tendency of Classification Based on Negative Association Rules

Long Zhao[1] Xiangjun Dong[1, 2] Runian Geng[1] He Jiang[1]

1 School of Information Science and Technology, Shandong Institute of Light Industry,
Jinan 250353, P.R. China

2 School of Management and Economics, Beijing Institute of Technology,
Beijing 100081, P.R. China

Email: zxcvbnm9515@163.com; dxj@sdili.edu.cn; grn@sdili.edu.cn; jianghe@sdili.edu.cn

Abstract

One application of association rule mining (ARM) is to identify classification association rules (CARs) that can be used to classify future instances from the same population as the data being mined. Associative classification is a well-known technique which uses association rules to predict the class label for new data object. This model has been recently reported to achieve higher accuracy than traditional classification approaches like C4.5.Only limits in view of positive association rules used in classification. This paper introduces main methods and the present situation of classification based on association rules, expatiates the present situation and main technologies of negative association rules, introduces the problem of exploding frequent itemsets and points out the development tendency of classification based on negative association rules.

Keywords: Negative Association Rules; Classification Method; Frequent Negative Itemsets

1 Introduction

Association rules describe the co-occurring relationships among data item in the large transaction database. They have been extensively studied in the literature for their usefulness in many real world areas such as market baskets analysis, expert system, stocks trend prediction, even public health surveillance, etc. So association rule mining is always a major topic in data mining research community. There have been discovered many efficient algorithms and their variants of association rules, such as: Apriori, FPgrowth, etc. In recent years, a new classification technique, called associative classification, is proposed to combine the advantages of association rule mining and classification. Classification is a supervised machine learning method, which aims to build a classifier and make prediction for new data object whose class label is unknown. Traditional association rule mining has been mainly focused on identifying the relationships strongly associated among itemsets that are frequent and high-correlation. The form AÞB was called positive association rules (PARs). As an important supplement, the other three forms AÞ¬B, ¬AÞB, ¬AÞ¬B call negative association rules (NARs). NAR catch mutually exclusive correlations among items; they can provide much useful information and play an important role in decision-making. How to use both PARs and NARs to class is an urgent problem.

2 Main Technologies and Status of Classification Base on Association Rules

Classification rules mining and association rules mining are two important data mining technologies. The goal of association rules is not pre-determined, but the goal of classification rules is identified in advance, that is category labels. Classification and association rules mining are indispensable in practice. Therefore, data mining technology has used association rules mining for classification. The combinations of these two technologies create a new classification method: classification based on association rules. (CARs)

2.1 Problem states

After Agrawal[1]proposed mining association rules in database in 1993, association rules no longer confined to the concept of functional description. Association rules mining algorithm and applications developed rapidly. Then Ali proposed the thought of partly classification. At the same year, Brin put forward to the thought of finding out strong related rules in the large dataset for the first time and leaded into the parameter χ^2 to measure the related degree among rules.

Till 1998, Liu proposed an algorithm called CBA [2] that integrates association rule mining with classification. The algorithm used a similar Apriori algorithm to generate the association rules. Set the threshold of support is 1%, the threshold of confidence is 50%.The result of experiment shows the classification error rate and compare to the decision tree algorithm C4.5[3].However, the main problem of this study is generating too many association rules by similar Apriori algorithm. The consumption of system resources is too much. In the literature [4], the author created a revealed classification model. At the same year Li. used jumped revealed model to class as the improvement of revealed model.

W.Hopkins discussed the calculate method of the correlation coefficient’s value and improved the accuracy of the classification in 2002. ARC-AC [5] and ARC-BC [6] have been introduced in the aim of text categorization. They generate rules similar to the Apriori algorithm and rank them in the same way as do CBA rules ranking method. ARC-AC and ARC-BC calculate the average confidence of each set of rules grouped by class attribute in the conclusion part and select the class attribute of the group with the highest confidence average.

A greedy associative classification algorithm called CPAR [7], which adopts the Foil strategy to generate rules, was proposed by Yin and Han. CPAR seeks the best rule that brings max Foil Gain among the available ones in the dataset. The accuracy of CPAR is similar to CMPR, but the number of generated rules is less than CMAR .Liu.et al used complete classification association rules set to improve the classification accuracy by scoring unknown examples .But the practicality of it is poor. To increase its practicality, Elena Baralis proposed basic rules, few items and a support-based compression rules algorithm. This method well compressed the original rules and allows for a smaller support.

Keivan Kianmehr proposed CARSVM [8] in 2006. CARSVM is a general classification framework that attempts to make the classification task more understandable and efficient by integrating association rule mining and SVM technique. Rule-based feature vectors present a high-quality knowledge extracted from the training sets .In addition to single association rules classification algorithm, W.LI discussed classification algorithm CMAR [9] which was based on multiple association rules. This technique uses improved frequent pattern of growth (FP-growth) to mine the association rulers and identify the category labels by using strong association rulers based on different weights.

CMAR generates rules in a similar way as CBA with the exception that CMAR introduces a CR-tree structure to handle the set of generated rules and uses a set of them to make a prediction using a weighted $\chi 2$ metric. The latter metric evaluates the correlation between the rules. However, there are two problems exist in this algorithm: 1, although FP-growth is faster than Apriori, but FP-growth don’t suit for high-dimensional and large databases.2, it is difficult to predict the weight of strong association rules.

2.2 The main classification technologies

Association rules mining in data mining includes the following steps:

（1）If the attribute values are continuous then discrete them

（2）Generate all class association rules satisfying certain user-specified threshold as candidate rules by association rule mining algorithm. These discovered rules have the form (attributes ⇒ class label).

（3）Establish a classifier based on association rules that have been generated

（4)Assign a class label for a new data object. When a new data object without a class label comes, the classifier ranks the fitness of these rules and selects some or all suitable rules to make a prediction.

As figure 1 show, we assume that the datasets is a normal relationship table. This table contains *N* cases with *V* different attribute values which have been classified to *C* known categories. Attribute values may be discrete or continuous. To the continuous attribute values, we firstly separated them into many small intervals, then bring these small intervals mapping to a series of consecutive integer values.

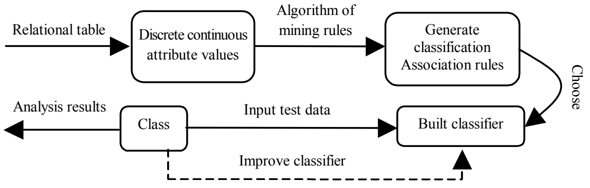

Relational table
Discrete continuous attribute values
Algorithm of mining rules
Generate classification Association rules
Choose
Analysis results
Class
Input test data
Built classifier
Improve classifier

Figure 1 associated classification flow chart

Let *D* be the dataset. Let *I* be the set of all items in *D*, and *Y* be the set of class labels. A classification association rule (CAR) is an implication of the form $A \Rightarrow B$. where $A \in I$, $B \in Y$. The rule $A \Rightarrow B$ holds in the transaction set *D* with support s, where s is the percentage of transactions in *D* that contain $A \cup B$. This is taken to be the probability, $P(A \cup B)$. The rule $A \Rightarrow B$ has confidence *c* in the transaction set *D* if *c* is the percentage of transactions in *D* containing *A* that also contain *B*. This is taken to be the conditional probability, then establish a classifier by CARs.

Current methods of building association classifiers could be divided into 3 categories.

1. Mining frequent item sets. This is also the traditional way to build association classifiers. CBA, CMAR, CPAR, L3, L3M, and CSFP are classifiers belong to this category. They either mine frequent itemsets, or mine rules with left part only made up of attribute information, and right part only made up of category information for classification.

2. Mining EPs (Emerging Pattern). EPs are defined as event associations whose supports change significantly from one dataset to another. CAEP, iCAEP, and DeEPs belong to this category.

3. Mining associations with help of SVM. To the best of our knowledge, only DRCBK and ACIK belong to this category.

For CBA and L3, their algorithms for selecting one rule for classification are heuristic. For CPAR, L3M, CSFP, DeEPs, and so on, the algorithms to set weight to association rules (EPs), and the algorithm for classification, are also heuristic. These algorithms lack strong mathematical background, which limits their accuracy. ACIK takes SVM with item set kernel as learning engine, which means that ACIK is also based on structural risk minimization theory, and that ACIK also makes classification decision by the optimal hyper plane in the feature space.

3 The Present Situation and Techniques of Mining NARs

3.1 Present Situation

NARs was first noted by Brin[10]. Since then there have been several attempts to solve the problem of NARs mining. Savasere [11] proposed a method to mine strong negative association rules. They discovered PARs and combined them with domain knowledge to constrain the search space to mine interesting NARs.

Xindong Wu introduced a method mining both positive and negative association rules [12]. This method mine both frequent items and infrequent items based on traditional Apriori algorithm. It can clearly mine $A \Rightarrow \neg B$, $\neg A \Rightarrow B$ and $\neg A \Rightarrow \neg B$ in infrequent items.

Antonie and Zaıane[13] introduced an automatic progressive threshold process. They used the Pearson's φ correlation coefficient to measure the strength of an association. If there are no strongly correlated rules found in any round, the threshold is lowered progressively until some rules are found with moderate correlation .In [14],a rapid and effective FP-tree based NARs mining algorithm MNARA was proposed. This algorithm not only can mine all NARs, but also only need scanning database *D* twice. It is Effective and feasible.

[15]proposed a PNARC(Positive and Negative Association Rules on Correlation) method and proposed a simple but efficient method to calculate the support and confidence of the NARs through positive association rule's. Furthermore this model can detect and delete those self—contradictory rules by correlation. A method of mining PARs and NARs based on multi-confidence and Chi-squared test is proposed this method solutes a dilemmatic situation of using a single confidence threshold. [16] proposed a method mining PARs and NARs based on militia confidence and χ2 test named PNARMC. This algorithm can correctly produce PARs and NARs and flexibility control the number of association rules.

3.2 The techniques of mining NARs

PNAR_MLMS [17] use known support and confidence of the positive rules to calculate support and confidence of the NARs, .Set *A*，*B*⊂ *I*, *A*∩*B*=Φ then

（1） $\text{supp}(\neg A) = 1\text{-supp}(A)$

（2） $\text{supp}(A \cup \neg B) = \text{supp}(A)\text{-supp}(A\ B)$

（3） $\text{supp}(\neg A \cup B) = \text{supp}(B)\text{- supp}(A\ B)$

（4） $\text{supp}(\neg A \cup \neg B)=1\text{-supp}(A)\text{-supp}(B)+ \text{supp}$ $(A\ B)$

（5） $$\text{conf}(A \Rightarrow \neg B) = \frac{supp(A)\text{ - }supp(A \cup B)}{supp(A)}$$
$$=1 - \text{conf}(A \Rightarrow B)$$

（6） $$\text{conf}(\neg A \Rightarrow B)=\frac{supp(B)\text{- }supp(A \cup B)}{1\text{-}supp(A)}$$
$$=\frac{supp(B)\text{- }supp(A)*conf(A \Rightarrow B)}{1\text{-}supp(A)}$$

（7） $$\text{conf}(\neg A \Rightarrow \neg B)=$$
$$\frac{1\text{ - }supp(A)\text{ - }supp(B)+supp(A \cup B))}{1\text{ - }supp(A)}$$
$$=1\text{- conf}(\neg A \Rightarrow B)$$

Literature [18] definite the relationship of item sets. Set *A* and *B* as two random events, s*upp(A),supp(B)* as the support of *A* and *B*, *P(A),P(B)* as the probability of the occurrence. The correlate coefficient of them is:

$$\rho_{AB}=\frac{supp(A \cup B)\text{ - }supp(A)supp(B)}{\sqrt{supp(A)(1\text{ - }supp(A))supp(B)(1\text{ - }supp(B))}}$$

There are three possible scenarios of ρ_{AB}

（1）If ρ_{AB} >0, then *A* and *B* are positive correlated, the more *A* occurs, the more *B* occurs

（2）If ρ_{AB} = 0, then *A* and *B* are independent the appearance of *B* unrelated to *A*

（3）If ρ_{AB}<0, then *A* and *B* are negative correlated, the more *A* occurs, the less *B* occurs

Set *mc* as minsup, *α* as measure of related strength in [19]. If ρ_{AB}⩽*α* (0≤α≤1) then:

（1）$\rho\neg AB \leqslant -\alpha$; （2）$\rho A\neg B \leqslant -\alpha$; （3）$\rho\neg A\neg B \leqslant \alpha$;

PNAR_MLMS proposed a new measure VARCC which combines correlation coefficient *α* and minimum confidence *mc* to mine positive rules in frequent item sets and negative rules in both frequent and infrequent item sets. A rule *A*⇒*B* can be extracted as a valid association rule if it meets *VARCC* (*A, B*,α, *minconf*)=1.

3.3 Exploding number of frequent itemsets in the mining of NARs

If the data set contains 100 items and every of them has a corresponding negative item, the gross of items is 200, the ceiling count of 2-items is 200×200 =40000, the ceiling count of *k*-items is 200^k. The system is likely to break down because of the rapid expanding of the items' count if all NARs are mined in the data set. That is question of explosion frequent patterns.

In [19] an algorithm is proposed in which the generating of frequent items is restricted by minimum support, maximal support and the large count of negative items comprised in the itemsets. The experiment results demonstrate two supplemental parameters are necessary especially when the data set contains much items. If the additional parameters are set properly, the mining can be carried through successfully.

Supposed the maximal support under which the interesting association rules are generated is *smax*, and a item contains negative, if *s (A)* > *smax*, the items is called over frequent itemsets

Axiom 1 If item *A* is over frequent itemsets, then the association rule which item *A* is educed to form former is not significant

Theorem 1 Interesting associations rules are impossibly generated by the over frequent itemsets or their super sets.

The association rules that *A* and it's sub sets are educed from the front items is meaningless, according to Axiom 1. IF association rules were generated by *W*, only following two forms may be done:

1) $A \Rightarrow X \cup Y$;

2) $A \cup X \Rightarrow Y$ (Or $A \cup Y \Rightarrow X$).

It is very important to generate the frequent items in mining the NARs .The two parameters of the maximal support and the large count of negative items contained in the item sets play a vital role, especially when data set contains so many items. If two parameters have been not set, the system is likely to break down because of the rapid expanding of the items' count, whereas the proper values of the two parameters can make the negative association rules mining successful.

4 The Development Tendency of Classification Based on NARs

From 1998 until now, the researches and applications on the classification of association rules have been never interrupted. It is a very important domain of data mining study nowadays. At the same time, study on NARs opens a new area for the traditional association rules. With the economy development and people's eager on the enhances of information, the study based on the classification of association rules becomes more and more important. The researches on the classification of association rules exists some limitation in the national area, and a lot of work need to do. We have carried on the forecast from the following several aspects of based on the classification of association rule and negative association rules.

1. Classification based on negative association rules

Classifier strategy is the method that based on NARs, and which is based on the classification of association rules, in order to demonstrate more effective direct connection among realistic events. The union between the classification method and PARs can carry on the classification effectively and accurately, and also can analysis inner links among each kind of hided factor more comprehensively. The researcher based on the classification of NARs will certainly to be the hot spot in the future study.

2. Classification in multi-databases

Multi-databases contain different kind of databases, how to mine association rules in it and class the itemsets effectively are important. Besides the extant attributes in the database, it is necessary to think about other attributes, such as localization. For example, when the MacDonald opens a branch store in some place, Kentuckey will open a Kentuckey branch store in the coming two months and in a mile scope. Therefore, how to mine the NARs in the multi-databases from the different databases will be a new research direction.

3. The study of classification not based on the support

Some works already pointed out support threshold based method may not be enough accurate in some case. And it is not easy to decide on a good value. Some recent novel studies don't use support threshold to evaluate the information gain of the current rule, such as Foil Gain. This gain values are calculated by using the total weights in the positive and negative instance sets instead of simply counting up the number of records in the training sets. By using, Compared to the traditional method which establish support-confidence structure, this gain measure is economic and accurate. How to calculate and set an appropriate Foil Gain has become a hot researcher spot.

4. Classified mining under the network and the distributed environment

With the development of internet, facing on the environment of distributed and networks, both data dimensions and data samples are growing at a high speed. This needs enhancing the efficiency of the data mining. But existed the methods of the neural network data mining, generally use the neural network training, cutting, and training again, cutting again…. The repeated method carries on the attribute choice and the rule extraction. For the database with small attribute dimension, this method can mine efficiently. But for the big dimension data, the shortcoming of this method appearance completely. In order to mine the high dimension data, it is necessary to statement new attribute choice method and rule extraction method to overcome the limitation of the existing excavation algorithm. Nowadays, the research about multi-Agent system in artificial intelligence provides us beneficial model and helps.

References

[1] Agrawal R, Imielinski T, Swami A. Mining associations between sets of item s in large databases [J]. Proceeding of the 93ACM 2S IGMOD international conference on management of data [Washington: Springer -Verlag, 1993.207—216

[2] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. Proceedings of the KDD[R], New York, NY, 1998.80-86

[3] J. R.Quinlan. Programs For Machine Learning. Morgan Kaufmann[R], San Francisco, 1993

[4] G.Dong, X. Zhang, L. Wong, and J. Li. CAEP: Classification by aggregating emerging patterns[C]. In Procaine and International Conference on Discovery Science, 1999

[5] Antonie, M., Zaiane, O. Text Document Categorization by Term Association.In[C]: Proc. of the IEEE International Conference on Data Mining (ICDM'2002), Maebashi City, Japan (2002) 19–26

[6] Antonie, M., Zaiane, O. Classifying Text Documents by Associating Terms with Text Categories [C]. In: Proc. of the Thirteenth Austral-Asian Database Conference (ADC'02), Melbourne, Australia (2002)

[7] X. Yin, J. Han. CPAR: Classification based on predictive association rules.[C]Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, CA, USA, 2003

[8] Keivan Kianmehr1 and Reda Alhajj: Effective Classification by Integrating Support Vector Machine and Association Rule Mining [J], E. Corchado et al. (Eds.): IDEAL 2006, LNCS 4224, pp. 920–927

[9] W. Li, J. Han, J. Pei. CMAR: Accurate and efficient classification based on multiple class-association rules.[C] Proceedings of the 2001 International Conference on Data Mining (ICDM), 2001.369-376

[10] Brin, S., Motwani, R., and Silverstein, . Beyond Market: Generalizing Association Rules to Correlations. In[C]: Processing of the ACM SIGMOD Conference (1997) 265-276

[11] Savasere, A., Omiecinski, E., Navathe, Mining forStrong Negative Associations in a Large Database of Customer Transaction[C]. In: Proceedings of the 1998 International Conference on Data Engineering (1998) 494-50

[12] Wu, X., Zhang, C., and Zhang,. Efficient Mining of both Positive and Negative Association Rules, ACM Transactions on Information Systems [J], 2004, 22(381-405)

[13] Antonie, M.L., Za¨ıane, O. Mining positive and negative association rules: an approach for confined rules [D]. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD, 2004

[14] Zhu, Y., Sun, L., Yang,. Algorithm for Mining Negative Association Rules Based on Frequent Pattern Tree [J], Computer Engineering, Vol.32, No.22

[15] Dong,X..,Wang,S..,Song,H. Study of Negative Association Rules[J], Beijing Institute of Technology Journal 2004,24（11）: 978-981

[16] Dong, X., Sun, F., Han, X., Hou, R. Study of Positive and Negative Association Rules Based on Multi- confidence and Chi-Squared Test [J]. LNAI 4093, Springer-Verlag Berlin Heidelberg, 2006: 100-109

[17] Dong, X..,Niu,Z..,Shi,X..,Zhang, X..,Zhu,. Mining Both Positive and Negative Association Rules from Frequent and Infrequent Itemsets[J]. ADMA 2007, LNAI 4632, Springer-Verlag Berlin Heidelberg (2007) 122–133

[18] Cohen, J.: Statistical Power Analysis for the Behavioral Sciences (2nd.)[M]. Lawrence Erlbaum, New Jersey, 1988

[19] Ma, Z., Lu, Y.,Exploding number of frequent itemsets in the mining of negative association rules [J],JTsinghua U niv (Sci & Tech) ,2007, Vol. 47, No. 7

An Application of FCM Cluster in Hand-written Numeral Recognition Based on Zernike Moment

Xiaojun Tong[1,2] Shan Zeng[1] Qin Jiang[1] Kai Zhao[1]

1 Department of Mathematics and Physics, Wuhan Polytechnic University, Wuhan 430023, China

2 Department of Control Science and Engineering, Huazhong University of Science and Technology
Wuhan 430074, China

E-mail: tongxiaojun1998 @yahoo.com.cn

Abstract

This article first used four steps of the compatible examination, the diversity examination, the relevant examination and the principal components analyzes to screen the handwritten digit Zernike moment, then we make use of the Fuzzy C-Mean cluster recognition of the hand-written numerals based on the screened Zernike moment. The example testified this method effective, and provides the theory basis for further establishment of the hand-written numeral recognition standard library. In this process, we have also discovered two reasons which cause the Fuzzy C-Mean cluster to be unefficient.

Keyword: Zernike moment; FCM cluster; handwritten numeral recognition

1 Introduction

The handwritten numeral recognition has been the hot research issue for many years. It is also one of the most successful research topics in the domain of image processing and pattern recognition. In the past dozens of years, the researchers have proposed many recognition methods. These methods can be divided into two categories according to the different application methods [1]: Method based on structural characteristics and Method based on statistical characteristics. The statistical characteristics usually include dot density measure, moment, characteristic region, character path and digital transformation method and so on. The structural characteristics obtain the geometry and typology characteristics of the numeral characters through the analysis of the outline or the skeleton of the numeral characters, which usually includes circle, end point, intersection, stroke, outline convex-concave and so on. Along with the development of information technology as well as the computer science, the traditional recognition methods have made a big improvement, such as: multistage recognition system, introduction of artificial intelligence into blending recognition system and so on. Since 1980s, the success application of fuzzy system theory in the pattern recognition has made it become one of leading methods in handwritten numeral recognition. But as for handwritten numeral recognition, it is still a long way to find out one kind of high recognition rate and low probability of misrecognition as well as light computation burden method. Therefore, the research of high performance hand-written numeral recognition system is a challenge topic.

The Zernike moment concept was first introduced by Teague in 1980[2]. The Zernike moment is the orthogonal function system which obtains based on the Zernike multinomial. Compared with the geometry moment and the Legendre moment, its computation is a little complex. But the Zernike moment has its own good natures: Orthogonal, image revolving invariability and low noise sensitivity.

Fuzzy clustering method receives universal welcome among the multitudinous classified methods based on the objective function, namely summing up the cluster as a nonlinear programming problem with belt restrains and obtaining the fuzzy division and cluster of the data through the optimized solution. This method is simple, and can solve the question to be a broad scope, and also may be transformed into the optimized question with the aid of the classical mathematics nonlinear programming theory, and can be realized easily on the computer. Based on the objective function cluster algorithm, namely Fuzzy c-Means cluster (FCM), is established by Bezdek in 1981[3]. Its theory was mostly perfect, and the application is wide.

The FCM cluster algorithm objective function is the generalized Euclidean distance of the sample characteristic vector, containing various components as the independent variable. The traditional numeral recognition methods do not have the orthogonality, making the FCM cluster inapplicable. Because of the Zernike moment has the orthogonality; it is reasonable to apply the FCM cluster in theory based on the Zernike moment. We generally selects the first 47s of the Zernike moment, but the Zernike moment still appears excessively many as the recognition characteristic which can cause the FCM cluster computation load to be big. Solving with the computer can result in the morbid state matrix, and can produce bad effects on the cluster result [4]. This article carries on the Zernike moment screening successfully used in statistics the random variable correlation coefficient and region sparse in unit plane of digitization of handwritten numeral. The example indicated this kind of screening is reasonable. We will carry on the induction of the handwritten numeral library based on the screened Zernike moment, for the further establishment of handwritten numeral recognition standard sample library. These works will be published later.

2 Zernike Moment and Screening

2.1 Zernike moment

Regarding a density image function $f(x,y)$, its n step Zernike moment definition is

$$Z_{nm}=\frac{n+1}{\pi}\iint_{x^2+y^2\le 1}V_{nm}^{*}(x,y)f(x,y)dxdy\ .$$

In the formula, * means taking the conjugate. The Zernike multinomial $V_{nm}(x,y)$ is given by the equation below:

$$V_{nm}(x,y)=V_{nm}(r,\theta)=R_{nm}(r)e^{jm\theta}$$

In the formula, n is the non-negative integer. $|m|\le n$ and $n-|m|$ satisfies for the even number, $R_{nm}(r)$ is defined as:

$$R_{nm}(r)=\sum_{s=0}^{(n-|m|)/2}\frac{(-1)^s(n-s)!}{s![\frac{(n+|m|)}{2}-s]![\frac{(n-|m|)}{2}-s]!}r^{n-2s}\ .$$

The Zernike multinomial $V_{nm}(x,y)$ satisfies the following orthogonal relations:

$$\iint_{x^2+y^2\le 1}V_{nm}^{*}(x,y)V_{pq}(x,y)dxdy=\begin{cases}\dfrac{\pi}{n+1} & n=p;m=q\\ 0 & other\end{cases}$$

Z_{nm} under the polar coordinate form may de expressed as:

$$Z_{nm}=\frac{n+1}{\pi}\int_0^1\int_0^{2\pi}R_{nm}(r)e^{-jm\theta}f(r,\theta)rdrd\theta \qquad （1）$$

From the above equation, regarding a solid two-dimensional picture, its Zernike moment Z_{nm} is a plural number, defines its real part and the imaginary component separately for C_{nm} and S_{nm}, then we have [5]:

$$C_{nm}=\frac{2n+2}{\pi}\int_0^1\int_0^{2\pi}R_{nm}(r)\cos(m\theta)f(r,\theta)rdrd\theta$$

$$S_{nm}=-\frac{2n+2}{\pi}\int_0^1\int_0^{2\pi}R_{nm}(r)\sin(m\theta)f(r,\theta)rdrd\theta \qquad （2）$$

According to the orthogonality, the inversed transformation of Z_{nm} is

$$f(r,\theta)\approx\frac{C_{n0}}{2}R_{n0}(r)+\sum_{n=1}^{M}\sum_{m>0}[C_{nm}\cos m\theta+ s_{nm}\sin m\theta]R_{nm}(r)$$

In the formula, M means the highest exponent number of the moment used.

Through formula （1）, we may also prove the revolving invariability of Zernike moment. Supposing the image revolves by angle α , after the revolving, Zernike moment uses the expression Z_{nm}^{r} , and then the

Zernike moment revolving invariability refers to:

$$Z^r_{nm} = Z_{nm} e^{-jm\alpha} \quad (3)$$

Through formula （3）, after the revolving image, $Z^r{}_{nm}$ and Z_{nm} then only have difference on the phase, no difference on the amplitude. This is the revolving invariability of Zernike moment.

2.2 Zernike moment rapid calculation

Literature [6] has studied the Zernike moment rapid calculation, namely using the Zernike multinomial iterative algorithm to reduce the calculation of the Zernike moment in the gradation image. Using the item $R_{n-2,m}(r)$ and $R_{n-4,m}(r)$ to calculate the Zernike multinomial $R_{nm}(r)$, the iterative relationship outset item may be $R_{mm}(r) = r^m$.

$$R_{n,m}(r) = \frac{(k_2 r^2 + k_3) R_{n-2,m}(r) + k_4 R_{n-4,m}(r)}{k_1}$$

Where $k_1 = (n-1)(n+1)(n-2)/2$;

$k_2 = 2n(n-1)(n-2)$; $k_3 = -(n-1)^3$;

$k_4 = -n(n-1)(n-3)/2$

Regarding the different starting value m, all must duplicate the whole process.

2.3 Zernike moment screening

Regarding the characteristic selection, the present methods may be divided into two kinds generally: the characteristic extraction and the characteristic choice. The characteristic extraction is to transform high Uygur characteristic correlation in space to low Uygur characteristic correlation in space based on the mapping method. And the new characteristic is the combination of its some primitive characteristic. The characteristic choice is to choose some most effective characteristic from the primitive characteristic to reduce the characteristic space dimension.

At present to Zernike moment, we generally select the first 47s among the Zernike moment. But the Zernike moment taken as the recognition factor theory, is oversized. Through analysis, we discovered certain redundant information in 47 Zernike moments. These redundant information do little good to enhance the digital recognition rate. On the contrary, it has enhanced the FCM cluster complexity, causing the FCM cluster computation load to be big and the morbid state matrix if solved by computers which produces bad effect on the cluster result. Regarding the characteristic value screening, mostly, use rough collection on the reduction. We attempt to use rough methods and FCM cluster method and so on to carry on the reduction of the Zernike moment. But we find that using the rough collection theory to carry on the reduction of the Zernike moment, but it needs much knowledge and many formulas, the reduction process is extremely complicated, and using the FCM cluster to carry on the reduction of the Zernike moment, because between the Zernike moment characteristic has the orthogonality, the effect is not ideal.

In classical information theory, because the information contained in the variable depends on the measurement of its variable variance. The bigger the variable variance is, the more classified information this variable contains. Based on the region sparse in unit plane of digitization of handwritten numeral, This article will introduce the following four steps to carry on the discussion: talks about the characteristic choice first to remove the characteristic with no classified information. Then it deals with the characteristic extraction method to carry on the mapping to lower the dimension.

1) The compatible examination. Carry on the variance analysis of 47 Zernike characteristics among 200 samples of each digital. Extract the average value and the variance of each Zernike characteristic. Retain the Zernike characteristic with small variance and eliminate Zernike characteristic variance with big variance. This step is to guarantee the characteristic relative dispersion to be small.

Takes each digital variance average value 3/2 time of achievement valve value, is bigger than the variance the valve value the characteristic deletion. The characteristic which retains is:

$$Z_{3,1}, Z_{4,4}, Z_{5,3}, Z_{6,0}, Z_{7,1}, Z_{7,3}, Z_{7,7}, Z_{8,2}, Z_{8,4}, Z_{8,6}, Z_{9,3},$$
$$Z_{9,5}, Z_{10,0}, Z_{10,4}, Z_{10,10}, Z_{11,3}, Z_{11,7}, Z_{11,11}, Z_{12,0}, Z_{12,6}, Z_{12,8}$$

2) The diversity examination. Make the average value of the ten numeral Zernike characteristic from step

1 screening the object of study. Extract the average value and the variance of each Zernike characteristic. Retain the Zernike characteristic with big variance and eliminate Zernike characteristic variance with small variance. This step is to guarantee the characteristic relative dispersion to be big and diverse.

Takes ten digital Zernike characteristic average value the variance average value 3/2 time of achievement valve value, is smaller than the valve value characteristic deletion the variance. The characteristic which retains is:

$$Z_{5,3}, Z_{8,2}, Z_{8,4}, Z_{8,6}, Z_{10,0}, Z_{10,4}, Z_{11,3}, Z_{11,7}$$

3) The relevant examination. Make the average value of the ten numeral Zernike characteristic from step 2 the standard Zernike characteristic values of the ten digitals. Carry on the degree analysis of correlation of each Zernike characteristic. Combine the Zernike characteristic with high degree of correlation to remove excessive information.

$$COV = \begin{bmatrix} 1 & 0.9905 & 0.7111 & 0.5258 & 0.9735 & 0.6532 & 0.9568 & 0.7413 \\ 0.9905 & 1 & 0.6875 & 0.5405 & 0.9954 & 0.6450 & 0.9868 & 0.6833 \\ 0.7111 & 0.6875 & 1 & 0.1723 & 0.6635 & 0.5665 & 0.6428 & 0.6989 \\ 0.5258 & 0.5405 & 0.1723 & 1 & 0.5631 & 0.8589 & 0.5814 & 0.6693 \\ 0.9735 & 0.9954 & 0.6635 & 0.5631 & 1 & 0.6470 & 0.9977 & 0.6450 \\ 0.6532 & 0.6450 & 0.5665 & 0.8589 & 0.6470 & 1 & 0.6488 & 0.8840 \\ 0.9568 & 0.9868 & 0.6428 & 0.5814 & 0.9977 & 0.6488 & 1 & 0.6180 \\ 0.7413 & 0.6833 & 0.6989 & 0.6693 & 0.6450 & 0.8840 & 0.6180 & 1 \end{bmatrix}$$

Eight Zernike characteristics retains which to step 2 through relevant examination, obtains the correlation coefficient matrix COV , May see from the matrix COV , $Z_{5,3}$ with $Z_{8,2}, Z_{10,0}, Z_{11,3}$ the correlation coefficient is very big. The minimum value is 0.9568, the maximum value reaches 0.9905, therefore will merge $Z_{5,3}$, $Z_{8,2}$, $Z_{10,0}$, $Z_{11,3}$ to a characteristic, and will retain $Z_{8,4}$, $Z_{8,6}$, $Z_{10,4}$, $Z_{11,7}$ these four characteristics.

4) The principal components analyzes. We analyses principally the average value of the ten numeral Zernike characteristic from step 3 to produce some new variables which are not related with each other. Select a few several new variables and enable them to include the information which primitive variables have as much as possible to lower the Uygur characteristic correlation and extract.

Obtains various principal components matrix PC and the variance matrix characteristic value *latent* through the computation.

$$PC = \begin{bmatrix} 0.5161 & -0.2925 & 0.7835 & -0.0832 & 0.1652 \\ 0.3865 & -0.6218 & -0.3683 & 0.4728 & -0.3237 \\ 0.3490 & 0.6529 & 0.1674 & 0.3128 & -0.5711 \\ 0.4390 & 0.3150 & -0.2892 & 0.3237 & 0.7209 \\ 0.5194 & 0.0484 & -0.3726 & -0.7529 & -0.1489 \end{bmatrix}$$

$$latent = \{\ 0.6565,\ 0.1508, 0.0620,\ 0.0214, 0.0041\}$$

Various principal components quantity is multiplied by the matrix PC by the primitive characteristic to be possible to result in. *latent* has reflected various characteristics weight. Through the principal components analysis, takes the first four characteristics the principal components transforma tion to take the new characteristic.

The above mentioned steps 1-3 belong to the characteristic choice, and the step 4 belongs to the characteristic extraction.

To the above step computed result analysis, through Figure 1 may discover the Zernike moment $Z_{5,3}$ 、 $Z_{8,2}$ 、 $Z_{10,0}$ 、 $Z_{11,3}$ has the enormous similarity. Figure 2 expression screening Zernike moment $Z_{5,3}$, $Z_{8,4}$, $Z_{8,6}$, $Z_{10,4}$, $Z_{11,7}$ company graph, two chart contrasts not difficult to discover that, After the screening characteristic has the enormous separation property.

In practical computation, as the data magnitude is dissimilar, we must carry on the normalized processing of the primitive Zernike characteristic. The data normalization is a bit difficult. This article uses is various characteristics divides its average value to achieve the magnitude the normalization.

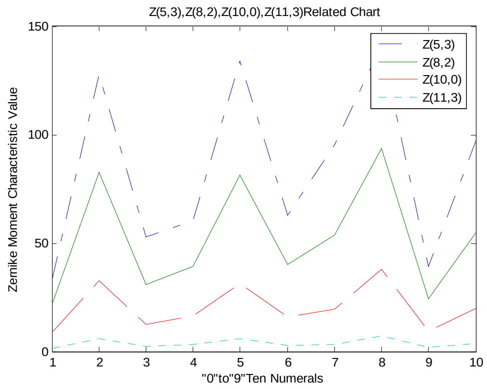


Figure 1 $Z_{5,3}$、$Z_{8,2}$、$Z_{10,0}$、$Z_{11,3}$ related chart

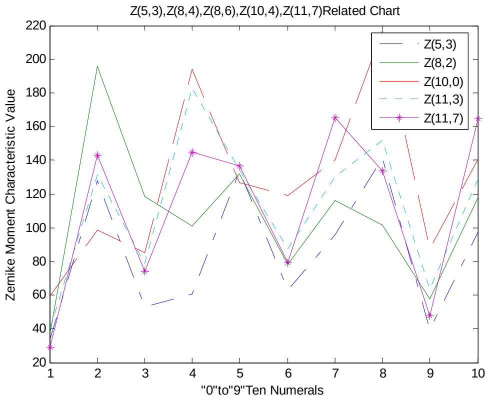


Figure 2 After screening Zernike moment segment

3 Based on Zernike Moment Fuzzy C Means Cluster

The fuzzy c-means Clustering algorithm(FCM), which is widely used in classifications. An objective function J_m is defined as follows:

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

Where $U\in M_{fc}$,

$(d_{ik})^2=\|x_k-p_i\|_A=(x_k-p_i)^T A(x_k-p_i)$, and $P=(p_1,p_2,\cdots,p_c)^T\in R^{cp}$, which is cluster center vector. The positive $(d_{ik})^2$ is a kind of distance between $k-th$ vector x_k and $i-th$ cluster center vector p_i; the positive definite symmetric matrices A is deciding the matrix. $m\in(1,2,\cdots,\infty)$, m means a smoothing weight. The u_{ik} is membership of the $k-th$ data point in $i-th$ class.

The cluster criterion for takes is $J_m(U,P)$ Minimum $\min\{J_m(U,P)\}$.

Because in matrix U each row all is independent, therefore

$$\min\{J_m(U,P)\}=\min\left\{\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^m(d_{ik})^2\right\}$$
$$=\sum_{k=1}^{n}\min\left\{\sum_{i=1}^{c}(\mu_{ik})^m(d_{ik})^2\right\}$$

The extreme value of above equation with constraint condition is equality blow:

$$\sum_{i=1}^{c}\mu_{ik}=1$$

Its solution by using Lagrange's method of multipliers is blow:

$$① \quad \mu_{ik}=\frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} \qquad \text{When } I_k=\varphi$$

$$\mu_{ik}=0,\forall i\in\bar{I}_k,\ \sum_{i\in I_k}\mu_{ik}=1 \qquad \text{When } I_k\neq\varphi,$$

$$② \quad p_i=\frac{1}{\sum_{k=1}^{n}(\mu_{ik})^m}\sum_{k=1}^{n}(\mu_{ik})^m x_k$$

Regarding $\forall k$, Definition set I_k and $\bar{I}_k$,

$I_k=\{i\,|\,1\le i\le c, d_{ik}=0\}$; $\bar{I}_k=\{1,2,\cdots,c\}-I_k$

The objective of the clustering is to minimize the objective function with respect to the partition matrix and cluster center. This kind of optimized question may use the Iterative algorithm to solve.

From the point view of objective function, we generally use the weighted Euclid distance. This distance is suitable for each independent component. The Zernike moment orthogonality has happened to satisfy its request. The traditional numeral recognition methods do not have the orthogonality. We think it is not proper to apply FCM cluster.

We obtained 2000 hand-written numeral samples from the internet（http://www.ics.uci.edu/~mlearn）, in order to establish the standard handwritten numeral storehouse. Each numeral has 200 different hand-written numeral samples and we carry out the FCM cluster separately to them. Take numeral “4” as an example. Table 1 is based on Zernike moment FCM cluster result.

Seeing from Table2, we can conclude that the screened Zernike moments cluster result is better than 47 Zernike moments cluster result.

Table 1 The FCM cluster result Based on Zernike moment which has not been screened (only to list partial samples)

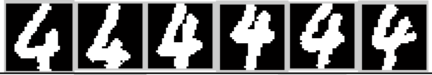
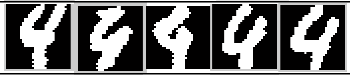
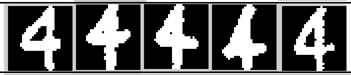
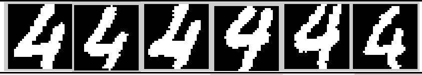
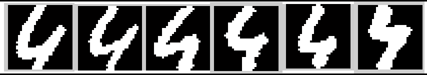
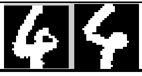
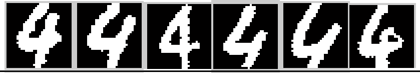
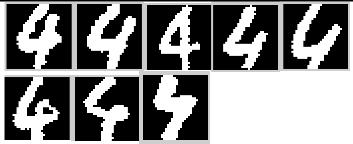
Category	Cluster sample
1	
2	
3	
4	
5	
6	

Table 2 Based on 47 Zernike moments and the screened Zernike moment cluster center contrast

47 Zernike moments	
The screened Zernike moment	

In the sample library, may be recognized as“4” or “9”. So is unrecognizable which will be put in the unrecognizable library. About unrecognizable library, this question will be discussed later.

In the recognition process, each digital image transformation is the 32*32 element constitution two value image. In this article, takes the matching valve value E_f=0.85, “0”~“9” The cluster sample number is[3、4、3、3、5、4、3、4、5、3]. Because the cluster sample is selects through the matching, therefore the cluster counts c=10.After the principal components analysis Zernike moment characteristic took the FCM cluster the characteristic vector, in the computation takes q=1.1, e=0.001. After 12 iterations, classified coefficient $F_c(R)$ = 0.9645, average fuzzy entropy $H_c(R)$ = 0.0812.Explained the cluster effect is good.

Carries on the recognition to 2000 samples, the recognition uses the Euclid matching. Above asks the best cluster center matrix is $V^* = [V_1^*, V_2^*, \cdots, V_{10}^*]^T$, regarding treats recognition sample A, if

$$\sigma(A, V_i^*) = \max_{1 \le j \le 10} \left\{ 1 - \left(\frac{1}{47} \sum_{s=1}^{47} | A(s) - V_j^*(s) | \right) \right\}$$

Then belongs to sample A the kind i .Table 3 has given the recognition result.

Table 3 uses the fuzzy clustering analysis the experimental result (ε =0.001)

Pattern	47 zernike moment recognition			8 zernike moment recognition			After principal components analysis zernike moment recognition		
	Accuracy	Probability of misrecognition	Resists to know rate	Accuracy	Probability of misrecognition	Resists to know rate	Accuracy	Probability of misrecognition	Resists to know rate
0	86.27	10.03	3.70	94.27	1.03	2.70	98.00	1.00	1.00
1	84.29	11.05	4.66	95.65	2.15	2.20	97.50	1.50	1.00
2	76.57	18.23	5.20	95.79	2.49	1.72	97.50	1.50	1.00
3	80.46	15.53	4.01	96.12	0.93	2.95	99.00	1.00	0.00
4	73.28	20.25	6.47	95.54	3.34	1.12	96.00	3.00	1.00
5	88.97	9.53	2.50	95.66	1.24	3.20	98.00	1.00	1.00
6	83.39	10.23	6.38	96.41	2.09	1.50	97.50	1.00	1.50
7	87.24	8.73	4.03	95.86	1.62	2.52	98.00	1.00	1.00
8	82.57	14.33	3.10	92.82	4.36	2.82	95.00	1.00	4.00
9	83.39	10.23	6.38	96.41	2.09	1.50	97.50	1.00	1.50
mean	82.643	12.814	4.643	95.453	2.134	2.223	97.4	1.3	1.3

4 FCM Problems

We have discovered two problems in the establishment of standard sample library during the application of FCM Cluster based on Zernike Moment. （1）regarding the same number with different hand-written sample FCM cluster, because the separation distance is so small that F_C is little, which is the weight of cluster affection, thus the recognition object will be quite close and the cluster effect will reduce. （2）when regarding the different digital recognition, although the Zernike moment degree of separation is big, but along with the sample size increase, it can also produce the morbid state matrix, and the FCM cluster effect will also be influenced. We will publish works on this issue later.

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] D.Cheng and H.Yan,Recognition of handwritten digits based on contour information, Pattern Recogni- tion,1998,31（3）:235~255

[2] Teague M R.Image analysis via the general theory of moments. OptSoc Am, 1980, 70（8）:920-930

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

[4] TONG Ji-jin,LIU Zhong TIAN Xiao-dong.An Application of Neutral Network Based on Zernike Moments and Rough Sets for Number Recognition, Electrical Measurement and Instrumentation. 2005, 12:50-53

[5] Khotanzad A,Hong Y H.Rotation invariant image recognition using features selected via a systematic method .pattern Recognition.1990,23（10）:1089-1101

[6] XU Danhua Gu Jia LI Songyi Shu Huazhong Fast algorithm for computation of Zernike moments. JOURNAL OF SOUTHEAST UNIVERSITY 2002,3: 189-192

Handwriting Verification Based on Fusion of Feature Extracting Algorithms

Mingge Li[1,a] Hui Wang[2,b] Yuzhen Zhong[1,a] Kai Zheng [2,b]
Xiangang Chen[1,a] Hongwei Zhao[2,b]

1 School of Information Technology, Changchun Vocational Institute of Technology, Changchun 130033, China
2 College of Computer Science and Technology, Jilin university, Changchun 130025, China
Email: [a] limingge66@163.com; [b] email_wanghui@126.com

Abstract

The handwriting verification information fusion algorithm based on advanced majority rule was put forward. Several methods of the off-line, text-dependent writer verification were combined with that fusion algorithm, and the ideal of giving the similarity degree of two handwritings while judging was proposed. We also give calculation method of the similarity degree. The result of combination and the similarity degree of two handwriting images can assist document examiners in making determination. Document examiners can make the final determination according to the request of specific application and the confidence degree of the system. Experimental results indicate that compared with current methods, verification accuracy is greatly raised with this approach.

Keywords: Computer Application， Writer Verification, Handwriting Verification， Information Fusion, Similarity Degree

1 Introduction

Writer verification based on handwriting is the task of determining whether or not a handwritten text has been written by a certain person[1]. Usually, people regard the handwriting as an important personal sign of the writer. Writer verification based on handwriting has a great significance in many real-world applications, such as bank checks, case solves, and etc. Handwriting-based writer verification can be classified into on-line (also called dynamic) and off-line (also called static) writer identification based on different input methods[2]. The former assumes that a transducer device is connected to the computer, which can convert writing movement into a sequence of signals and then transmit this signal sequence into computer. And the latter usually deals with handwriting materials directly scanned into computer and thus much dynamic information of writing process is lost. As a result, despite continuous effort, off-line handwriting-based writer verification still remains highly challenging research issue. Further, the off-line writer identification can also be divided into two types: text-dependent and text-independent[3]. Text-dependent methods only match the same character but text-independent methods do not.

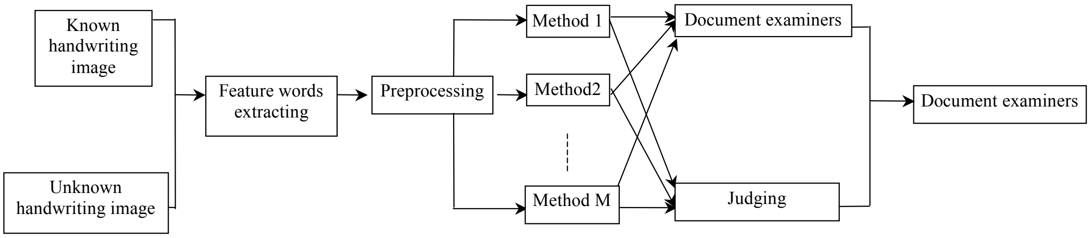


Figure 1 Writer Verification Process

This research was supported by Specialized Research Fund for the Doctoral Program of Higher Education of China(NO20050183032)

Our research work focuses on the off-line, text-dependent writer verification based on handwriting. In the last several decades, some researchers have touched this field[4]. Unfortunately, the appraisal effect of a single several decades, some researchers have touched this algorithm is not very ideal. Considered several methods of off-line, text-dependent writer verification based on handwriting reflected the handwriting characteristic separately from the different aspect, they are complementary in a certain degree. To improve the accuracy of the verification, we can extract features of a word using kinds of methods, and integrate the result of each one with feature parameter, give the verification result finally.

2 Writer Verification Process

It is known that computer handwriting-based writer verification can't substitute for documentary examiners completely but only can assist in giving determination. While the traditional computer methods only give final result with so many details lost, the author proposed computing similarity degree of two handwritings while judging at the same time which realized in the part of similarity computing module.

The writer verification process is as figure 1 shown, we extract features of two handwriting images which have been preprocessed with M methods of writer verification, then match them and calculate match distances, input the match distances to the similarity computing module and judging module. In similarity computing module, similarity degree of two handwritings will be given with the certain algorithm. In judging module, the system carries on the synthesis determination in form of advanced majority rule which be proposed, and produces result of the machine determination. Finally the system outputs two parts of results together, which can assist document examiners in making determination. Document examiners can make the final determination according to the request of specific application and the confidence degree of the system.

On the following part, we explain the application methods of feature extracting and matching, multiple classifiers combination on separately.

3 Feature Extracting and Matching

Suppose there are two handwriting images X,Y ,with N same characters between them. This determination is a classification problem with two classes, w_1 which means two documents are written by the same person and w_2 which means not.

We extract features from N same characters of two handwriting images with method k in set of M. And then obtain feature vectors $x_i^k=(x_{i1}^k,x_{i2}^k,\cdots,x_{in}^k)^T$ and

$y_i^k=(y_{i1}^k,y_{i2}^k,\cdots,y_{in}^k)^T$ $(i=1,2,\cdots N)$ $k=1,2,\cdots M$,

n is the number of features which extracted from character with this method. Then calculate the distance d_{ki} in view of the different method with different match algorithm, and design classifiers appropriately.

Let us explain this process with multi-channel decomposition and match method[5] for instance:

1) First, we can obtain 336 length feature vectors from same words in two handwriting images with MCD method. They are $x_i=\left(x_{i1},x_{i2},\cdots,x_{i336}\right)^T$ and $y_i=\left(y_{i1},y_{i2},\cdots,y_{i336}\right)^T$

$\left(i=1,2,\cdots N\right)$.

2) Then distance d_i $\left(i=1,2,\cdots N\right)$ between feature vectors can be calculated according to Eq.(1) and it is normalized,

$$d_i\left(x_i,y_i\right)=\frac{\sum_{j=1}^{256}\left|x_{ij}-y_{ij}\right|+4\sum_{j=257}^{320}\left|x_{ij}-y_{ij}\right|+16\sum_{j=321}^{336}\left|x_{ij}-y_{ij}\right|}{\sum_{j=1}^{256}\left(\left|x_{ij}\right|+\left|y_{ij}\right|\right)+4\sum_{j=257}^{320}\left(\left|x_{ij}\right|+\left|y_{ij}\right|\right)+16\sum_{j=321}^{336}\left(\left|x_{ij}\right|+\left|y_{ij}\right|\right)} \tag{1}$$

3) For the within-class and between-class probability distribution of distance measures d is approximate to the normal distribution[6], we can estimate them with normal distribution. Towards feature word i, suppose d_{ij} $\left(j=1,2,\cdots N_1\right)$ are the distances between two handwritings of N_1 same person and $d_{ij}^{'}$ $\left(j=1,2,\cdots N_2\right)$ are the distances between two

handwritings of N_2 different person. With maximum a posteriori discriminate method[7], suppose that the prior probability of two kinds of patterns is equal, then discriminate threshold value t_i is equal to the distance value when conditional probability of two kinds of patterns is equal. So, t_i is the solution of Eq.(2).

$$\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{\left(x-\mu_i\right)^2}{2\sigma_i^2}\right)=\frac{1}{\sqrt{2\pi}\sigma_i^{'}}\exp\left(-\frac{\left(x-\mu_i^{'}\right)^2}{2\sigma_i^{'2}}\right) \tag{2}$$

and $\mu_i=\frac{1}{N_1}\sum_{j=1}^{N_1}d_{ij}$, $\sigma_i=\frac{1}{N_1}\sum_{j=1}^{N_1}\left(d_{ij}-\mu_i\right)^2$,

$$\mu_i^{'}=\frac{1}{N_2}\sum_{j=1}^{N_2}d_{ij}^{'},\ \sigma_i^{'}=\frac{1}{N_2}\sum_{j=1}^{N_2}\left(d_{ij}^{'}-\mu_i^{'}\right)^2$$

Here, we estimate correct rate of classifier c_i and error rate f_i by Leave-one-out method[7] for behind.

We can design many classifiers for other feature extracting methods by the same way. Each classifier's threshold value and it's correct rate all may use the above method to obtain.

4 Information Fusion

4.1 Similarity computing module

In this module, total similarity degree of two handwriting images will be calculated with distance measures received before. The process is as follows:

1) Produce confidence matrix according to the method of calculate c_i in 3 section

$$C=\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \vdots & \vdots & & \vdots \\ c_{M1} & c_{M2} & \cdots & c_{MN} \end{bmatrix} \tag{3}$$

c_{ki} is the correct rate of classification of feature word i $\left(1\le i\le N\right)$ based on method k$\left(1\le k\le M\right)$. The bigger of the value represents the more correct of the result of classification. Weighting with matrix C when we are calculating similarity degree of two handwriting images can avoid bad affection of some individual instable character and therefore get more accurate result.

2) Suppose the distance matrix of two handwriting images is

$$D=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & & \vdots \\ d_{M1} & d_{M2} & \cdots & d_{MN} \end{bmatrix} \tag{4}$$

It's easy to see that the smaller of the distance of two handwriting images, the bigger of the similarity degree. So the similarity matrix which be weighted is

$$A=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ a_{21} & a_{22} & \cdots & a_{2M} \\ \vdots & \vdots & & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MM} \end{bmatrix}=C\cdot\left(1-D\right)^T \tag{5}$$

The element a_{kk} on the diagonal of matrix A is the similarity degree of two handwriting images which be calculated by method k. The sum of all similarity degrees of two handwriting images divided by the sum of confidence degrees is equal to the total similarity degree.

3) Total similarity degree

$$S=\frac{\sum_{k=1}^{M}a_{kk}}{\sum_{k=1}^{M}\sum_{i=1}^{N}c_{ki}} \tag{6}$$

4.2 Judging Module

In this module, the determination of computer which is used to assist document examiners in making determination will be given by writer verification information fusion based on advanced majority rule proposed. The method of writer verification information fusion based on advanced majority rule is as below: Firstly, we make each classifier gives a determination result r_{ki} (see Eq.(7)), then have liner weight on two types of patterns w_1 and w_2 with confidence degree. Finally, the determination of computer depends on the result of determination function which is bigger.

$$r_{ki}=\begin{cases}1, & d_{ki}\le t_{ki} \\ 0, & otherwise\end{cases} \tag{7}$$

$k=1,\cdots,M$, $i=1,\cdots,N$.

If $d_{ki}\le t_{ki}$, the handwritings is written by the same person, it belongs to pattern w_1,so the determination function of w_1 is

$$f_1\left(X,Y\right)=\sum_{k=1}^{M}\sum_{i=1}^{N}c_{ki}r_{ki} \tag{8}$$

Appropriately the determination function of w_2 is

$$f_2\left(X,Y\right)=\sum_{k=1}^{M}\sum_{i=1}^{N}c_{ki}\left(1-r_{ki}\right) \tag{9}$$

At last, it's two-class recognition problem, if $f_1\left(X,Y\right)\geq f_2\left(X,Y\right)$， X,Y are written by the same person; else they're written by the different persons.

5 Experimental Results

Considered geometric moment method[5] can reflect the shape of characteristic well, distance transform method[5] can reflect the similarity degree of two handwritings from the distribution of pen delimits, multi-channel decomposi- tion and match method[6] can reflect texture feature. The article extract features with this three methods of writer verification. Test them in same situation separately. Meanwhile design classifier appropriately and estimate error rate of classifier.

In the experiment, we use handwriting images written by 15 persons which contained 109 same persons handwritings and 198 different persons handwritings, there are more than 8 same characters between each two handwritings. Take multi-channel decomposition and match method as instance, eight feature words and their error rate as shown in table 1.

Table 1 Feature words and their error rate with multi-channel decomposition and match method

Words	Type Ⅰ error	Type Ⅱ error
"义"	7.4%	15.6%
"大"	1.0%	27.5%
"学"	7.4%	15.6%
"法"	10.6%	19.0%
"的"	18.7%	8.8%
"张"	13.9%	5.4%
"日"	4.2%	25.8%
"敬"	12.3%	13.9%

From table 1 we can see that type one error and type two error of different feature words are large different, this may because of the extent of the changes in writing, and it is impossible to know in advance, the only solution must be integrate a number of words results to offset this effect；On the other hand, data given in table indicates that the error rate of one word using in handwriting verification is high, this is because changes of one word when written and similar part of one word written by different person can easy be confused. From above, determine the identity of writer with only a word, it certainly will not be too accurate. The results of verification with the above eight feature words as shown in table 2.

Table 2 results of multi-channel decomposition and match method with the combination of eight feature words

Correct rate	Type Ⅰ error	Type Ⅱ error
92.4%	3.5%	4.1%

Fuse the results of three methods with advanced majority rule (M=3,N=8)，estimate error rates.

The results as shown in table 3 under the circumstance of not considering refused rate. The experimental data is available by closed set test, therefore the estimate of error rate is optimistic.

The result indicated that method in this article is much better than the single method (only thinking about determination by the machine in the information association module),in such a condition not thinking about the refusal rate experimental results shows by table 3.

Table 3 Experimental results

Feature extracting method	Geometric moment method	Distance transform method	Multi-channel decomposition and match	The author's method
Error rate	8.26%	12.40%	6.61%	2.65%

6 Conclusion

Through analysis of the experimental data, it's not difficult to discover that we have obtained a better effect when using the method of the multiple classifiers combination. Compared with verification result of the single method, there is a distinct enhancement. To give the similarity degree of the two handwritings with the determination made by the machine, it can provide more information for the document examiners and assist them in making final determination.

References

[1] R. Plamondon and G. Lorette. Automatic signature verification and writer identification-the state of the art[J]. Pattern Recognition, 22 , 1989 , pp.107-131

[2] Zhenyu He,Yuan Yan Tang,Bin Fang,Jianwei Du, Xinge You. A Novel Method for Off-line Handwriting-based Writer Identification [J]. Eighth International Conference on Document Analysis and Recognition (ICDAR'05), Seoul, Korea, 31 August to 1 September 2005, pp. 242-246

[3] H.E.S.Said, T.N.Tan. Personal identification based on handwriting [J]. Pattern Recognition, vol.33, no. 1(7) , January 2000, pp.149-160

[4] Y.Zhu and T.Tan. Biometric personal identification based on handwriting [J]. Proc. 15th International Conference on Pattern Recognition, Barcelona, Spain, Sep3-7 2000, pp.801-804

[5] Liu Chenglin,Lu Yingjan,Dai Ruwei. Writer Identifi- cation By Multi-channel Decomposition and Matching [J]. Acta Automatica Sinica, 23(1), 1997, pp. 56-63

[6] Liu Chenglin. The Theory and Method of Computer Writer Identification[D].Beijing：Institute of Automation, Chinese Academy of Sciences, 1995, pp. 29-53

[7] Bian Zhaoqi, Zhang Xuegong. Pattern Recognition [M]. Second edition. Beijing: Tsinghua University Press, 2000, pp. 25-34.

[8] Kenneth R.Castleman.Digital Image Process,Peking Electronic Industry Press 2003,pp.66-90

[9] Yong Zhu,Tieniu Tan,Yunhong ,Identification Based on Handwriting, Journal of Automation,2001,pp.20-45

[10] GuanqingWang,The New Exploration of the Theorical and Pratical Handwriting Identification,Peking University Press,2003,pp.80-118

State-of-the-art on Cluster Analysis of Gene Expression Data

Qianqian Gao Jun Sun

School of Information Technology, Jiang Nan University, Wuxi, 214122, China
Email:missgaoqian@163.com

Abstract

With the development of DNA technology, there are huge volumes of gene expression data to be generated. How to effectively organize and analyze these data has become an urgent problem to be solved. At present, cluster analysis is an effective and practical tool to analyze the flood of gene expression data to gain important biological and genetic information. In recent years, many improved conventional clustering algorithms as well as new clustering algorithms have been proposed to process the gene expression data. This paper simply introduces how to produce gene expression data firstly, and then discusses some new cluster algorithms applied in gene expression data. For gene-based clustering, we present advantages and disadvantages of its methods in detail and simply introduce sample-based clustering and biclustering.

Keywords: DNA microarray, gene expression data, cluster analysis

1 Introduction

Due to advancement of DNA technology, the massive of gene expression data are produced, which needs automated analysis techniques and tools to solve this problem. Cluster analysis [1] as an efficient data analysis and exploratory tool has been applied to data mining, image processing, information retrieval and other fields. Now, it has a wide range of applications in analysis of gene expression data. For example, through clustering of genes, we can find out functions of unknown gene; though clustering of samples, we could know phenotype structure of samples and it automatically classifies pathology characters or experimental conditions by it; through biclustering, we may find some genes that involve in regulating under some conditions.

The rest part of the paper is organized as follows. Section 2 gives a brief introduction for obtaining of gene expression data, followed by Section 3 presenting some new algorithms for gene-based clustering. In Section 4, we simply describe sample-based clustering and biclustering. The paper is concluded in Section 5.

2 Acquisition of Gene Expression Data

Gene expression data is accessed mainly through technologies that cDNA [2, 3, 4] microarrays and oligonucleotide microarrays(also known as gene chip, DNA chip)[5,6,7] . It reflects the abundance of cells by direct or indirect measurements.

Schean put forward the gene chip technology firstly, which principles refer to use many specific oligonucleotide fragments or gene fragments which compose DNA microarray as probes and arrange them on the fixed supporting thing. According to pairing cross bred principle, DNA microarray hybridizes with marked fluorescent that has not been measured. To scan the chip with confocal laser fluoroscope system, then, comparing and testing fluorescence signal of each probe with computer systems. Gene expression level and biological information can be obtained from this kind of method.

cDNA microarray that have successfully applied in gene expression data is firstly proposed by Stanford university in 1993. Its principle is that mRNA of cells retrovirus into cDNA, which is sampled on the membrane or glass by Arrayer. During this course, the size of samples can'tguarantee all the same and the order of the points may not in rules. So, we can not directly compare the fluorescence intensity of different points. At preparing the samples, we need to use two samples, one is a control sample whose cDNA commonly marked by green fluorescence (cy3) and another one is measurement sample whose cDNA used red fluorescence (cy5) to mark. Then, using this two marked cDNA to hybridize with cDNA microarray and scanning with laser the microarray that have hybridized, which we record the fluorescence intensity from and compare the relative gene expression levels.

In this paper, the gene chip and cDNA microarray all called DNA microarray, excepting for special circumstances. Gene expression data derived from the DNA microarray can't directly apply in cluster analysis. Because there are negative or incomplete data, as well as obvious noise in gene expression level. Sometime we need to estimate missing data [8-10]. During the data processing, normalization of data is necessary and we don't describe strictly here. In this context, let gene denote the data that is measured in experiment and samples as the experimental conditions.

Clustering is the process of grouping the data into classes or clusters. The elements or objects within clusters have high similarity in comparison to one another, but are very dissimilar to objects in other clusters [11]. As the amount of experimental data in molecular biology grows exponentially each year, new efficient and effective clustering algorithms must be proposed to process this growing amount of biological data. According to different objects and goals of gene expression data in processing, clustering algorithms can divide into three categories: gene-based clustering, sample-based clustering and biclustering, which will be introduced one by one.

3 Gene-Based Clustering

The meaning of gene-based clustering is that it refers genes as objects to be clustered and samples as characters of genes. Though gene clustering, we can find out genes which have similar expression patters. As we known, genes that lie in the same class may have the same functions. So, we can infer functions of unknown genes from the known genes [12]. With the help of traditional clustering algorithms, such as, K-means, Hierarchical clustering, SOMs and so on genes are partitioned into groups based on the similarity between their expression profiles. Of course, there are many new cluster algorithms produced, such as Artificial Neural Network-based clustering, fuzzy clustering and other novel clustering algorithms.

3.1 Artificial Neural Network-Based clustering

Self -organizing Maps (SOMs), invented by Teuvo Kohonen, which reduce the dimensions of data through the use of self-organize neural networks. They reducing dimensions is by producing a map of 1 or 2 dimensions which plot the similarities of the data by grouping similar data items together [13]. SOMs have advantages and disadvantages. The best thing about SOMs that they are very easy to understand, another advantage is that they work very well. But, there is a problem that every SOM is different and finds different similarities among the sample vectors. Another problem is that you need a value for each dimensions of each member of samples in order to generate a map [13].

The Self-organizing Tree Algorithm (SOTA) [14] is based both on the SOM and the growing cell structures [15]. For one good thing of SOTA is that the binary topology produced a nested structure in which nodes at each level are averages of the items below them; which makes it straight forward to compare average patters of gene expression at different hierarchical levels even for large data sets and has more robust. According to different neurons defining the different hierarchical levels represent the averages of the gene expression patters contained in the clusters. Despite the advantages that SOTA presents, but the output structures of SOTA is binary tree, which will contain the structure of data presenting only is binary tree.

In literature [16], the author proposed a new tree-structure Dynamically Growing Self-organize Tree (DGSOT) that overcomes drawbacks of SOTA. The DGSOT algorithm is a tree structure self-organize neural network designed to discover the proper hierarchical structure of the underlying data, not confine in binary tree. The DGSOT grows vertically and horizontally and the complexity of it is $o\left(m\log_d^m\right)$, where m is the number of data, d is the branch factor of the DGSOT. The DGSOT algorithm combines the horizontal growth and vertical growth to construct a mutlifurcating hierarchical tree from top to bottom to cluster the data. In each vertical growth, the hierarchical tree expands one more hierarchical level. In addition, in each hierarchical level the DGSOT employs a horizontal growth step to optimize the number of the subclusters [16]. Compared to other self-organizing algorithms, DGSOT has a larger improvement on efficiency.

3.2 Fuzzy Clustering Algorithms

Many proteins serve different roles depending on the demands of the organism, and therefore the corresponding genes are often coexpressed with different groups of genes under different situations [17]. In other words, gene may have different functions in different conditions.

Fuzzy clustering method can able to identify clusters of genes that were not identified by hierarchal or standard K-means clustering. The fuzzy clustering algorithms generate a fuzzy partition providing a degree of membership u_{ij} of each data X_i to a given cluster F_j.Since fuzzy algorithm makes soft decisions in each of iteration through the use of membership functions, it is less prone to local minimum than crisp clustering algorithms [18]. The most widely used fuzzy algorithm is Fuzzy C-means (FCM) algorithm that proposed by Bezdek, 1981. The principal of FCM is similar to K-means, but FCM needs to consider the membership when it calculates centroid of each cluster. We suppose the number of genes is m and the number of clusters is *K*. Firstly, initializing the matrix $U=[u_{ij}]$, where u_{ij} denotes the value of membership that gene X_i to the cluster $F_{j,}$ $1\le i\le m$, $1\le j\le k$ $0\le u_{ij}\le 1$. Secondly using

$$c_j=\left(\sum_{i=1}^{m}u_{ij}^q x_i\right)\Bigg/\left(\sum_{i=1}^{m}u_{ij}^q\right) \tag{1}$$

to calculate centroid c_j of each cluster, where q is fuzziness parameter, then, using the new centriod to calculate the membership

$$u_{ij}=1\Bigg/\left(\sum_{s=1}^{k}\left(\frac{d\left(x_i,c_j\right)}{d\left(x_i,c_s\right)}\right)^{\frac{2}{q-1}}\right) \tag{2}$$

of each gene of cluster. When the membership matrix changes little, it stops to repeat. Now, many literatures for example, in [17], [18], [23], use FCM algorithm in gene expression data analysis.

In [17], the author used a heuristically modified version of fuzzy K-means clustering to identify overlapping clusters of yeast genes and post the different traits of function and regulation of each gene. It reflects the relationships between transcription factor and environmental conditions. According to membership of genes, they are arranged in rules. In addition, compared to K-means algorithm, this algorithm isn't too sensitive to K and it isn't affect performance of the algorithm, if the value of K isn't too larger. The algorithm also has several flaws: (a). the algorithm couldn't identify all the clusters. In dealing with gene expression data of yeast, the algorithm identify out 90% of the known clusters and couldn't guarantee recognize clusters that identified by hierarchical clustering. (b). It's difficult to design the parameter of the algorithm. If the input data is cosmically or the number of clusters too big, the capability of FCM will distinctly decline.

The authors think that a major problem in applying the FCM method for clustering microarrary data is the choice of the fuzziness parameter m, in literature [19]. In this paper, it shows that commonly used value $m=2$ doesn't appropriate for some data sets and that optimal values for m vary widely from one data to another. The main contribution of this paper is that give a method to design fuzziness parameter m. In classical FCM, it need *m*>1, but in [19], it needs m between 1and 2 and it's more better, if the value of m is close to 2.

The classical FCM algorithm and its variants require the user to pre-define the number of clusters. Selections of a different number of initial clusters result in different clustering partitions. Thus, it is necessary to validate each of the fuzzy *c*-partitions once they are found. This evaluation is called cluster validity [18]. In this paper, a new cluster validity index for fuzzy clustering is proposed. The proposed validity index exploits an inter-cluster proximity between fuzzy clusters. The inter-cluster proximity is used to measure the degree of overlap between clusters. A low proximity value indicates well-partitioned clusters. Through experimental testing, the performance of the proposed index on various data sets demonstrates good effectiveness and reliability.

It's well known that fuzziness parameter m is difficult to design. In [20, 21], which give a theoretical approach to choose the appropriate fuzziness index. The experimental results show that these theoretical rules are effective.

Fuzzy J-Means (FJM) and Variable Neighborhood Search (VNS) [23] were presented to process gene expression data in [22]. FJM has been recently developed (Belacel, et al. 2002). The FJM method was inspired by the local search heuristic J-Means, developed for solving the minimum sum-of-squares clustering problem (Hansen and Mladenovic, 2001). In FJM, defining a goal function

$$\left(\min_{V}\right) R_m\left(V\right) = \sum_{i=1}^{n}\left[\sum_{k=1}^{c}\left\|X_i - V_k\right\|^{2(1-m)}\right]^{(1-m)} \tag{3}$$

Where $V = \left\{v_{1,} v_2......v_k\right\}$ is a set which is composed by centriods of K clusters, $\left\{X_1, X_2......X_n\right\}$ is a set composed by gene expression vector. Both F-CM and F-JM are local heuristics, which can not guarantee that the final result is the overall optimal clustering solution, even when using several different starting points. The VNS is a previously developed Metaheuristic for solving combinatorial and global optimization problems (Hansen and Mladenovic 1997). The principal of VNS is similar to genetic algorithm, through randomly chosen initialized input of FCM, new input are give FJM to compute the optimize solutions. The stopping criterion may be set either to the maximum CPU time or a maximum number of iterations allowed. The innovation of [26] is that Fuzzy J-Means, embedded into the Variable Neighborhood Search metaheuristic for the clustering of microarray gene expression data, which effectively improve the performance of the algorithm, but largely increase computational complexity.

The conventional methods are limited to identify different shapes of clusters because they use a fixed distance norm when calculating the distance between genes. The fixed distance norm imposes a fixed geometrical shape on the clusters regardless of the actual data distribution. Thus, different distance norms are required for handling the different shapes of clusters [24]. The authors present the Gustafson-Kessel (GK) clustering method for microarray gene-expression data to detect clusters of different shapes in a dataset in [24].

3.3 Others Clustering Algorithm

There also are novel clustering algorithms in [25] [26][27] to analyze gene expression data. The main idea of the proposed algorithm [25] is to avoid the problems and disadvantages of the commonly used clustering algorithms. In this paper, the authors try to overcome the problem that most of the used clustering is the full dependence on a cluster center as the represent metaheuristic ive of each cluster including all the data points in it rather than using a single cluster center as the cluster prototype. The other new clustering algorithm is Genetic Weighted K-means Algorithm (denoted by GWKMA), which is hybridization of a genetic algorithm (GA) and a weighted K-means algorithm (WKMA). In GWKMA, each individual is encoded by a partitioning table which uniquely determines a clustering, and three genetic operators (selection, crossover, mutation) and a WKM operator derived from WKMA are employed [26]. The results of the GWKMA perform better than the k-means in terms of the cluster quality and the clustering sensitivity to initial partitions. In literature [27], the distinct characteristic of new algorithm is that it integrates the validation technique to improve the quality of clustering and Principal Component Analysis to reduce the dimensionality of the data set to the clustering process [27]. One contribution of this algorithm is that it incorporates computation intelligence technique to classify gene expression data efficiently.

4 Clustering of Samples and Biclustering

4.1 Clustering of samples

Clustering of samples refers to let genes as traits and samples as objects for clustering. Through clustering of samples, we can discover the phenotype structure of samples, which is used to classify automatically experimental conditions or pathological features. The most important thing is that we can find gene regulation mechanism under different pathological features or experimental conditions.

At literatures [28], [29], they discuss the problem about model-based approach for sample- based clustering. The principal of them is almost similar. In [30], it adds principal component analysis (PCA) to reduce dimension. However, it uses factor analysis to reduce dimension. The procedures followed: firstly, selecting genes that correlate with samples. Secondly, using the algorithm of literature [30] to cluster the genes that have larger relationship with samples and clustering to samples with genes of cluster, which take the method that mixtures of factor models [31],[32]., then, calculating average vector of each gene which is used to cluster samples. In literature [33], it proposed a algorithm for clustering of samples, which is more accurate in finding the true number of clusters in situations that are relevant to current and future microarray studies. Compared to method based on EM+BIC, variational bayesian mixture model are more accurate to cluster.

4.2 Biclustering

Biclustering also be called subspace clustering, coclusteting or direct clustering, which refers to clustering of genes and samples at the same time. Due to only certain genes that lie in certain conditions involved in biological processes need to be checked. So, the goal of biclustering is to find out conditions that make some genes to associate with gene cluster that involve in regulation.

It is a NP problem that makes biclustering optimal in gene expression data. Hence, proposed biclusteing algorithm mainly is heuristic. In the course of biclustering, genes and samples are considered equally. Some algorithms of biclustering, such as Coupled Two-Way Clustering (CTWC)[33], [34], plaid model [35] have successfully using in gene expression data analysis. Now, a new cluster method, projected clustering have also used in gene expression data. Projected clustering firstly proposed by Aggarwal [36], its principal is that objects projected in certain dimension firstly, then to cluster these objects.

HARP (Hierarchical approach with Automatic Relevant attribute selection for Projected clustering) is proposed by Kevin [37], [38]. In projected clustering, clusters exist in subspaces of the input space defined by the dimensions of the data set. The similarity between different members of a cluster can only be recognized in the specific subspace. A data set can contain a number of projected clusters, each form in a possibly distinct subspace [37]. In HARP, it needs to define relevance index of each property that belongs to cluster, the bigger relevance index, the more partional likelihood of the property with other clusters. Merge score is used to measure pros and cons of the two clusters merged into one cluster. The algorithm firstly think that every object is a cluster, then calculating merge score of two clusters and merging two cluster that there are the biggest merge score. Finally, computing the relevance index of each property of cluster which is assigned as subset of the cluster, if the value of property larger than given threshold $R_{\min}$. Repeat these procedures until one cluster to remain. Experimental results on synthetic and real data suggest that HARP has a higher accuracy and usability than the projected and nonprojected algorithms being compared, and it remains highly accurate when handling noisy data [37]. But, HARP doesn't overcome the fatal drawbacks: easily affected by noises of hierarchical algorithm.

5 Conclusion

From these new algorithms in recent years, we can see cluster algorithms of gene expression data are no longer remaining in the early traditional algorithms. People pay more attention to new algorithms which are more suitable for characteristics of gene expression data. The main purpose to research the methods of gene expression data is to provide a tool that used to mine gene expression data for biological researchers. So, the flexibility and easy using of algorithm are the key issues to be considered.

With the large number of gene expression data to be generated and more cluster algorithms are proposed, biological researchers face an important issue is how to choose reasonable algorithm. In fact, there isn't a absolute optimize algorithm and standard used to evaluate the capability of a algorithm.

The ultimate goal of cluster analysis is to help user to access more biological information, so, when we select cluster algorithm and analyze results of clustering, we should make full use of known biological knowledge and some effective aids, of course, good human-computer interaction must be good at analyzing in results of clustering. At present, people commonly use visualization technology such as, heat map, dendrogram to help analyzing the results of clustering. In short, new algorithms will be useful to mine much more valuable information.

References

[1] Jain A K, Murty M N, Flynn P J. Data Clustering: A Review. ACM Computing Surveys, 1999, 31(3): pp.264-323

[2] Schena M, Shalon D, Davis R W, Brown P O. Quantitative monitoring of gene expression patterns with complementary DNA microarray. Science, 1999, 270(5235): pp.467-470

[3] Schena M, Scalon D, Heller R. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proceedings of the National Academy of Sciences (USA), 1996, 93(20): pp.10614-10619

[4] Ramsay G. DNA chips: State-of-the art. Nature Biotechnology, 1998, 16(1): pp. 40-44

[5] Lockhart D J, Dong H, Byrne M C, Follettie M T, Gallo MV, Chee M S et al.Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology, 1996, 14(13): pp.1675-1680

[6] Lipshutz R J, Fodor S P A, Gingeras T R, Lockhart D J.High density synthetic oligonucleotide arrays. Nature Genetics, 1999, 21(1 Suppl): pp.20-24

[7] Harrington C A, Rosenow C, Retief J. Monitoring gene expression using DNA microarrays. Current Opinion in Microbiology, 2000, 3(3): pp.285-291

[8] Hyunsoo Kim, Golub G H, Park H. Missing value estimation for DNA microarray gene expression data: local leasts quares imputation. Bioinformatics, 2005, 21(2): pp.187-198

[9] Tuikkala J, Elo L, Nevalainen O S, Aittokallio T. Improving missing value estimation in microarray data with gene ontology. Bioinformatics, 2006, 22(5): pp.566-572

[10] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R. Missing value estimation methods for DNA microarrays. Bioinformatics, 2001, 17(6): pp.520-525

[11] B.Sathiyabhama,N.P. Gopalan.MINING GENE EXPRESSION DATA USING ENHANCED INTELLIGENCE CLUSTERING AND MEMORY REDUCTION TECHNIQUES DOI 10.1109/ICCIMA.2007.358

[12] Eisen M B, Spellman P T, Brown P O, Botstein D. Cluster analysis and display of genome-wide expression patterns.Proceedings of the National Academy of Sciences(USA),1998, 95(25): pp.14863-14868

[13] Kohonen T. Self-Organizing Maps. Berlin: Springer Verlag, 2001

[14] Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics, 2001, 17(2): pp.126-136

[15] Fritzke B. Growing cell structures-a self-organizing network for unsupervised and supervised learning. Neural Networks, 1994, 7(9): pp.1141-1160

[16] Feng Luo, Khan L, Bastan F, I-Ling Yen, Jizhong Zhou. A

dynamically growing self-organizing tree (DGSOT) for hierarchical clustering gene expression profiles. Bioinformatics, 2004, 19(16): pp.2605-2617

[17] Gasch A U, Eisen M B. Exploring the conditional coregulation of yeast gene expression through fuzzy K-means clustering. Genome Biology, 2002, 3(11): pp. 1-22

[18] Dae-Won Kim，Kwang H. Lee，Doheon Lee，Fuzzy cluster validation index based on inter-cluster proximity.Pattern Recognition Letters November 2003, Pages pp.2561-2574

[19] Dembele D, Kastner P. Fuzzy C-means method for clustering microarray data.Bioinformatics, 2003, 19(8): pp. 973-980

[20] Jian Yu. On the Fuzziness Index of the FCM Algorithms. Chinese Journal of Computers, 2003, 26(8): pp.968-973

[21] Jian Yu, Qiansheng Cheng, Houkuan Huang. Analysis of the weighting exponent in the FCM. IEEE Transactions on Systems, Man and Cybernetics-part B: Cybernetics, 2004,34(1): pp. 634-639

[22] Belacel N, Cuperlovi′c-Culf M, Laflamme M, Ouellette R.Fuzzy J-Means and VNS Methods for Clustering Genes from Microarray Data. Bioinformatics, 2004, 20(11): pp.1690-1701

[23] Belacel N, Hansen P, Mladenovic N. Fuzzy J-Means: a New Heuristic for Fuzzy Clustering. Pattern Recognition, 2002, 35(10): pp.2193-2200

[24] Dae-Won Kim, Kwang H. Lee, Doheon Lee. Detecting clusters of different geometrical shapes in microarray gene expression data. Bioinformatics, 2005, 21 (9): pp.1927-1934

[25] Hossam S. Sharara， Mohamed A. Ismail.CORR: A novel algorithm for clustering gene expression data 2007 IEEE

[26] Fang-Xiang Wu .A Genetic Weighted K-means Algorithm for Clustering Gene Expression Data DOI 10.1109/ IMSCCS.2007.22

[27] B.Sathiyabhama, N.P. Gopalan. MINING GENE EXPRESSION DATA USING ENHANCED INTELLIGENCE CLUSTERING AND MEMORY REDUCTION TECHNIQUES International Conference on Computational Intelligence and Multimedia Applications 2007

[28] McLachlan G J, Bean R W, Peel D. A Mixture Model-based Approach to the Clustering of Microarray Expression Data.Bioinformatics, 2002, 18(3): pp.413-422

[29] Ghosh D, Chinnaiyan A M. Mixture Modelling of Gene Expression Data from Microarray Experiments. Bioinformatics, 2002, 18(2): pp.275-286

[30] Yeung K Y, Fraley C, Murua A, Raftery A E, Ruzzo W L.Model-based clustering and data Transformations for Gene Expression Data. Bioinformatics, 2001, 17(10): pp.977-987

[31] McLachlan G J, Peel D. Finite Mixture Models, 2000

[32] McLachlan G J, Peel D. Mixtures of factor analyzers. In:Proceedings of the Seventeenth International Conference on Machine Learning. San Francisco, USA: Morgan KaufmannPublishers. pp.599-606

[33] Getz G, Gal H, Kela I, Notterman D A, Domany E. Coupled two-way clustering analysis of breast cancer and colon cancer gene expression data. Bioinformatics, 2003, 19(9): pp.1079-1089

[34] Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. Proceedings of the National Academy of Sciences (USA), 2002, 97(22): pp.12079-12084

[35] Lazzeroni L, Owen A. Plaid models for gene expression data.Statistica Sinica,2002,12(1): pp.61-86

[36] Aggarwal C C, Procopiuc C, Wolf J L, Philip S. Y, Jong Soo Park. Fast Algorithms for Projected Clustering. In: Proceedings of ACM International Conference on Management of Data. New York, USA: ACM Press, 1999. pp.61-72

[37] Yip K Y. HARP: A Practical Projected Clustering Algorithm for Mining Gene Expression Data. [Master dissertation], The University of Hong Kong, 2004

[38] Yip K Y, David W C, Michael K N. HARP: A Practical Projected Clustering Algorithm. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(11): pp.1387-1397

A Chinese Web Documents Clustering Method Based on the Suffix Tree

Chiwen Wu

School of Information Technology, Jiangnan University, Wuxi Jiangsu China
Email:dldldl536@hotmail.com

Abstract

An improved Chinese documents clustering method based on suffix tree is reported in this paper. The foundation of the improved method is the keyword, at the same time, joined the POS (part of speech) judgment. The STC (Suffix Tree Clustering) is introduced first, and then, its improved method including its realization details are given, accordingly, a cluster evaluation method is used to measure the performance of the proposed method. The experiments demonstrate that the proposed method can achieve better results than the STC does.

Keywords: STC, Chinese documents clustering, keywords, weight, POS

1 Introduction

With the fast development of Internet technology, people feel the impact which is brought by the information more and more intensely, and the documents are the important carrier of information. About 80% of the information in people’s daily life exists in the form of document. The diversification and complication of the information content and form, also the quick information refreshed rate, causes a problem for people to find out what they are interested. Therefore how to collect, manage the Internet's large information, how to search information which the user needs quickly, become the question which is worth researching. Therefore, the high efficient search engines appear unceasingly, Yahoo, Google, Baidu, and so on. However, because the huge information as well as the current search engine which is based on “key word inquiry”, the results that the search engine returns are numerous, while what the user cares actually is often only a very small part. If we can carry on the classification to the inquiry results, we may facilitate the user rapidly to locate the interest content. At present a search engine called “Meta Search Engine”, can carry on the classification to the search results, with the aim of giving the user the information fast and accurately from Internet's magnanimous information.

As a non-surveillance machine learning method, the cluster technology already became an important method to effectively organize, abstract and guide the information, it attracts attention by more and more researchers. There are many cluster methods, for instance, Agglomerative Hierarchical Clustering, K-Means Clustering as well as Suffix Tree and so on. In these algorithms, the Suffix Tree Clustering is especially prominent in the time complexity, it is one linear algorithm, it starts to work while the search engine returned the first result, demonstrates the result when it receives the last result, the user can hardly feel the obvious sluggish phenomenon, and its rate of accuracy is higher than the classical cluster algorithm. Therefore, this article takes it as the method to realize Chinese documents cluster.

The structure of this article is as follows: The second part is the introduction of the Suffix Tree Clustering, the third part introduces the improvement, the fourth part is the introduction of the evaluation method and the experimental result, the fifth part is the conclusion.

2 Algorithm Introduction

2.1 Related Work

The cluster algorithms can be divided into two kinds approximately: Hierarchical Clustering and Partition Clustering. Although their applications are very widespread, there are also some shortcomings, first: It needs to determine the good condition beforehand, for instance the K-means Clustering needs to determine the K value beforehand. Second: In practical application, it has the possibility that an article is belonged to several different categories, for example an introduction about the army life's soap opera article, may be belonged to the entertainment article, it may also be belonged to the military article. This is the situation which will happen frequently. Taking this into consideration, by Oren Zamir and Oren Etzioni in "Web Document Clustering: A Feasibility Demonstration" proposed, the Suffix Tree Clustering is the actual need automatic cluster method. This method's merit is that it does not need to assign classified number beforehand, and can use the common phrases to describe these articles. In addition, it allows a document to appear in many categories.

2.2 The Brief Introduction of the Suffix Tree

The nodes of the suffix tree are drawn as circles. In Figure 1, each suffix-node has one or more boxes attached to it designating the string(s) it originated from. The first number in each box designates the string of origin; the second number designates which suffix of that string labels that suffix-node.

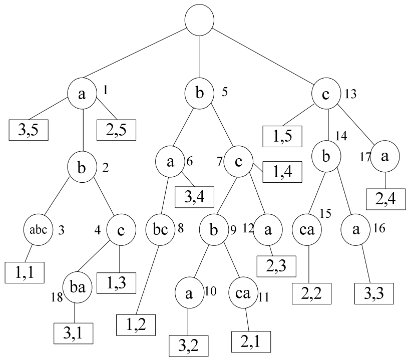


Figure 1 suffix tree

2.3 The Suffix Tree Algorithm

The suffix tree algorithm is one kind of linear complexity documents clustering algorithm, the main idea is: each document is a string, and then construct the suffix tree, the same short-strings which appeared in the suffix tree are considered as the basic clusters, then merge the basic clusters, and finally finish the clustering process.

Input：String S [1…*N*].

Output：Suffix Tree *T*.

```
1 .T= {}
2 .Create root node T={root}；
3. for i=1 to N do
4. CALL Procedure_Add_S[i,N]_TO_T
Procedure_Add_S[i,N]_TO_T
5 .find the longest prefix S[i,r]of S[i,N]which exists in T
6. if this prefix ends at a node(A)then
7. create a leaf node(B)
8. create an edge from A to B and label this edge with S[r,N]
9. else
10. create  a node which path is S[i,r](A)
11 .create a leaf node(B)
12. create  an edge from A to B and label it with S[r,N]
```

In the algorithm, it calls process Procedure_Add_ S[i, N]_To_T to increase suffix S[i, N], i = 1….N. The process is as follows: The line 5 started in T from the root node searches S[i, N] the longest prefix string, if S[i, r] (i≤r≤N one 1) is existed in the T S[i, N] longest prefix string, then there are two situations : If S[i, r] is the side pointed to (A) , then the line 7-8 produces a node point (B) newly, and produces one side to connect A and B , marks the side by S[r+1, N] ; Otherwise, S[i, r] ends on one side, the line 10 increases a point (A) in the termination spot, the line 11 produces new point (B), the line 12 produces connection between point A and B and marks this side S[r+1, N]. when r=i, the point A is a root node.

After completing the suffix tree structure, we can start the base clusters merging, In Table 1,the 1 to 18 points are called the basic clusters (including the leaf node).

Table1 base clusters

Nodes	Words	Atricles
1	a	1,2,3
2	ab	1,3
3	ababc	1
4	abc	1,3
5	b	1,2,3
6	ba	1,3
7	bc	1,2,3
8	babc	1
9	bcb	2,3
10	bcba	3
11	bcbca	2
12	bca	2
13	c	1,2,3
14	cb	2,3
15	cbca	2
16	cba	3
17	ca	2
18	abcba	3

We define a binary similarity measure between base clusters.:

$$|a\cap b|/|a|>0.5 \tag{1}$$

$$|a\cap b|/|b|>0.5 \tag{2}$$

Given two base clusters Bm and Bn, with sizes |Bm| and |Bn| respectively, and |Bm∩Bn| representing the number of documents common to both base clusters, we define the similarity of Bm and Bn to be 1 iff: |Bm∩Bn|/|Bm| > 0.5 and |Bm∩Bn|/|Bn| > 0.5 Otherwise, their similarity is defined to be 0. Next, we look at the base cluster graph, where nodes are base clusters, and two nodes are connected iff the two base clusters have a similarity of 1. A cluster is defined as a connected component in the base cluster graph. Each cluster contains the union of the documents of all its base clusters.

3 The Improvement of the Chinese Clustering Method Based on the Suffix Tree

The suffix tree clustering algorithm is applies in English firstly. As a result of the language difference, when we apply this algorithm in Chinese documents clustering, we needs to do some work to the Chinese documents. We may find from the above introduction that the suffix tree cluster algorithm in English is used the word as a unit. When we apply this method in Chinese documents clustering, we can also use the similar method. Certainly, there are also some algorithms based on the characters and sentences. The algorithm based on characters, its suffix tree produced is too huge, too complicated, unfavorable to the subsequent suffix trees operation, there are also too much space and time spent, meanwhile, its classification produced finally described the result is unsatisfactory. And suffix tree clustering based on sentences, the suffix tree is viewed from space and time, will have reducing by a large margin and shorten, however, even in the same type of articles ,there are few same sentences appeared .So, the author of this document regards that the algorithm based on word is the suitable choice.

Algorithm of suffix tree regards a document as a continuous string, carried on the composition of the suffix tree based on this. Because based on the word, so before carrying on a document in the suffix tree clustering, do the participle at first, calculate the weight of the keywords, all keywords make their sequence according the weight, thus forms a new string, then we carry on relevant subsequent operation, the advantage of adopting this method is:

(1) The weight of the keywords can represent the meaning of the articles more accurately.

(2) Adopted behind the method, the same keyword that dispersed among one article appearing many times can show its importance in the new string.

(3) Because the keyword only appears in the new string once, so it can bring a lot of advantages. First of all, it can make the construction of the suffix tree much easier. Second, after structure finished, it can make the optimization function to the merging of the base clusters. This algorithm has been joined the dynamic array in the nodes of the suffix tree, saved the execution time of the algorithm. From the experimental result, we can find out this method can embody the improvement.

(4) The summary of the cluster can achieve better result after the merging of the base clusters finished.

After the re-queue of all words finished in the documents, the length of the documents will be shorten in some extent, however, we can know by the experimental result, if only carrying on the treatment to the keywords of the documents, we can reduce the time and space of the algorithm, but the final classification description is not very well. For example it will appear “of”, “on”. These clusters can hardly help user to search for the results, it will also have very strong interference effect to some results. It is the problem how to solve the interference of the characteristic. After it is reconfigured according to keyword-weight, it is similar to the traditional algorithm in choosing the characteristic, however, first: Because the documents are numerous, choosing the characteristic according to a certain computing technology can hardly represent the meaning of the article very accurately. Second: Because algorithm is automatic cluster, so classification that produced is confirmed by the words that included in the base clusters. It can not be better to adopt the traditional characteristic to choose the method to describe the final cluster result. So, the author of this document consider from linguistics, the words can have morphological feature, some morphological feature such as noun, verb, adjective, etc. more accurately express an intact meaning, some words for example conjunction, modal particle, function word, etc. almost have any meaning, there is no very great help to express the centre meaning of the articles. Therefore, after arranging keywords to a file, we can introduce the judgment of the morphological feature. The advantage of doing it in this way is:

(1) In numerous keywords, we can choose the words which show the article meaning accurately.

(2) Find the characteristic words which can help users to get necessary information in relevant clusters.

Of course, the judgment of introducing the morphological feature has some shortcomings too, at first, have increased the execution time of the algorithm, second, can not guarantee the words that elected can sign the meaning of the article completely. However, because carrying on the treatment to the keywords in the articles before this, it saved certain execution time and space, so the author thinks that in order to improve the result of the clustering, this time increasing is worthy. And after carrying on the morphological feature, the remaining keywords quantity is further reduced, structure time and space spent of the suffix tree is reduced too. This can be found out from experiment result.

After the keywords in the file finished with morphological feature judgment, we succeed in getting a string composed by keywords, and arranged by the weight. We can construct a suffix tree according the new string. Finally, we can get a suffix tree and finish the one article’s construction. Then, we do the same thing to the rest articles, take them into the suffix tree built ceaselessly and finally, we can get a suffix tree that included the whole articles. After this, we can merge the base clusters and get the summary of the results, and finish the whole algorithm

4 Experimental Result

For our experiments, we constructed document collections by Sogou's documents. These articles are collected from the Web. We choose not to use standard IR collections, as we are interested in the performance of document clustering on the Web.

This experiment chooses 10, 20, 40, 60, 80, 100 articles while classifying at random each time from ten clusters. The machine CPU of this experiment is AMD Athlon 3000, memory 1G, 80G hard disk, Windows XP operating system, the program runs under the environment of Microsoft Visual C ++ 6.0.

In order to evaluate the new STC, we compared it to two algorithms .The algorithms used in the comparison were STC and The Same Word Clustering. We measure the result with the method [14] based on the average precision ratio and the average recall ratio

We can see while we compared with The Same Word Clustering algorithm, the improved suffix tree clustering algorithm has slight deficient rate in average precision ratio, compared with traditional suffix tree clustering algorithm, raised range is not very much large. In the result of the experiment, it will present the rising or reducing in a certain extent, because the judgment of the morphological feature in this experiment is mainly noun, proper noun (including name, name of the place, organization's name, other proper nouns etc.). And after adopting the morphological feature to judge, the average recall ratio rate will be promoted. This shows the filtration of the noise characteristic of the morphological feature is still functional with a certain extent. Secondly, execution time and space of the experiment get the shorting and reducing in a certain extent, We can find out the corresponding figure from Figure 2, and the growth trend slows down. Third, according to final cluster result, this can better describe classification characteristic. For example, if we inquire “China”, we can look for the information needed from different categories, such as “geography” , “history”, “economy”, “military”,. Therefore it can be used in finding the information of the user’s interest fast and conveniently. All about this improvement in these classifications can’t be embodied in experimental data.

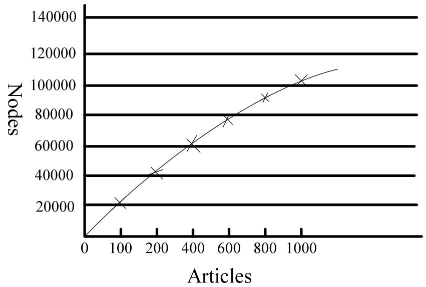


Figure 2　the articles number and the nodes number

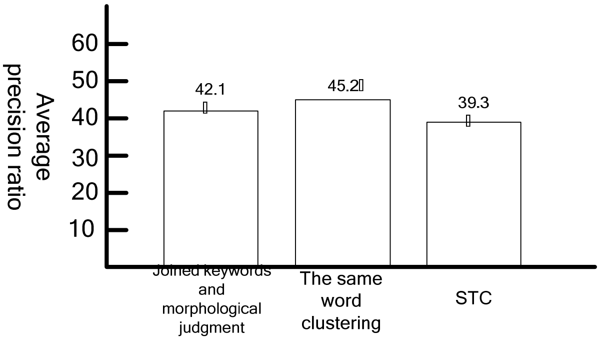


Figure 3　average precision ratio

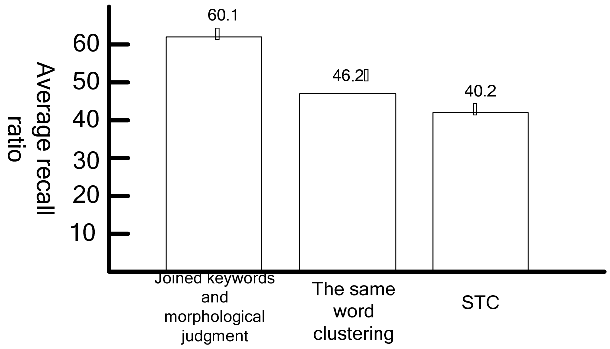


Figure 4　average recall ratio5　Summary

This article has proposed an improvement of the clustering method, in the course of dealing with the algorithm, have joined certain linguistics method. The advantage of the improvement is: first, improve the characteristic choice result, can except some useless words in a certain extent which are not useful, have reduced its negative impact to the clustering. Second, there is certain help in saving space of algorithm and time complexity. Third, in using actually, as to users in intuitionistic choice of the results, can play some improved function.

The experiment proves, there is a certain extent promotion to the suffix tree Chinese documents clustering with this improvement method.

Certainly, this method has not be completed, for example, in practical application, it can not only rely on the noun and proper noun to signify an article, add too much morphological feature judgment can increase the influence of the noise characteristic. So how to balance the choice condition is a problem to be solved. Second, though some nouns can be selected, but it has interference effect to the results, it is not easy to be removed, such as this word of “reporter”, it will appear in news many times, but this word can not express the meaning of the article correctly. The concept of the suffix tree similarity measure is very simple, but the implementation is quite difficult. Our work presented in this paper is mainly focused on improving the effectiveness of document clustering algorithms. Efficiency optimization of the algorithm has been a target of our current work, both the time efficiency and the space efficiency.

References

[1] Oren Zamir , Oren Etzioni . Web Document Clustering: A feasible demonstration[Z].University of Washington, Seattle, U.S.A, 1998

[2] Yang jing-wu . A Chinese Document Clustering Method Based on the Suffix tree.[J]. Wuhan University Journal of Natural Sciences,2004,9(5),pp:817-822

[3] Esko Ukkonen . On–line Construction of Suffix Trees[Z]. Department of Computer Science,University of Helsinki. 1995

[4] GuoLi , Zhang Ji , Tang Jian-long . Research and Implementation of On-line Text Categorization System Based on STC. [J]. Journal Of Chinese Information Processing, 2005 ,19(5), pp: 16-23

[5] Ge Jian , Wang Guo-Ren , Yu Ge . Parallel Construction of Suffix Trees[J]. Computer Science,2004,31(5):96-99

[6] Yan Li-li, Zhang Yan-ping .A Class - Based Feature Selection Algorithm For Test Clustering .[J]. Computer Engineering andApplications,2007,43(12),pp:144-146

[7] Chang Hao-hen . Using Summarization Techniques For Web Content Mining[J]. Control and Automation, 2006, 22(8-3), pp:302-304

[8] Yang Xue-ming. Research and Implementation of Chinese Web-text Clustering[J].New Technology of Library and Information Service,2006,12,pp:81-84

[9] He Hui-yi, Yao Li-xiu , Shen Hong-bin , A New Subspace Clustering Algorithm[J].,2007,5,pp:813-817

[10] Deng Han-cheng , Wang Min-fang. Theoretical Study of the Relationship between Recall and Precision Ratio[J]. Information Journal,2000,19(4),pp:359-362

[11] WuSai,YangDong-qing,Han,Jin-Qiang,Zhang Ming,Wang Wen-qing,. WRM: A Novel Document Clustering Method Based on Word Relation [R]. School of Electronics Engineering and Computer Science, Peking University,2004

[12] Liu Yuan-chao,Wang Xiao-long, Xu Zhi-ming, Guan Yi. A Survey of Document Clustering [J]. Journal of Chinese Information Processing,2006，20(3):55-63

[13] BaoXiao-yuan ,TangShi-wei,Yang, WangTeng-jiao. Suff-Index-An XML Index Structure Based on SuffixTree[J], Journal of Computer Research and Development. 2004, 41(10), pp:1794-1803

[14] Jin yu, Zhihong Deng, Jing Tian, Shiwei Tang.Pinky Search: A Clustering Based Meta-Search Engine. Peking University.2006

[15] Wu Chi-wen(1983 -) , male, Jiangsu ZhanJiagang, master student,main research fields: artificial intelligence and pattern recognition

SWSPMiner: Efficient Mining of Weighted Sequential Patterns from Traversals on Weighted Directed Graph Using Statistical Theory*

Runian Geng[1,2] Wenbo Xu[1] Xiangjun Dong[2]

1 School of Information Technology, Jiangnan University; Wuxi, Jiangsu 214122, China

2 School of Information Science and Technology, Shandong Institute of Light Industry
Jinan, Shandong 250353, China
Email: gengrnn@163.com, xwb_sytu@hotmail.com, d-xj@163.com

Abstract

To solve the problem of mining weighted traversal patterns (*WTP*s) with noisy weight from weighted directed graph (*WDG*), an effective algorithm, called *SWSPMiner* (*S*tatistical theory-based *W*eighted *S*equential *P*atterns *M*iner), is proposed. The algorithm undergoes two phases to discover *WSP*s from the traversals on *WDG*. In the first phase, it adopts the weight's confidence interval (*CI*) to delete the vertices with noisy weights from the traversal database (*TDB*), which reduce remarkably the size of *TDB*. In the second phase, based on the property that the items in a traversal pattern are consecutive, the algorithm regards a traversal pattern as a sequence pattern. Then the algorithm adopts an improved weighted prefix-projected pattern growth approach to decompose the task of mining original sequence database into a series of smaller tasks of mining locally projected database and pushes the weight constraint into the mining process so as to efficiently discover fewer but more important *WTP*s. Comprehensive experimental results show that the algorithm is efficient and scalable for mining sequential patterns from traversals on the *WDG*. Moreover, the algorithm can be applied to various applications which can be modeled as a *WDG*.

Keywords: Statistical Theory, Weighted Directed Graph, Traversal Pattern Mining, Sequential Pattern Mining

1 Introduction

Data mining on graph traversals have been an active research field during recent years. The structure of Web can be considered as a directed graph (*DG*) in which a vertex represents one Webpage and a directed edge represents a hyperlink between two WebPages. Users' navigations on the Web can be regarded as traversals on a *DG*. Capturing users' access patterns in such environments is referred to as mining traversal patterns. Clearly, comparing with other common patterns, the traversal patterns have a sequential property. To reflect different importance of each Webpage, we can assign a weight to each vertex in *DG*. Consequently, we can simulate the traversal behavior on weighted Web structure by a *WDG* model, and the analysis of Website access can be translated into the problem of mining *WTP*s. However, traditional algorithms of traversal patterns mining hardly considered weighted traversals on the graph [1,2]. In additional, it is subjective to assign a weight to each vertex since different users have different importance measure of the same vertex. So we must find a method to balance this difference and remove those vertices whose weights are not confident, here called *noisy weights*. For the weight constraint mining, most of previous works are related to the

* This work was supported in part by the Natural Science Fund of Shandong Province (No.Y2007G25), the Excellent Young Scientist Foundation of Shandong Province, China (No.2006BS01017).

mining of association rules and frequent itemsets [3,4]. Although they take the notion of weight into account as explored in this paper, they only concern with the mining from items which are not ordered, but not from traversals whose items are sequential. A traversal pattern can be regarded as a sequence pattern whose all elements only contain one item since the items in a traversal pattern are consecutive. Thus we can mine interesting traversal patterns by sequential pattern mining approach.

Sequential pattern mining, which was first introduced by Agrawal et al. [5] has been an important and active research topic in data mining. Most previous sequential pattern mining algorithms [5-7] use Apriori-based pruning strategy to prune search space. However, when the minimum threshold is small or mining long sequential patterns, this strategy generates huge candidate sequences in a large sequence database and a large amount of the original sequence database must be repeatedly scanned in order to check if a candidate is frequent. This is inefficient and ineffective. To overcome problems of Apriori based sequential pattern mining algorithms, sequential pattern growth methods [8,9] have been developed. However, they do not consider weights-constraint. For weighted patterns, the traditional pattern growth algorithms lose their utilities since there usually do not exist a '*downward closure*' property among the weighted patterns.

In this paper, we extend previous works [1,2] by attaching weights to the traversals and propos a new effective & scalable algorithm called *SWSPMiner* to discover *WTP*s from traversals on a *WDG*. It adopts the weight's confidence level (*CL*) to remove the vertices with noisy weights, as reduce remarkably the size of *TDB*. Based on the property that the items in traversal pattern are consecutive, the algorithm regards a traversal pattern as a sequence pattern. Then the algorithm adopts an improved weighted prefix-projected pattern growth approach to decompose the task of mining original sequence database into a series of smaller tasks of mining locally projected database. In addition, the algorithm pushes the weight constraint into the mining process so as to efficiently discover fewer but more important *WTP*s.

The remaining of this paper is organized as follows. Section 2 gives the related definitions and model of the problem. The algorithm named *SWSPMiner* is proposed in Section 3. In Section 4, we present experimental results. Finally, Section 5 gives the conclusion and suggestions for future works.

2 Problem Statement

2.1 Preliminaries

Let $I = \{i_1, i_2, ..., i_n\}$ be a set of n instinct items， where i_k (k=1,2,...,n) is unique item. An itemset is a subset of I. A sequence S is an order list of itemsets, denoted as $<s_1, s_2...sm>$, where s_j (j=1,2,...,m) is an itemset, and $s_j \subseteq I$. s_i is also called an element of S, and denoted as $(i_{j1}i_{j2}...i_{jl})$, where i_{jk} is an item in s_j, l is called size of s_j. The brackets are omitted if an element has only one item. An item can occur at most once in an element of a sequence, but can occur multiple times in different elements of a sequence. The size of S, $|S|$, is the number of itemsets or elements in S. The size of an element s_j, denoted as $|s_j|$, is the number of items in s_j. The length of sequence S, $l(S)$, is the sum of all elements size in it, i.e., $l(S) = \sum_{j=1}^{m} |s_j|$. A sequence with length l is called a l-sequence. A sequence $\alpha=<a_1a_2...a_n>$ is contained by another sequence $\beta =<b_1b_2...b_m>$ if there exist integers $1\le i_1 < i_2 <... <i_n \le m$ such that $a_1 \subseteq bi_1$, $a_2 \subseteq bi_2$, . . . , $a_n \subseteq bi_n$. If sequence α is contained by sequence β, then we call α a *subsequence* of β and β a *supersequence* of α. A SDB (sequence database) D is a set of tuples $<sid, S>$, where *sid* is a sequence id and S is a sequence.

2.2 Related Definitions and notions

Definition 1 (*Weighed directed graph*). A *WDG G* is a finite set of vertices and edges, in which each edge joins one ordered pair of vertices, and each vertex or edge is associated with a weight value [10].

There are two kinds of *WDG*s. One is *VWDG* (*V*ertex-*WDG*) which assigns weights to each vertex, and the other is *EWDG* (*E*dge-*WDG*) which assigns weights to each edge. The two *WDG*s are essentially equivalent [10], so we only study *VWDG* in this paper.

Definition 2 (*Traversal on graph, length of traversal, traversal transactions database*). A traversal on graph is a sequence of consecutive vertices along a sequence of directed edges on a *G*. A traversal is a path and can be regarded as a pattern. To easily consider, we assume that each path (i.e., traversal) has no repeated vertices. The length of a traversal is the number of vertices in it. A *TDB T* is a set of traversal transactions, where each transaction, denoted as a tuple <*tid*, *T*>, contains a set of sequential vertices and is associated with a unique traversal identity *tid*.

From the definition of traversal, we know that a traversal discussed in this paper can be regarded as a sequence whose element only contains one item. Consequently, a *TDB* can be considered as a sequence database denoted as *SDB*, and then the problem of mining traversal patterns is converted to that of mining sequential patterns.

Definition 3 (*Sup_count & support*). The support count of a pattern *S,* denoted as *sup_count*(*S*), is the number of traversals, in *TDB*, containing the pattern *S*. The support of a pattern *S*, *supp*(*S*), is the fraction of traversals containing the pattern *S*, denoted as: (|*T*| be the number of traversals.)

$$\mathrm{supp}(\mathrm{S})=\frac{\mathrm{supc}(\mathrm{S})}{|\mathrm{T}|}. \tag{1}$$

Definition 4 (*Frequent sequence pattern*). Given a threshold minimum support *min_sup*, a traversal pattern *S* is a frequent sequence (traversal pattern) if *supp*(*S*) ≥*min_sup*, also called sequential patterns.

Definition 5 (*Weighted sequence, weight of a sequence*). A weighted sequence is a sequence in which each item has a weight. The weight of a sequence is an average value of weights of all items in it.

Given a weighted sequence $S=<s_1s_2...s_m>$,where $s_j=(i_{j1}i_{j2}...i_{jl})$, $j\in\{1,\ldots,m\}$, $l=|s_j|$, the weight of item i_{jk}, $k\in\{1,\ldots,|s_j|\}$) is denoted as $w(i_{jk})$, then the weight of *S* is represented as follows:

$$\mathrm{weight}(\mathrm{S})=\frac{\sum_{j=1}^{m}\sum_{l=1}^{|s_j|}w(i_{jl})}{l(S)}=\frac{\sum_{j=1}^{m}\sum_{l=1}^{|s_j|}w(i_{jl})}{\sum_{j=1}^{m}|s_j|}. \tag{2}$$

Definition 6 (*Weighted support*). The weighted support of a sequence *S*, denoted as *wsupp* (*S*), is defined as follows:

$$\mathrm{wsupp}(S)=weight(S)*(\mathrm{supp}(S)). \tag{3}$$

Definition 7 (*Weighted frequent sequence*). A sequence *S* is said to be a weighted frequent sequence or sequential pattern when its weighted support is no less than a user-specified minimum weighted support threshold called *minwsup*, i.e.,

$$\mathrm{wsupp}(S)\geq \mathrm{minwsup}. \tag{4}$$

Definition 8 (*Weighted prefix*). Given a weighted sequence α = $<e_1e_2...e_n>$ (where each e_i corresponds to a weighted frequent element in the SDB), a sequence β $=<e'_1e'_2...e'_m>$ ($m<n$) is called a weighted prefix of the sequence α if (1) $e'_i=e_i$ for($i\leq m$-1), (2) $e'_m\subseteq e_m$ and (3) all the weighted frequent items in (e_m- e'_m) are alphabetically listed after those in e'_m [9].

Example 1. By definition 8, the weighted sequence <*a*>, <*aa*>, <*a*(*ab*)>, and <*a*(*abc*)> are weighed prefixes of sequence α=<*a*(*abc*)(*ac*)*d*(*cf*)>, but neither <*ab*> nor <*a*(*bc*)> is considered as a weighted prefix if every item in the prefix <*a*(*abc*)> of sequence α is weighted frequent in *SDB* since they do not satisfy the condition (3) of the definition 8.

Definition 9 (*Weighted suffix*). Given a weighted sequence α = $<e_1e_2...e_n>$ (where each e_i corresponds to a weighted frequent element in SDB). Let β $=<e'_1e'_2...e'_m>$ ($m<n$) is the weighted prefix of the sequence α. Sequence γ = $<e''_m\ e_{m+1}...e_n>$ is called a weighted suffix of α with regards to weighted prefix β, where e''_m =(e_m -e'_m). If e''_m is not empty, the suffix is also denoted as <(_items in e''_m) $e_{m+1}...e_n$>. If β is not a subsequence of α, the weighted suffix of α with regards to β is empty [9].

Example 2. By definition 9, for the weighted sequence *S* =<*a*(*abc*)(*ac*)*d*(*cf*)>, the weighted sequence <(*abc*)(*ac*)*d*(*cf*)> is the weighted suffix with regards to the weighted prefix <*a*>, the weighted sequence

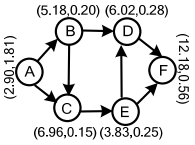
<(*_bc*)(*ac*)*d*(*cf*)> is the weighted suffix with regards to the weighted prefix <*aa*>, and the weighted sequence <(*_c*)(*ac*)*d*(*cf*)> is the weighted suffix with regards to the weighted prefix <*a*(*bc*)>.

Clearly, toward the traversal sequence discussed in this paper, because the size of its element is one, so in definition 8, we need not check the condition (3) and $e'_m = e_m$, and in definition 9, e''_m is empty.

Definition 10 (*Weighted projected database*). Given a *WSP* α in a weighted sequence database *D*. The weighted α-projected database, denoted as $D|_\alpha$, is the collection of weighted suffixes of sequences in *D* about the prefix α..

Definition 11 (*Individual traversal pattern, whole traversal pattern*). Usually, mining the traversal patterns based on the same *WDG* mainly involves two kinds of tasks, i.e., mining the individual traversal pattern and the whole traversal pattern. There are some users to travel the same environment (e.g. Websites) modeled as a *DG*, and they have divers interesting to different content, so they can respectively assign weight to each vertex to reflect the importance degree of themselves. Thus, if there are *n* users to travel the same *DG*, after attaching weight phase, we can get *n WDG* models. We can study a certain user's traversal patterns by his *WDG*, and we can research *k* ($1<k\leq n$) users' whole traversal patterns as well. The former is called the *individual* traversal pattern and the latter is called the *whole* traversal pattern.

In this paper, our work is to mine the whole *WTP*s, i.e., the whole weighted sequential patterns (*WSP*s) from a *WDG*. Figure 1(b) shows a *TDB* with weights given to each vertex by the different users. Note that the whole traversal pattern should not a simple combination of all individual traversal patterns. We had better discovery a whole traversal pattern which can reflect the most users' traversal interesting. That is to say, to mine the whole traversal pattern, we must devise a strategy to accept or reject some individual influence on the whole traversal pattern. We use the following confidence interval to accept or reject the individual impacting.



(a) A revised VWDG

TID	Traversals	Vertex weight
1	<A,B>	2.1,5.3
2	<B,C,E,F>	4.8,6.85,3.4,13.0
3	<A,C>	1.7,7.05
4	<B,C,E>	5.2,7.0,4.0
5	<A>	6.5
6	<A,C,E,D>	1.9,6.7,3.7,6.0
7	<C,E,D,F>	7.0,3.9,5.7,11.6
8	<B,D,F>	5.2,6.1,12.3
9	<A,C,E,D,F>	2.3,6.9,4.2,6.5,12.5
10	<B,C,E,D,F>	5.4,7.2,3.8,5.8,11.5

(b) A TDB with weights

TID	Traversals	Vertex weight
1	<A,B>	2.1,5.3
2	<~~B~~,C,~~E,F~~>	~~4.8~~,6.85,~~3.4,13.0~~
3	<A,~~C~~>	1.7,~~7.05~~
4	<B,C,E>	5.2,7.0,4.0
5	<~~A~~>	~~6.5~~
6	<A,~~C~~,E,D>	1.9,~~6.7~~,3.7,6.0
7	<C,E,~~D,F~~>	7.0,3.9,~~5.7,11.6~~
8	<B,D,F>	5.2,6.1,12.3
9	<A,C,~~E,D~~,F>	2.3,6.9,~~4.2,6.5~~,12.5
10	<~~B,C~~,E,D,~~F~~>	~~5.4,7.2~~,3.8,5.8,~~11.5~~

(c) Revised TDB

Figure 1 The revised VWDG and TDB

Definition 12 (*Confidence interval*, *confidence level*). A confidence interval (*CI*) is an interval in which a measurement or trial falls corresponding to a given probability by a confidence level (*CL*).

For example, for an unknown parameter θ, θ' is its estimated value, than the probability of $|\theta-\theta'|$ less than ε ($\varepsilon>0$) is:

$$P\left(\left|\theta' - \theta\right| < \varepsilon\right) = 1 - \alpha. \tag{4}$$

Thus, $(\theta'-\varepsilon, \theta'+\varepsilon)$ is called the *CI* which indicates the accuracy of θ' and $(1-\alpha)$ is called *CL* which indicates the reliability of θ'. To reflect the real Web browsing characteristics as far as possible, the distribution of weight is generated from normal distribution. For a normal distribution $N(\mu,\delta^2)$, the *CI* of ensemble weight average μ is:

$$\bar{x} - \frac{\delta}{\sqrt{n}}\mu_{\frac{\alpha}{2}} < \mu < \bar{x} + \frac{\delta}{\sqrt{n}}\mu_{\frac{\alpha}{2}}. \tag{5}$$

Here, $\bar{x}$ is the average weight of the sample, and δ is the standard deviation of the sample's weights. Thus, the length of *CI* is the smallest, and the accuracy of estimation value of μ is the highest. In our problem, we adopt the notion of *CI* to classify the weights into the confident ones and abnormal ones.

Definition 13 (*Noisy weight, outlier vertex*). To get a whole traversal pattern which reflect the most users' traversal interesting, we must remove some vertices whose weight overrun the tolerable *CI* of the weight. These removed weights are called noisy weights, and the corresponding vertices are called outlier.

If a weight exists within the *CI*, then it is regarded as a confident one to be accepted, but if it lies outside the *CI*, it is considered as an outlier to be removed.

Thus, the problem concerned with in this paper is stated as follows. Given a weighted directed graph *G*，a minimum weighted support threshold *minwsup* and a set of path traversals on the graph —traversal database *T*, we find the whole traversal patterns (i.e., sequential patterns) with weight constraint in *T* by sequence patterns mining approach. However, by Eq. (2), (3) and (4), if a weighted sequence is weighted infrequent, but its weighted supersequence is possible to be weighted frequent. That is to say that the weighted support measure satisfies neither the '*anti- monotone*' property nor the '*monotone*' property. So we cannot directly use the '*anti-monotone*' property of weighted support to prune weighted infrequent candidate sequence patterns.

2.3 Revised weighted support

To let weighted support satisfy the '*anti-monotone*' property so as to prune the weighted infrequent candidate sequence pattern, we revise the representation of weighted support in mining process. Given a *G* containing *n* nodes (i_1, $i_2 \ldots i_n$) whose weight must satisfy: $min(W) \le w(i_j) \le max(W)$, where $W=\{w(i_1), w(i_2) \ldots w(i_n)\}$, *min*(*W*) and *max*(*W*) are the minimal weight value and the maximal weight value of *W* respectively. To let weighted patterns $S=<s_1, s_2 \ldots s_m>$, where s_j ($j \in \{1,\ldots,m\}$) is an element containing some vertices on *G*，satisfy the '*anti-monotone*' property (i.e., if $wsupp(S) \le minwsup \Rightarrow wsupp\ (S') \le minwsup$, where $S \subset S'$), we revise $w(i_j)$ as the following two representations: $w(i_j)=min(W)$ or $w(i_j)=max(W)$. Then, the weight of the sequence pattern *S* is revised a*s* follows.

$$\text{weight}(\mathrm{S}) = \frac{\sum_{j=1}^{m}\sum_{l=1}^{|s_j|} w\left(i_{jl}\right)}{l(S)} = \frac{\sum_{j=1}^{m}\sum_{l=1}^{|s_j|} \min(W)}{\sum_{j=1}^{m}|s_j|}. \tag{6}$$

or

$$\text{weight}(\mathrm{S}) = \frac{\sum_{j=1}^{m}\sum_{l=1}^{|s_j|} w\left(i_{jl}\right)}{l(S)} = \frac{\sum_{j=1}^{m}\sum_{l=1}^{|s_j|} \max(W)}{\sum_{j=1}^{m}|s_j|}. \tag{7}$$

Because $S' \supset S$, so support(*S'*) ≤ support(*S*). If we adopt the revised weight of sequence pattern (i.e., Eq. (5) or (6)), we can get w*supp* (*S'*) ≤w*supp* (*S*), i.e., the revised weighted support meets the '*anti-monotone*' property. However, if we adopt Eq. (6), clearly we could prune some patterns which should have been weighted frequent to lead to incorrect mining results since the weighted support is too small. To avoid this flaw, we adopt Eq. (6) to compute revised weight of each sequence pattern. However, the weighted support value computed by Eq. (6) is only an approximate value, so in final step, we should check if each mined result sequential pattern *S* is really a weighted frequent sequential pattern by its real weight value, i.e., we must check if each mined result sequential pattern satisfies the following inequality:

$$\sum_{j=1}^{m}\sum_{l=1}^{|s_j|} w\left(i_{jl}\right) \Big/ l(S) * \operatorname{supp}(S) \ge \text{minwsup}. \tag{8}$$

3 Mining Weighted WSPS From WDG by Improved Weighted PrefixProjected Pattern Growth Approach

We devise an efficient and scalable algorithm, called *SWSPMiner* to mine the whole *WSP*s from *TDB*. The algorithm is mainly composed of two ordered phases: (1) The phase of revising the *TDB* and *VWDG*, and (2) The phase of discovering the whole *WSP*s from the revised *TDB* generated in the previous phase.

3.1 Revision of VWDG and TDB

This phase is a pre-procession phase, in which each vertex in the *VWDG* is revised by adding its average weight (denoted as w_μ) and standard deviation of weights (denoted as w_σ). Then, base on the w_μ and w_σ, we get the weight *CI* of each vertex and remove the vertices with noisy weight from *TDB*. For example, for the vertex '*A*' in Figure 1 (b), different users' weight values of it is: 2.1, 1.9, 6.5, 1.9 and 2.3. Then after calculating, we get the supplementary information of '*A*' as follows: w_μ (*A*)=2.90 and w_σ(*A*)=1.81. Given (1-α)=90%, then $u_{\alpha/2} = u_{0.05}$= 1.645, by Eq. (2), the *CI* of weight for '*A*' is (2.90-1.81/$\sqrt{5}$ *1.645, 2.90+1.81/ $\sqrt{5}$ *1.645) =(1.57,4.23). Based on the *CI* of the weight for '*A*', we find the weight of vertex '*A*' in the #5 traversal transaction lies outside the *CI* of ‘*A*’, i.e., it is a outlier of vertex ‘*A*’, so the vertex ‘*A*’ in #5 traversal transaction should be deleted and the support of vertex ‘*A*’ is revised to be 4 from 5. Toward other vertices, we adopt the above similar method to update *VWDG* and *TDB*. After the first phase, we get the revised *TDB* and *VWDG* shown as Figure 1(a) and (c) respectively.

3.2 Discovery of the whole WSPs

The *WSP*s discovery phase is the main phase, in which the whole *WSP*s are mined from the amended *VWDG* and *TDB* on graph generated in the first phase.

As we described above, for the revised weighted setting, it is true that all the supersequences of a weighted infrequent sequence pattern are weighted infrequent, i.e., revised weighted support has an ‘*anti-monotone*’ property. Based on this property, we devise an efficient and scalable algorithm called *GTWSPMiner* which exploits a d*ivide-and-conquer* strategy with an improved weighted prefix-projected pattern growth method to efficiently mine fewer but more important *WSP*s.

Algorithm *GTWSPMiner* is given in Figure 2. Figure 3 gives the procedure *GTWSPM* ($SD|_{\alpha}$, *GTWSP*, α, *l*), in it, *GTWSP* is used to store so far found *WSP*s.

```
SWSPMiner(Statistical theory-based Weighted Sequential Patterns Miner)
Inputs: (1)A traversal database SD on the WDG G,
        (2)A minimum weighted support threshold minwsup,
        (3) The weights of each vertex of G.
Output: All the real weighted sequential patterns (WSPs) in SD
Method:
1.To scan SD once to find all the global weighted frequent items b which
satisfies: supp(b)*Max(W)≥minwsup;
2.To remove all items c which satisfies: supp(c)*Max(W)<minwsup;
3.To initialize the set SWSP. //SWSP to store the real WSPs;
4.To Call SWSPM (SD, SWSP, <b>, 1).
```

Figure 2 Algorithm *SWSPMiner*

Example 3. For the *TDB* shown in Figure 1, given *minwsup*=0.5 (clearly, *max*(*W*)=12.18), we show how to mine *WSP*s based on *WDG* traversals by using an improved weighted prefix-projected pattern growth approach. The mining process is as follows.

(1) Scan *TDB* once to find all the weighted frequent item *b* in sequence transactions which satisfies *supp* (*b*)**Max* (*W*) ≥ *minwsup*. Each of these frequent items is an approximate weighted frequent 1-sequence. In our example, they are <A>: 4, <B>: 3, <C>: 4, <D>: 3, <E>: 4 and <F>:2, where the notation “<*pattern*>*: count*” represents the pattern and its associated support count. The weighted infrequent items can be removed (here no item is removed).

(2) Division of searching space of *WSP*s. The complete searching space of *WSP*s can be partitioned into the following six subspaces according to the six prefixes: <A>, <B>, <C>, <D>, <E> and <F>.

(3) Mining the *WSP*s from six subspaces. The *WSP*s in six subspaces can be mined by constructing the corresponding *prefix-projected database* and mine each recursively.

(4) For the above mined result sequential patterns *S*, we must check if they are the real sequential patterns by the inequality: $\sum_{j=1}^{m}\sum_{l=1}^{|s_j|} w\left(i_{jl}\right) \Big/ l\left(S\right) * \mathrm{supp}\left(S\right) \geq$ minwsup and remove the sequence which does not satisfy the above inequality from *SWSP*.

```
Procedure SWSPM (SD|α, SWSP, α, l)
Inputs: (1) SD|α is the α-weighted projected database if α≠<>, otherwise, it is
SD, (2) Set of real WSPs, SWSP, (3) A WSP α, (4) l is the size of α, i.e., |α|
Output: the real WSPs set SWSP
Method:
To Scan SD|α once to find all weighted frequent items β which satisfies:
supp(<β>)*Max(W)≥ minwsup, such that
  (a) β can be assembled to the last element of α to form a new WSP or
  (b) <β> can be appende to α to form a new WSP.
for (∀β∈ SD|α)
    add β to α to form a new WSP α’;
    SWSP=SWSP∪<α’>.
end
for (∀S∈SWSP)
    if (Σ_{j=1,m}Σ_{l=1,|Sj|} w(i_{jl})/l(S)*supp(S)<minwsup)
        SWSP=SWSP-S;
    else
        t=check-item-order(S); //to check if the items of S is consecu-
        tive, if yes return 1, otherwise return 0
        if (t==1)
            break;
        else
            SWSP=SWSP-S;
        end
    end
end
for (∀α’)
    construct α’-weighted projected database SD|α’;
    call SWSPM(SD|α’, SWSP, α’, l+1);
end
output SWSP;
```

Figure 3 Procedure *SWSPM*

(5) Finally, toward each real *WSP* minded in the (4) phase, we must also check if it is included in *TDB* shown in Figure 1. Note, each traversal pattern discussed in this paper can be regarded as a sequence pattern the size of whose element is one, and the items

in the traversal is ordered and consecutive. In addition, according to the original permutation relations of items in traversal transactions of *TDB*, the relative position between any two items of mined traversal pattern is not changed in the mining process by the prefix-projected pattern growth approach, and only the neighbor relations among items of mined traversal patter could be changed. So we can decide if each real *WSP* mined in (4) phase is included in *TDB* by the above discussed facts. The checking method is as follows: If the mined *WSP* changes the neighbor relations of items in original traversal transactions, then this *WSP* must not be included in *TDB*, and we must remove it from *GTWSP*. For example, toward the prefix <*C*>, after the above several mining phases, we get a real *WSP* <*C,F*>:1, however, this *WSP* changes the respective neighbor relations of items '*C*' and '*F*', so the pattern <*C,F*>:1 is not included in the original SDB and must be pruned from *GTWSP*.

After the above five ordered mining phase, we ultimately get each prefix-projected database and corresponding real *WSPs* contained by *TDB*. They are shown in Table 1.

Table 1 Projected database and corresponding WSPs

prefix	Prefix-projected database	Corresponding WSPs
<A>:4	<B>,<E,D>,<C,F>	<A>:4
<B>:3	<C,E >, <D,F>	<B>:3,<B,C>:1, <B,D>:1,<B,D,F>:1
<C>:4	<E >, <E>, <F>	<C>:4,<C,E>:2
<D>:3	<F>	<D>:3,<D,F>:1
<E>:4	<D>, <D>	<E>:4,<E,D>:2
<F>:2	∅	<F>:2

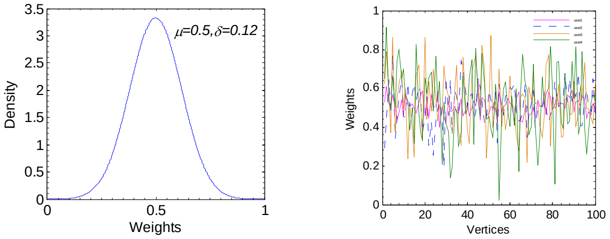
4 Experimental Evaluation

Because there are not real datasets about *WDG* currently, we test the algorithm performance by using synthetic dataset. The experiments were performed on Pentium IV PC at 2.93 GHz with 768MB memory and Windows XP installed. We used Microsoft SQL Server 2000 database to generate simulation of *WDG* and the traversals on it, and implemented our algorithm with C++ language. All the reported runtimes are in seconds.

4.1 Generation of synthetic datasets

During the experiment, the *WDG* is generated mainly according to following parameters: number of vertices and max number of edges per vertex. And then, we assigned random weight to each vertex of the graph. The characteristics of these datasets are summarized in Table 1.

The distributions of weight of all users are generated from Gauss distribution. Figure 4(a) shows the Gauss distribution (μ=0.5,σ=0.12), and Figure 4(b) shows the weight distributions of 4 users for the situation of vertices=100. The σ in it is 0.05, 0.1, 0.15 and 0.17 respectively.



(a) Gaussian distribution density (b) Four users' weight distribution

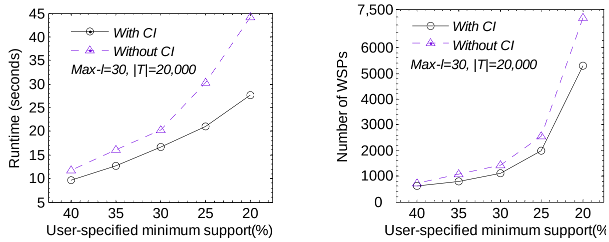
Figure 4 Weight distribution

Toward vertices=100, we implemented experiment to test algorithm's performance. In experiment, we fist used the weight distributions of four different users, shown in Figure 8(b), to found a revised *VWDG*, and then we generated 8 sets of *TDB* on the revised *VWDG* funded above, in which the maximum length of traversals varies from 10 to 60. All experimental results are average value of 8 sets of synthetic datasets. In addition, we executed the scalability test for vertices=100 by varying the number of traversals from 10k to 50k and for the fixed |*T*| and *minsup* by varying the number of vertices from 100 to 500.

4.2 Experimental results

4.2.1 Impact of CI

Figure 5 shows the effect of *CI* on the runtime and number of *WSP*s. From the figures, we can see that the runtime is faster and the number of *WSP*s is much less when considering *CI*. This means that the exclusion of outliers by the *CI* makes us more efficiently discover *WSP*s from the *TDB*.

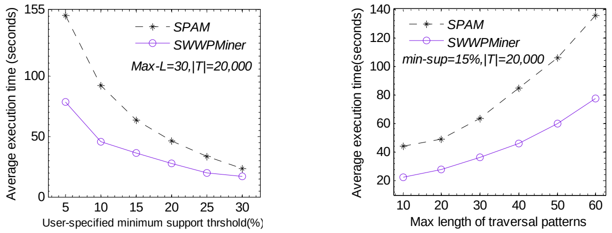


(a) Runtime (b) Number of WSPs

Figure 5 Impact of CI

4.2.2 Effectiveness comparison of SWSPMiner and SPAM

As [11] shown, SPAM is by far the fastest algorithm when mining to get the whole set of the sequential patterns, So we only explore our experimental results on the performance of *SWSPMine*r in comparison with *SPAM*. Figure 6 shows the trend of the execution time of *SWSPMine*r and *SPAM* with respect to different *min-sup* and *Max-L* based on |*T*|*=20,000*. As shown in Figure 6(a), the average runtime of two algorithm increases along with *min-sup*'s decreasing. The lower *min-sup* is, the larger performance difference between them becomes. In all case of *min-sup*, *SWSPMine*r outperforms algorithm *SPAM*. This is because *SWSPMin*er is a weight constraint-based sequential patterns mining algorithm, and it can mine more important and has fewer search space by pushing weights constraints into the process of mining. However, algorithm *SPAM* is not constraint-based one, and it has a larger searching space than *SWSPMiner*. Figure 6(b) shows *SWSPMiner* is faster than *SPAM* and the difference between them becomes larger as *Max-L* becomes longer.

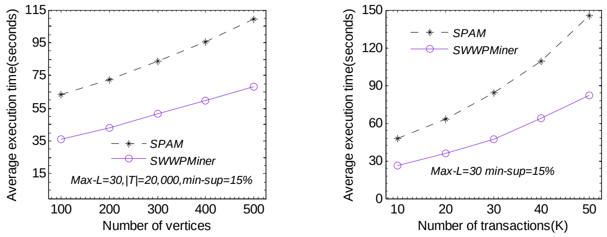


(a) Different min-sup w.r.t runtime (b) Different Max-L w.r.t runtime

Figure 6 Runtime comparison

4.2.3 Scalability study

To evaluate how the performance of *SWSPMiner* scales with the size of the database, we performed an experiment in which we respectively varied the number of vertices from 100 to 500 and the number of traversal transactions |*T*| from 10 to 50k based on the *min-sup*=15.0% and *Max-L*=30. Figure 7 shows the experimental results. From Figure 7 we can see *SWSPMiner* approximately scales linearly with the size of the vertices and traversal transactions. And although itself runtime also increases, *SWSPMiner* has much better scale-up properties than *SPAM*.



(a) Number of vertices scale-up test (b) Number of vertices scale-up test

Figure 7 Scalability test

5 Conclusions

This paper explores the problem of mining frequent traversal patterns from *WDG* by the prefix-projected pattern growth approach. It regards a traversal pattern as a sequence pattern and proposes the algorithm *SWSPMiner*. Firstly, the algorithm adopts the weight's confidence level to reduce the *WDG* and *TDB*. Then, it makes the weighted support of the traversal patterns possess the '*anti-monotone*' property by revising the weight of patterns. In the mining process, *SWSPMiner* adopts a '*divide-and-conquer*' strategy to decompose the total mining task from original sequence database into a series of smaller tasks of mining locally projected database. In addition, it carries out the validity check about weighted support of the mined approximate *WSP*s, and then checks if the *WSP*s are included in the *TDB*, and ultimately mines the *WTP*s contained by the *TDB*. The extensive performance analysis shows: (1) Taking *CI* into consideration, we can mine more reliable *WSP*s. (2) Algorithm *SWSPMiner* is efficient and scalable. There are a lot of cases which can be modeled as a *WDG*. How to efficiently put the model and algorithm devised in this paper into practice, and can we deeply optimize the algorithm will be our future research topics.

References

[1] M.S. Chen , J.S. Park and P.S. Yu, "Efficient Data Mining for Path Traversal Patterns", IEEE Trans. on Knowledge and Data Engineering, Vol.10, No.2, 1998, pp. 209-221

[2] A. Nanopoulos, Y. Manolopoulos, "Mining Patterns from Graph Traversals", Data & Knowledge Engineering, Vol.37, No.3, 2001, pp. 243-266

[3] W. Wang, J. Yang, P.S. Yu. "Efficient mining of Weighted Association Rules (WAR)", Proceedings of the 6th SIGKDD'00, New York, NY, USA, ACM, 2000, pp. 270-274

[4] C.H. Cai, A.W.C. Fu, C.H. Cheng et al. "Mining Association Rules With Weighted Items". Proceedings of the ISDEA'98, Washington, DC, USA, IEEE Computer Society, 1998, pp. 68-77

[5] R. Agrawal, R. Srikant. "Mining Sequential Patterns", Proceedings of the Eleventh International Conference on Data Engineering, Los Alamitos, CA, USA, IEEE Computer Society, 1995, pp. 3-14

[6] M. Zaki. "SPADE: An Efficient Algorithm for Mining Frequent Sequences". Machine Learning, Vol.42, No.1-2, 2001,pp. 31-60

[7] R. Srikant, R. Agrawal. "Mining Sequential Patterns: Generalizations And Performance Improvements", Proceedings of the 5th Int. Conf. on Extending Database Technology (EDBT'96), London, UK, Springer-Verlag,1996, pp.3-17

[8] J. Han, J. Pei, B.M. Asi, et al. "Freespan: Frequent Pattern-Projected Sequential Pattern Mining", Proceedings of the Sixth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, New York, NY, USA, ACM, 2000, pp.355-359

[9] J. Han, J. Pei, B.M. Asi et al. "Mining Sequential Patterns By Pattern-Growth: the Prefixspan Approach", IEEE Trans. on Knowledge and Data Engineering, Vol.16, No.11, 2004, pp. 1424-1440

[10] R. Geng, X. Dong and W. Xu, "Efficiently Mining Closed Frequent Patterns with Weight Constraint from Directed Graph Traversals Using FP-trees", Proceedings of ISIP'08, IEEE CS, Los Alamitos, USA, 2008

[11] J. Ayres, J. E. Gehrke, T. Yiu, et. al, "Sequential Pattern Mining Using Bitmap Representation" Proceedings of 2002 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'02), New York, NY, USA, 2002, pp. 429-435

A Logical Model of E-Commerce System by Applying the Idea of Business Process Reengineering

Dongbin Hu Feng Liu

School of Business, Central South University , Changsha, Hunan, China

Email: hdbin@163.com

Abstract

This paper first emphasizes the importance of the BPR in the e-commerce system by explaining the relationship between them, and gives the necessary of applying the idea of BPR in the e-commerce system's development, and then discusses when and how to apply it. To solve this problem, we give a logical model in the system analysis phase. To explain the feasibility and validity of the logical model, we define the detailed processes by applying this model in a typical example of a web-reimbursement system. And finally, we can deduce this logical model can integrate the BPR with e-commerce system's development better.

Keywords：BPR; E-commerce system; prototypal method; BPR practices

1 Introduction

The characteristic of modern e-commerce system (E-commerce systems have different appellations: some papers call it web applications, some call it hypermedia systems, and others call it e-commerce applications etc, and there we call it e-commerce system) is the high-integrated of enterprise interior systems with its web site. Since this new characteristic is transformed, researchers of the e-commerce have been gradually turning to their research on e-commerce system's development concepts, methodologies and tools, the methodologies and tools for business process redesigning (BPR) in the context of e-commerce, the way for applying the idea of BPR to the development of e-commerce system. The first two pieces of research have been discussed a lot, and are successfully tackled by researchers and practitioner. In [8], the article presents its exploitation concepts like modularity, containment and encapsulation and gives the strategies for e-commerce system design modeling by using these concepts. In [5], it summarizes a number of methodologies for e-commerce system's development: the object-oriented hypermedia design method (OOHDM), OOHDM-Web which exploits the prototypal method, the web site design method (WSDM), the object-oriented Hypermedia and UWA methodologies, web modeling language (WebML), and the Ubiquitous Web Applications design framework. OOHDM and UWA design framework is the two well established methodologies. In [6], it presents a methodology and some tools for reforming business processes in the context of e-commerce by using a critiquing approach---the critiquing methodology and tools such as a measurement framework, a set of critiquing tables, and a software environment. And so on.

However, there are few researches on how to apply the idea of BPR in the e-commerce system's exploitation. So, this paper presents a logical model of e-commerce system by applying the idea of BPR. The paper is structured as follows. In section 2, we discuss the relationship between the BPR and E-commerce system's exploitation. In section 3, we explain what is the inspiration of our logical model and how to build the logical model. In the last section, we give conclusions.

2 The relationship between BPR and e-commerce system's development

Many companies have found out the hard way that successful e-commerce requires more than a flashy web presence. Existing business processes must be seamlessly integrated with the new, electronic form of interaction with suppliers and customers. E-commerce promises to dramatically alter the structure and processes of commerce. On the one hand, e-commerce systems have been extended in order to support the design of the business processes. Different exploitation methodologies have different effect on the designing of the business processes. In [5], it presents a framework for analyzing and comparing e-commerce system's exploitation methodologies, with regard to their approach for designing business processes, highlighting their strengths and their weaknesses. This framework indicates that, on the whole, OOHDM lends itself better than other methodologies to the design of business processes. On the other hand, BPR can serve well for e-commerce system's exploitation. In [7], it summarizes the 13 selected best practices to redesign business processes for e-commerce system's development: task elimination, task automation, knockout, control relocation, parallelism, case manager, empower, outsourcing, contact reduction buffering, trust party, case types and case-based work. Then they apply the 13 practices in two case studies to show how these practices affect the e-commerce processes and support the e-commerce system's exploitation. Many known practices to improve a process may be used to make an e-commerce process better performed.

We know both of the BPR and the e-commerce exploitation are enterprise innovations. BPR is an effective method in the business process designing. BPR is an innovation, business process reengineering in order to be successful for e-commerce should be done the revolutionary way. it needs essential consideration and thoroughly design in the business process designing so as to improve the four targets: quality, service, cost and speed. Its success is involved in a lot of complicated factors which start with the cost and customer value.

The same as the other information system, e-commerce system's exploitation is a project, it has its own exploring methodologies and tools. At the same time, it is also an innovation which determines we must put a revolution point of view. The main barrier for the success of e-commerce is not technical, but have to combine with the idea of the BPR. Some report that 80% to 95% of corporations' e-commerce web sites are not even linked up with their back-office processes. Once again, the view on the entire process is missing, which prevents the new technology from becoming truly effective.

3 Applying BPR to the e-commerce system's development

To all appearances, e-commerce system's exploitation cannot be departed from the idea of BPR. But how can we apply the idea of BPR to the e-commerce system's developing? That is the question we will discuss in the paper.

3.1 The inspiration of the logical model

There are two types of integration of the BPR with e-commerce in the process of the e-commerce developing. One is first BPR then e-enabling, the other is BPR and e-enabling in one step[7]. The first type is exploited by some practitioner. Here, we illustrate an e-commerce system's development model attributed to the first type which is brought forward by Li Zhigang and Zhang Xianghong[4]. It's described in the Figure 1.

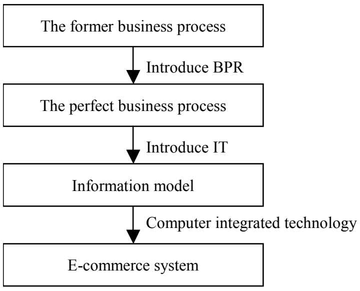

The former business process
Introduce BPR
The perfect business process
Introduce IT
Information model
Computer integrated technology
E-commerce system

Figure 1 E-commerce exploiting model of the first type

This application model is an idea that first carry BPR , then exploit e-commerce system by means of IT on the base of it. But we know both of them are not an easy job. Besides according to their implement steps, the supervisor's outstanding leadership and unfearing battle effectiveness is most important, they have to do with legal issues, reluctance of people in changing their way of doing things, lack of trust in e-transactions and so on. According to this application model, it seems that it's an ideal and perfect model, but it's excruciating for the supervisor and their employee. Because any innovation means the dread of new things and the yearn of old environment, this love knot is enough to make them who take part in the innovation in fear and trembling. Then what is the solution? How to alleviate their dread? This is the inspiration of our logic model. We can find the answer from the second type of integration----BPR and e-enabling in one step. If we deduce when and how to apply the idea of BPR in the process of enabling, it's a perfect type apparently.

3.2 The way to build the logical model

We know the most typical method to exploit information system is the prototypal developing method which is on the base of the system design life cycle (SDLC). This method can simplify the most complicated system's exploitation starting with a smaller prototype. That's to say, first exploiting a prototype which satisfies a small part of system requirements, then constantly extending it until it satisfies the whole complicated system requirements. In 1998, Schwabe melted the prototypal method with the OOHDM, which is OOHDM-Web. With this methodology we can educe an exploring model which has four phases: system programming, prototypal (which has four sub-phases: system analysis, system designing, system exploiting and system evaluating), running and maintenance. It's described in Figure 2.

It is necessarily to point out that this exploitation model can reduce uncertainties and mistakes. We have given the phases of the e-commerce system's exploitation, and apparently prototypal is the main phase of the exploitation. So the following question is when and how to introduce the idea of BPR to this e-enabling in the phase of prototypal. We know every phase has its own tasks. We can elicit when to introduce the idea of BPR according to the tasks of every phase. We know system analysis have the following tasks: system requirements and functions analysis, those analysis are the business process analysis in a word. So, we can introduce the idea of BPR in this phase apparently. Then how to introduce it in the phase of system analysis? In this phase, firstly we analysis the system requirements, secondly we analysis the functions which specific the system requirements, thirdly we divide the functions into many small sub-functions which can satisfy the whole system functions by integrating. We know functions are composed of a succession of business processes. So the new system's function is realized by a succession of business processes which are composed of the former processes and the new designing processes. In order to finish the new function perfectly, we must introduce the idea of BPR to redesign the former business process. Besides according to the BPR's implement steps, we also use the best BPR practices. When the amendatory business processes are designed, maybe it will drive some new system functions arisen by remounting to the phase of system function analysis. So, we build a logic model which is shown in Figure 3. This application model is the system analysis logical model which preferably applies the idea of BPR to the e-commerce developing in the phase of system analysis.

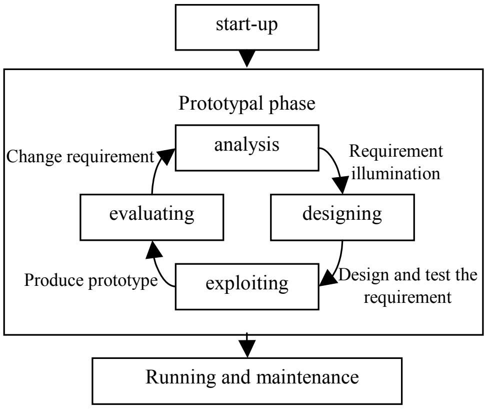


Figure 2 The phases of the e-commerce system's exploiting

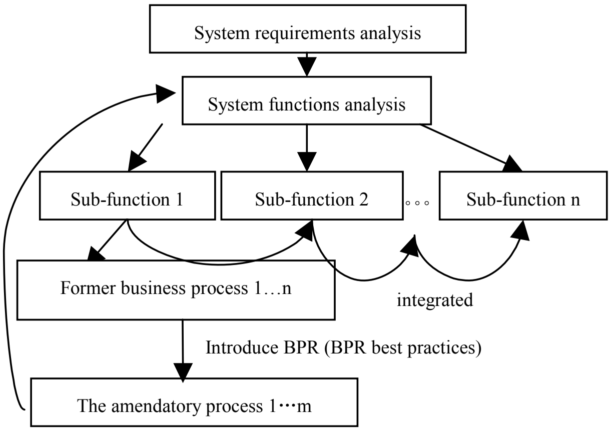

System requirements analysis
System functions analysis
Sub-function 1
Sub-function 2
。。。
Sub-function n
Former business process 1…n
integrated
Introduce BPR (BPR best practices)
The amendatory process 1⋯m

Figure 3 E-commerce system analysis logical model

3.3 The application of the logical model

We consider a web-reimbursement system. This system is aimed at solving the question of the reimbursement which is tangle some and time-consuming. By carrying the system functions analysis, the system has such main sub-functions: write the applications, examine and approve the applications, loan, write the expense accounts, examine and approve the expense accounts, reimburse. To explain the logical model, we discuss the function of examining and approving the applications. We can draw the former business process in table 1.

Table 1 The former processes of the reimbursement

put the applications to the applicants' department examinants
pack up the applications by the examinants
examine the applications by the examinants
approve the applications by the examinants
transform them to the financial department examinants
pack up them by the financial department examinants
examine them by the financial department examinants
approve them by the financial department examinants
transform them to the principal of the applicants' department
transform them to the applicants by the principal

We can redesign the former business processes by applying the best BPR practices. The amendatory processes are described in table 2. The system sub-function also must include: process the list automatically, send the list and applications and so on after the business process redesigning by remounting to the phase of system function analysis.

Table 2 The amendatory processes of the reimbursement

List the applications automatically by the system
Send the applications to the examinants by the system
The examinants receive the applications
The examinants examine and approve the applications
Send the applications to the applicant by the system

Therefore, we can ameliorate the system function continuously by applying the business process redesigning. That to say, this is a preferable logical model which melts the idea of BPR and the prototypal method of e-commerce system's development.

4 Conclusions

This article applies the idea of BPR to the exploitation of the e-commerce system and gives a logical model in the system analysis phase. This application model can serve well in the phase of e-commerce system analysis.

It is worthwhile to point out that BPR is not limited within the corporation, it expects to be integrated with the exterior of the corporation. It is the whole supply chain's integration, not just a part of it. So in the future the discussion is how to apply the idea of X reengineering to the whole supply chain's exploitation of e-commerce system.

References

[1] Xu Huahui, Zhou Xiaojun, Corporation, Management and Application, China machine press, 2005

[2] James Champy, .X-Engineering the Corporation, CITIC PUBLISHING HOUSE, 2002

[3] Luo Zhenhua, E－Business System Programming and Designning, TSING HUA University publishing house, 2006

[4] LiZhigang, Zhang Xianghong, The method and strategy of E-Business system's programming, China Management Informationization, vol（11）, 2006

[5] Damiano Distante, Gustavo Rossi , Gerardo Canfora, Modeling Business Processes in web Applications: An Analysis Framework,SAC'07, 2007-5-11, pp.1677-1682

[6] Lerina Aversano, Thierry Bodhuin, Gerardo Canfora,

Raffaele Esposito and Maria Tortorella, Evolution of Business Processes towards eBusiness using a critiquing approach, SAC'04, 2004-05-14, pp.1351-1358

[7] Dr. M.H. Jansen-Vullers, M. Netjes, Dr. ir. H.A. Reijers, Business Process Redesign for Effective E-Commerce, ICEC'04, 2004, pp.382-391

[8] Ahmet and N.Yasemin, Strategies for Hypermedia Design Modeling, MIS 2003, 2004, pp.150-157

[9] Cui Shuyin, Ren Hao, Probe on Design of Business Process Under E-business Environment, Value Engineering, No（9）, 2005, pp.64-66 [10] LIU Bing, Huang Xiaoyuan, Sun Shuang. Exploring Relationship Between Information Technology and Business Process Reengineering, China Metallurgy, No（4）, vol（9）, 2006,pp.47-50

Study on Architecture of Global Supply Chain with MAS

Jizi Li   Zhiping Zuo   Peilin Guo

School of Economics & Management, Wuhan University of Science & Engineering ,Wuhan, 430073，China

Email: Jisonli@yahoo.com.cn

Abstract

The last decades of the twentieth century witnessed a considerable expansion of supply chain into international locations. Effective and efficient global supply chain coordination is crucial way to sharpen the edges for the firms to win the international competition. To address this need, this paper used multi-agent theory to construct an architecture based on the requirement for enabling dynamic interoperation of members within a supply chain for successful global manufacturing. The definition and internal structure of single agent were presented in details and the communication structure of agents was explored. Furthermore, a multi-agent coordination mechanism was proposed to determine the best decisions.

Keywords: multi-agent system; global supply chain; architecture

1 Introduction

The last decades of the twentieth century witnessed a considerable expansion of supply chain into international locations, especially in the automobile, computer, and apparel industries. This growth in globalization, and the additional management challenges it brings, has motivated both practitioner and academic interest in global supply chain management. The interest in global operations management among researchers has been documented by Prasad & Babber[1], who noted both a long history of attention to global operational issues, as well as increase in the number of articles published in the leading operations management journals on this subject. Supply chain management is not just a just a domestic phenomenon—supply chains transcend national boundaries, imposing the challenges of globalization on managers who design supply chains for existing and new product lines.

Literature pertaining to supply chain management that uses multi-agent system (MAS) methods is reviewed. Through the review, it has been found that the fundamental issue of a supply chain is not addressed. Such limitation has triggered the proposal of an architecture, which is constructed based on the fundamental requirement for supply chain (i.e., promote interoperation of members within a supply chain). The idea is to enable all numbers within the supply chain to work as a whole in a coordinated manner eliminating unnecessary problems such as miscommunication, lack of or outdated information, bullwhip effect and the like. The architecture, the benefits of using the architecture, the future directions are also discussed in this paper.

2 Literature review

Many researchers viewed agent negotiation and bidding mechanism as the main function to facilitate communication between parties within a supply chain. In line with the view, different multi-agent based models that focus on agent negotiation/bidding mechanism are proposed to enable successful selling and buying operations. Existing methods for supply chain management are commented as not designed to produce optimal solutions but solutions that satisfy objectives and constraints based on theory of constraint. Hence, a supply chain method is proposed to evaluate alternative solutions via multi-agent interactions before an optimal solution is determined [2]. The dynamic nature of market demand has driven the development of a flexible and adaptable method termed Flexible conversation Model (FCM) for supply chain management[3]. The key principle of the proposed method is to enable agents representing different parties within a supply chain to exchange and acquire changing information (represented as new conversation policies) and make decisions based on received information. Likewise, a multi-agent coordination mechanism is proposed to address the dynamic variation of demand by representing changes as tokens[4]. Agents react to tokens and negotiate to determine the best decisions for different scenarios. Effective integration and sharing of information is also mandatory to facilitate decision optimization within a supply chain. To address this issue, a multi-agent based architecture that consolidates manufacturer, supplier, and production design centre as an integrated whole is proposed [5].

Besides, updating real time information concerning orders is also another significant issue to be considered within a supply chain, most notably in a make-to-order manufacturing environment. Realizing the importance of this issue, a multi-agent based approach focusing on effective updating and analyzing of real time information is proposed. Likewise, an agent based model for warehouse system is proposed to facilitate order update as well as information update concerning product delivery. Order information flows from customers to warehouses and finally to manufacturers, whereas product delivery information flows from manufacturers back to the warehouses and finally to the customers. A simulation study shows that the coordination of order and supply information keeps inventory at the minimal level i.e., just in time usage and/or delivery of inventory. The Multi-Agent System for Distributed Coordination of Supply Chains (MADC) is proposed to coordinate information sharing throughout the entire supply chain. This includes the sharing of market information between manufacturers and suppliers; sharing of inventory information between manufacturers and customers; as well as observing demand patterns to estimate/predict market demands. The feasibility of MADC is evaluated on a thin Film transistor-Liquid Crystal Display (TFT-LCD) manufacturing plants. Results obtained show MADC is highly dependent on the accuracy of demand estimation in order to effectively minimize the Bullwhip effect. In other words, only when market demand is estimated accurately, Bullwhip Effect is minimized.

The major limitation of existing supply chain related work is that the fundamental issue of a supply chain is not addressed. In its simplest form, a supply chain integrates activities/processes of suppliers, a manufacturer, and customers. Hence, the issue of interoperation between suppliers, the manufacturer, and customers is the key element for a successful supply chain operation and management. Ensuring a smooth negotiation/bidding and coordination mechanism; performing optimization to generate an optimal solution; developing adaptive system; and emphasizing on real time information sharing and updating are incontestably significant but are not sufficient to enable efficient global manufacturing when the fundamental issue is not addressed. Once the fundamental aspect is tackled, any other research will be the complementary to the research on fundamental issue.

3 The architecture of global supply chain

This paper is proposing a multi-agent based supply chain architecture that enables parties within a supply chain to operate in con-function with one another to promote dynamic optimization and reconfiguration of activities in a turbulent market environment. The proposed architecture resembles a multi-agent based network liked architecture with intelligent agents representing nodes of the network (refer to Figure 1). Individual nodes represent parties within a supply chain i.e., suppliers, a manufacturer, and customers. There are sub-nodes within each node and are denoted by sub-agents. The node that represents a manufacturer will have sub-nodes representing its manufacturing plants. For a sub-node that represent a manufacturing plant, there will be sub-nodes representing management, sales department, design department, production shop floor, etc. within the plant. The sub-node that represents shop floor will embed other sub-node representing workers, machine tools, cutters, storage area, etc.

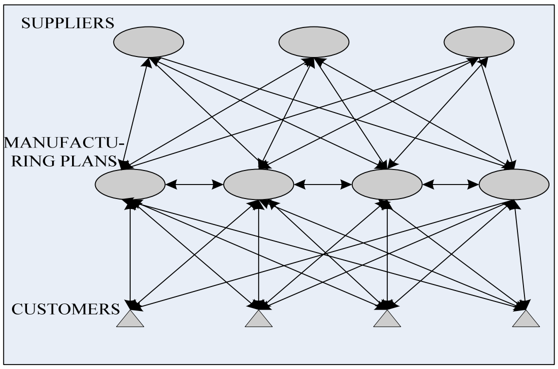


Figure 1 A Dynamic Multi-agent-Based Architecture for Global Supply Chain

In general the proposed network is a set encapsulating three major players within a supply chain-the suppliers represented as Supplier Agents(S), a manufacturer represented as a Manufacturer Agent (M), and customers represented as Customer Agents(C).

$$N = \{S, M, C\} \quad （1）$$

S is the set that encapsulates all Supplier Agents for the Manufacturer Agent, giving S as follow:

$$S = \{S_a, a = 1, 2, \cdots r\} \quad （2）$$

S_a represents a Supplier Agent and r represents the total number of Supplier Agent. Likewise, C is the set that encapsulates all Customer Agents for the Manufacturer agent, giving C as follow:

$$C = \{C_b, b = 1, 2, \cdots u\} \quad （3）$$

C_b represents a Customer Agent and u represents the total number of customer Agent. Within a Manufacturer Agent(M), there are many manufacturing plants represented as Plant Agents(P), hence, giving M as follow:

$$M = \{P_d, d = 1, 2, \cdots v\} \quad （4）$$

P_d represents a Plant Agent and v represents the total number of Plant Agent. Within a Plant Agent, there are multiple departments denoted as Departmental Agent(D), giving P as follow:

$$P = \{D_e, e = 1, 2, \cdots w\} \quad （5）$$

D_e represents a Departmental Agent and w represents the total number of Departmental Agent. The Departmental Agent that represents Production Shop Floor (SF) is a set encoding Resource Agents (R).

$$SF = \{R_f, f = 1, 2, \cdots x\} \quad （6）$$

R_f represents a Resource Agent and x represents the total number of Resource Agent.

The relationships between agents and sub-agents, as well as constraints exist between units within the supply chain will be identified to enhance interactive co-operation between agents for decision making and problem solving. Individual agents constituting the network will be equipped with knowledge/information of their environment. These agents will have the ability to learn from their peers, from changing market or production environment, from decisions they made and outcomes they produced to enhance their knowledge about an environment.

3.1 Internal Architecture of an Agent

The internal architecture of an agent is shown in Figure 2.

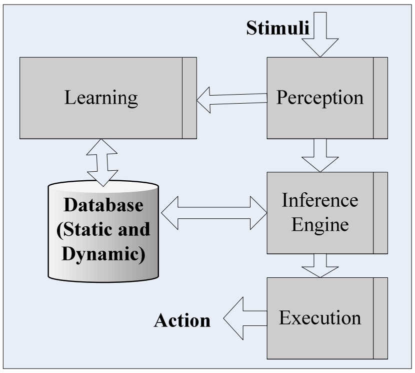


Figure 2 The Internal Architecture of An Agent

Perception Module (P) allows agents to perceive stimuli within its working environment and receive information from other agents or human user. The perceived/received information is forwarded to Learning Module and Inference Engine. Learning Module allows agent to perform self learning. Information is analyzed and new knowledge and/or rules are produced. These knowledge/rules are forwarded to corresponding database for storage or are u8sed to update existing information. Static database stores data that remain unchanged in the short term (i.e., typically in 3 months' time) such as locations of manufacturing plants, number of machines within a plant, etc. on the other hand, Dynamic database stores data that change in short term such as the quantity of weekly orders, the status of machines, etc. which these data require continuous updating. Inference Engine Module allows agents to examine received stimuli and make decisions using data reside within its Static and Dynamic Databases. Execution Module carries out decisions Made by Inference engine Module.

3.2 Communication architecture of agents

The communication architecture of agents is depicted in Figure 3. Supplier Agents (S), Plant Agents (P) and Customer Agents (C) communicate via Internet. Agents residing within a department (the Departmental Agents (D))or within a shop floor (the resource Agents (R)) communicate via Intranet.

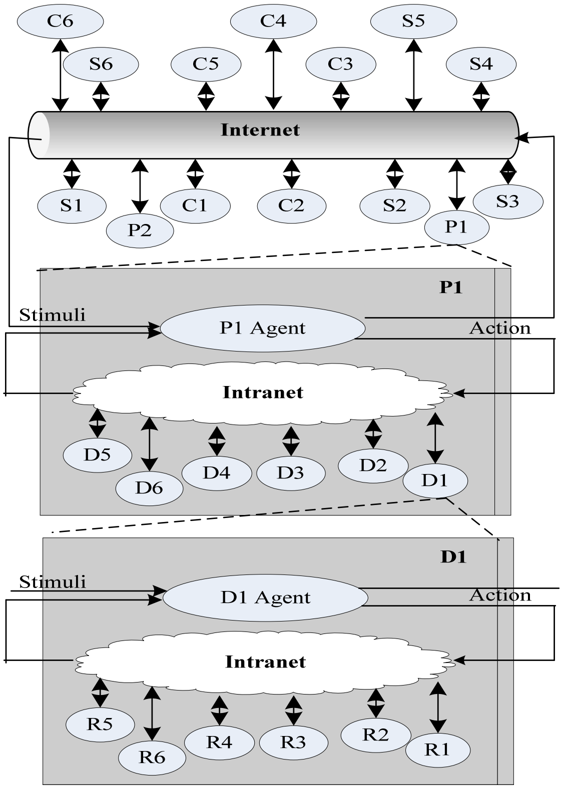


Figure 3 The Communication Architecture of Agents

3.3 The benefits offer by the proposed architecture

Benefits offer by the proposed architecture are:

(i) Improving communication between parties within a supply chain network: reduce communication time; minimize unnecessary delays and errors; improve logistics system; improve activity coordination

(ii) Reducing time to market: maximize productivity; improve customer satisfaction

(iii) Improving system's flexibility and respon siveness towards changing conditions: capitalize market opportunity i.e., new and/or existing market niches; enhance manufacturing efficiency

4 Conclusion

The complexity of global supply chain management with turbulent demand has required a control system to support effective activity integration and coordination. Responding to such need and the limitation of existing methods for supply chain management, a new architecture is proposed in this paper. The proposed architecture addresses the issue of dynamic interoperation of parties within a supply chain to enable efficient global manufacturing. The principle of the proposed architecture will be implemented and verified in a real global manufacturing environment.

References

[1] Prasad, S., Babbar, S. International operations management research. Journal of Operations Management, No.18, 2002, pp209-247

[2] Ulieru, M., Norrie, D., Kremer, R., Shen, W. A. multi-resolution collaborative architecture for web-centric global manufacturing. Information Sciences, Vol.127, （1）, 2000, pp.3-21

[3] Dumond, Y., Roche, C. Formal specification of a multi-agent system architecture for manufacture: the contribution of the π-calculus. Journal of Materials Processing Technology, Vol. 107,（3）, 2000,pp.209-215

[4] Turowski, K. Agent-based E-commerce in case of mass customization. International Journal of Production

Economics, Vo.75, （2）, 2002, pp. 69-81

[5] Kaihara, T. Multi-agent based supply chain modeling with dynamic environment. International Journal of Production Economics, Vol. 85,（2）, 2003,pp. 263-269

[6] Berenji, H. R., Vengerov, D. Learning, Cooperation, and Coordination in Multi-agent Systems. Intelligent Inference Systems Corp. Technical Report IIS-00-10, 2000

[7] Sandholm, T., Equilibrium of the Possibilities of Unenforced Exchange in Multi-agent Systems. University of Massachusetts, 1995

[8] Mark, S. F, MiHai, B., Rune. T., Agent-Oriented Supply Chain Management. The International Journal of Flexible Manufacturing System, 2000,（12）, pp.165-188

[9] Van, D., Parunak, H., What can Agents do in Industry and Why? Forthcoming at CIA.98[C], 1998

[10] Yu, F. X., Ping, T. X., CAO, C., etal., An Improved Mobile Agent Communication Algorithm, CHINESE J.COMPUTERS, 2002, 25,（4）, pp.357-364

Application of the Gray Relation Analysis in Positioning the Clients of Electric Power Market

Lixian Xing   Ling Ma   Cuie Zhang

School of Business and Administration, North China Electric Power University, Baoding, Hebei, 071003,China

E-mail: xlxwhy@sina.com

Abstract

By using the gray relation analysis, we position the clients of the electric power market to determine the different measures taken in order to develop the electric power market in accordance with different clients. High quality services ensure residents use electricity normally, thus promoting of the development of agriculture and electricity as well as verifying the gray relation analysis is practical and feasible in positioning the clients of the electric power market.

Keywords: electric power market; gray system theory; gray relation analysis

1 Introduction

There are many calculation methods [1][2] of positioning the clients of electric power market, but it has been long for researchers that how to make the quantitative analysis of electric power market more accurate. Former multivariate statistical analysis, such as variation analysis, main factor analysis, element analysis, regression analysis and so on, has many drawbacks:

① large amount of data

② normal-distribution-based samples

③ great calculation work, complex progress, high error rate

④ complicated theory, poor system, non-intuitional analysis results, results different from qualitative analysis emerge easily.

2 The process of calculating the degree of grey relation

2.1 Decide the sequence matrix[7]8]

Let Y_0 be the affected sequence and $x_1, x_2, \cdots\cdots x_i$ be the comparison sequence, forming the following matrix:[9][10]

$$\left(Y_0, x_1, x_2, \cdots\cdots x_i\right) = \begin{pmatrix} Y_0(1) & x_1(1) & \cdots & x_i(1) \\ \vdots & \vdots & \vdots & \vdots \\ Y_0(k) & x_1(k) & \cdots & x_i(k) \end{pmatrix}$$

2.2 Use initialization to make the data dimensionless

The initialization formula is

$$x_i'(k) = \frac{x_i(k)}{x_i(1)}, i = 1, \cdots\cdots n, k = 1, \cdots\cdots m \quad (1)$$

Then we get the dimensionless matrix

$$\left(Y_0', x_1', x_2', \cdots\cdots x_i'\right) = \begin{pmatrix} Y_0'(1) & x_1'(1) & \cdots & x_i'(1) \\ \vdots & \vdots & \vdots & \vdots \\ Y_0'(k) & x_1'(k) & \cdots & x_i'(k) \end{pmatrix}$$

While gray relation analysis(GRA) can avoid the shortages mentioned above and the results will not be affected by the partial unknown data in the system. So we can obtain a favorable effect using this analysis

While we are doing the market positioning of the clients of electricity supply enterprises by quantification method.

Gray relation analysis (GRA) is a method brought forward by gray system theory[3][4][5][6]. It is a kind of multi-factor statistical analysis and quantification of degree of correlation, which adopts the degree of similarity of geometric shapes to analyze the weights of all factors.

2.3 Find the difference value matrix

According to the difference sequence formula

$$\Delta_{0i}(k)=\left|Y_0'(k)-x_i'(k)\right|,$$
$$i=1,\cdots,\cdots,n \qquad k=1,\cdots,\cdots,m \tag{2}$$

We get the difference value matrix

$$\begin{pmatrix} \Delta_{01}(1) & \cdots & \Delta_{0n}(1) \\ \vdots & \vdots & \vdots \\ \Delta_{01}(m) & \cdots & \Delta_{0n}(m) \end{pmatrix}$$

Find the maximum value(the largest difference) and minimum value (the smallest difference)presented by $\Delta_{\max}$ and $\Delta_{\min}$

2.4 Calculate the correlation coefficient

$$\zeta_{0i}(k)=\frac{\Delta_{\min}+\rho\Delta_{\max}}{\Delta_{0i}(k)+\rho\Delta_{\max}} \tag{3}$$

in which ρ is the resolution coefficient $\rho\in[0,1]$,usually we let ρ =0.5, then we obtain the correlation matrix

$$\begin{pmatrix} \zeta_{01}(1) & \cdots & \zeta_{0n}(1) \\ \vdots & \vdots & \vdots \\ \zeta_{01}(m) & \cdots & \zeta_{0n}(m) \end{pmatrix}$$

2.5 Calculate the degree of correlation

$$r_{0i}=\frac{1}{m}\sum_{k=1}^{m}\zeta_{0i}(k) \tag{4}$$

2.6 Sort the degree of correlation and decide the weight of all factors

(The more degree of correlation there is the more it weighs, otherwise the contrary)

3 Example of market positioning of the clients of electric power supply enterprise in Shandong province

Let's take the consumption of country clients in Shandong Province published by Shandong Electricity Power Enterprise as example (see table 3-1).

Table 3-1 consumption of country clients in Shandong Province (hundred million kilowatt-hours)

Year	Gross Electricity Consumption	Agriculture Irrigation and Drainage	Agricultural sideline Processing
1997	163.58	51.93	14．98
1998	171.09	52.69	15.32
1999	195.63	62.98	16.98
2000	208.48	64.50	17.30
2001	235.30	74.21	16.24
2002	262.16	85.44	15.38
2003	285.25	79．00	15.13
Year	Country Enterprises	Other uses in the Country	Country Residents Living
1997	50.48	5.33	40.86
1998	54．00	5.78	43.31
1999	62.23	6.36	47.08
2000	70.13	5.97	50.58
2001	80.46	7.28	57.11
2002	91.83	7.97	61.54
2003	115.09	9.55	66.49

Use Gray relation analysis to analyze the market positioning of electric power clients and take different measures in accordance with different c.

3.1. Decide the sequence matrix of market position of electric power clients

Let affected sequence represent the gross country consumption and comparison sequence represent industries consumption. Then we set the sequence matrix:

$$\begin{pmatrix} 163.58 & 51.93 & 14.98 & 50.48 & 5.33 & 40.86 \\ 171.09 & 52.69 & 15.32 & 54 & 5.78 & 43.31 \\ 195.63 & 62.98 & 16.98 & 62.23 & 6.36 & 47.08 \\ 208.48 & 64.5 & 17.3 & 70.13 & 5.97 & 50.58 \\ 235.3 & 74.21 & 16.24 & 80.46 & 7.28 & 57.11 \\ 262.16 & 85.44 & 15.38 & 91.83 & 7.97 & 61.54 \\ 285.25 & 79 & 15.13 & 115.09 & 9.55 & 66.49 \end{pmatrix}$$

3.2 Get the dimensionless matrix of market position of electric power clients by initia lization

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1.046 & 1.015 & 1.023 & 1.070 & 1.084 & 1.060 \\ 1.196 & 1.213 & 1.134 & 1.233 & 1.193 & 1.152 \\ 1.274 & 1.242 & 1.155 & 1.389 & 1.120 & 1.238 \\ 1.438 & 1.429 & 1.084 & 1.594 & 1.366 & 1.398 \\ 1.603 & 1.645 & 1.027 & 1.819 & 1.495 & 1.506 \\ 1.744 & 1.521 & 1.010 & 2.280 & 1.792 & 1.627 \end{pmatrix}$$

3.3 Get the absolute difference matrix of market position of clients by difference sequence formula

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.031 & 0.023 & 0.024 & 0.039 & 0.014 \\ 0.017 & 0.062 & 0.037 & 0.003 & 0.044 \\ 0.032 & 0.120 & 0.115 & 0.154 & 0.037 \\ 0.009 & 0.354 & 0.155 & 0.073 & 0.041 \\ 0.043 & 0.576 & 0.216 & 0.107 & 0.097 \\ 0.223 & 0.734 & 0.563 & 0.048 & 0.117 \end{pmatrix}$$

By difference matrix we know $\Delta_{\max} = 0.734$ and $\Delta_{\min} = 0$.

3.4 Get the correlation matrix of market position by correlation coefficient formula

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0.922 & 0.941 & 0.939 & 0.904 & 0.963 \\ 0.956 & 0.855 & 0.908 & 0.992 & 0.893 \\ 0.920 & 0.754 & 0.761 & 0.704 & 0.908 \\ 0.976 & 0.509 & 0.703 & 0.834 & 0.900 \\ 0.895 & 0.389 & 0.630 & 0.774 & 0.791 \\ 0.622 & 0.333 & 0.406 & 0.884 & 0.758 \end{pmatrix}$$

3.5 Get the degree of correlation of market position of clients by degree of correlation formula

$r_{01} = 0.899$ Agriculture Irrigation and Drainage (principal client)

$r_{05} = 0.888$ Country Residents Living (major client)

3.6 Sort the degree of correlation of market position of clients

The result is $r_{01} > \gamma_{05} > r_{04} > r_{03} > r_{02}$,that is to say, the weight of market position of clients is: Agriculture Irrigation and Drainage, Country Residents Living, other uses in the country, country enterprises and Agricultural sideline Processing

4 Measures taken for clients in different market position

① In the country clients of Shandong Province, due to the electricity consumption, the part of agriculture irrigation and drainage grows fast with a degree of grey relation reaching 0.899, thus making it have a great affection on the gross electricity consumption. As top client in the country, it should be managed as the principal clients as well. Since Shandong is a one of the major agricultural provinces and the climate of most of its area is dry. With our state's inclining to agriculture, the electric power consumption of agriculture irrigation and drainage will be going up in future. So change the way how agriculture irrigation and drainage consumes electric power and require farmers to choose efficient electrical equipments, use efficient and energy saving material, make reasonable work schedule and change their consuming behaviors ,in order to reduce the fluctuation of the grid load and ensure agriculture and electric power develop in harmony.

② The degree of gray relation of country residents living is 0.888, which is also very important to the gross electricity consumption and should be treated as the major clients in the electric power consumption management. With the rapid growth of economy of our nation, more household electrical appliances (e.g. family lighting, cooking and heating) will come into families and the electricity consumption will go up dramatically. So it is a must for the electric power supply enterprises to cultivate country residents` management awareness of electric power consumption. By conducting them with education campaigns, require them to reduce the demand for electric power during the peak of grid load, according to the load property of electric power system, and transfer or add it to the low point to change the timing distribution of electric power consumption.

$r_{04} = 0.870$ Other uses in the country (important client)

$r_{03} = 0.764$ Country enterprises (general client)

$r_{02} = 0.683$ Agricultural sideline processing (potential client)

③ Although the proportion of electric power consumption of other uses in the country is not so much, however, since it has a high degree of gray relation, it should be regarded as the important client. Meanwhile, with the growth of our society and the people` s demand, there exists a big potential of this market .So while electric power supply enterprises are managing the electric power consumption of these uses, they must take a series of measures to balance the demand for electric power in this market.

④ On contrast with other uses in the country, country enterprise has a relatively low degree of gray relation but plays the very most important role in the gross electric power consumption, and it should be seen as the general client. This matter has something with the local commercial economy, which is still developing. And with the adjustment of our national economic structure and the rising of country enterprises, electric power supply enterprises should concern themselves with the trend, and offer high quality services to country enterprise.

⑤ Agricultural sideline processing has the lowest degree of gray relation and also takes the smallest share of the market, and this makes it the potential client. But since the country economic development is going on, the electric power consumption of agricultural sideline processing will must get some promotion, and the electric power supply enterprises should pay attention it.

Referencrs

[1] Qin Shoukang, Principle and Application of Comprehensive Evaluation [M].Publishing House of Electricity Industry, 2003

[2] Hu Yonghong Etc. Method of Comprehensive Evaluation [M]. Science Press,2000

[3] Deng Julong. Basic Method of Gray System [M].Wuhan, Huazhong: University of Science and Technology (UHST) Press,1987

[4] Tan Xuerui, Deng Julong. Gray Relation Analysis: New Method Multivariate statistical analysis[J]. Statistical Research, 1995,65（3）:pp.46-48

[5] Chen Li,Niu Dong-xiao,Li jun. Improved gray relevant analysis of influencing factors evaluation in electric power consumption[J]. Journal of North China Electric Power University, 2003,30（1）:pp.61-64

[6] Ji peirong The research of the characteristics of Gray forecast model, Systems engineering theory and practice 2001,9:105-108

[7] DAI Debao, CHEN Rongqiu. Journal of Grey Systems, 2001,Frame of IAGO Space[J].The Journal of Grey Systems 2001 (I): 9-12

[8] Luo Dang, Liu Sifeng, Dang Yaoguo. The optimization of grey model GM(1,1) [J] • Engineering Science, 2003, 5（8）: 50-53

[9] Mian-yun Chen. Principle of grey dynamic modeling[J]. SAMS, 1996, 26,69-79

[10] Deng Julong. Moving operator in grey theory[J].The Journal of Grey System,1999（1）:1-5

A Performance-optimizing Dependable Virtual Storage Scheme*

Ming Hu[1]   Minghua Jiang[2]

College of Computer Science, Wuhan University of Science and Engineering, Wuhan, Hubei, 430073, China

Email: 1 stereotype@263.net; 2 mhjiang@126.com

Abstract

A performance-optimizing dependable virtual storage scheme is proposed to improve both the data-accessing performance and availability for distributed storage systems to meet user's different requirements. The storage system improves the storage security by encrypting command message between servers and storage nodes. It also provides the two-level address-mapping and request-decomposing computation to hide the implementation of storage nodes, to increase the capability of distributed computation, and to implement two types of storage layout. The scheme takes not only both homo- and hetero-geneous storage nodes into account, but also protect the integrity of the storage system behaviors based on distributed environments.

Keywords：Distributed storage; dependable network; virtual storage; scheme; and performance-optimizing

1 Introduction

Disk array [1-2] for distributed systems is a popular method to improve their performance and reliability and its hardware cost is quite low. For the ordinary disk array, Plank J S and Xu Lihao [3] presented the optimization method for Reed-Solomon codes which require complex finite field arithmetic. J. L. Hafner [4-5] presented XOR-based WEAVER and HoVer codes with lower storage efficiency to improve the performance. A.Thomasian [6] presented multi-level RAID (MRAID), like RAID 1/5, RAID5/5 and RAID6/5, for storage nodes. RAID5x [7] further improved the performance of the double fault tolerant storage systems with higher storage efficiency.

Virtual storage has been defined as a type of technology which makes difference between the storage description of the server operating system and physical storage devices. The research on virtual storage technologies focuses on the storage- and file-level schemes against multiple failures to improve the data-accessing performance of storage devices in the distributed computing environments. In such a distributed storage system, the method to construct a disk array is to use single I/O space image which can be implemented at driver, user, and file system level.

This paper will use storage-zoning idea to generalize RAID5x and RAID0 into the distributed environment. However, some studies [8-10] shows that the distributed networked environment is not secure to the computation and request operations of distributed storage system against manual attacks. This paper presents the scheme of a distributed homogeneous storage node array, and then generalizes it into a dependable virtual storage scheme considering both the homo- and heterogeneous storage configuration. To implement such a dependable virtual storage scheme, it is necessary to support three-level accessing mechanism and two-level address mapping and request-decomposing, and to protect the integrity of

* Project Supported By HuBei Natural Science Foundation under Grant NO.2007ABA376 and Hubei Education Foundation under Grant NO. 200717005.

storage system behaviors. The virtual storage scheme provides the maximal flexibility for upgrading, scaling up, and configuring a distributed storage system to support the data-accessing performance and availability according to users' different requirements.

2 A Distributed Node Array

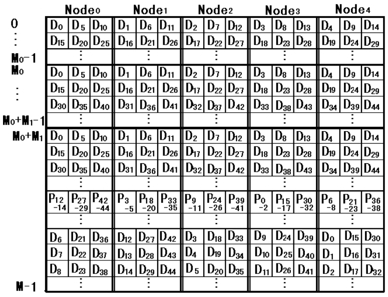
Without loss of generality, consider N independent homogeneous storage nodes, each of which contains G disks and each disk contains M stripe units denoted as S(i,j,k), where $0≦i≦N-1$, $0≦j≦G-1$, and $0≦k≦M-1$. For example, the j-th disk in the i-th node has M stripe units denoted as S(i,j,0), S(i,j,1), …, S(i,j,M-1). From user's point of view, some data is transient or comes from the other sources, so it is not necessary to implement fault tolerance based on data storage. Therefore, the storage system needs to provide two data storage mechanisms without and with redundancy. To implement two data storage mechanisms, we use storage-zoning ideas to generalize RAID0 and RAID5x into the distributed environments. Therefore, according to user's data requirements, user can store the data to one storage zone without redundancy or another with redundancy.

RAID5x presents a mirror-and-parity code mixing placement for disk arrays and is generalized into N independent storage node array to implement double node fault tolerance. Each N-2 logical data stripe units needs N nodes to store data and redundant information, where N-2 nodes are used to store data of N-2 stripe units, each of which is stored in an independent node, one of more two nodes stores the parity information about the N-2 data stripe units, and another stores copies of the N-2 data stripe units sequentially one after another to guarantee the accessing efficiency of small requests. In order to guarantee even distribution of parity information across nodes, a node-periodical stripe needs N parity stripe units, one in each node.

For a generalized RAID5x, each N-2 data stripe units generate a parity and N-2 copies, and N parity stripe units need $m=(N-2)×N$ for storing data and $f=N+(N-2)×N=(N-1)×N$ for storing redundant information, so these data, parities and copies form a node-periodical stripe with the size of $m+f=(2N-3)×N$. RAID0 is simple to be generalized into N storage node array. Because of no redundancy, we can simply designate $m=(N-2)×N$ for storing data and $f=0$ for storing redundant information. Let $M=M_0+M_1+M_2$, $M_0=L$ for storage system information, $M_1=H×(N-2)$ for RAID0, and $M_2=R×(2N-3)$ for RAID5x. For above storage array with N nodes, the number T of node-periodical stripe at most is $[(M_2×N×G)/((2N-3)×N)]=R×G$ where [x] is the greatest integer less than or equal to x. For such a scheme, the storage efficiency $u=(M-R(N-1))×N×G/(M×N×G)=1- R(N-1)/M$.

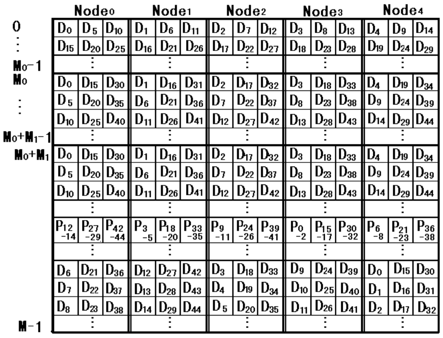
Logically, the storage node array has three storage area: system area with the range A_0, RAID0 area with the range A_1, and RAID5x area with the range A_2, where $0≦A_0≦ L×G×N-1$, $0≦A_1≦H×G×(N-2)×N-1$, $0≦A_2≦R×G×(N-2) ×N-1$. Physically, RAID5x area has three areas: data, parity, and mirrored area. Parity information and mirrored copies store on the same disk in a node. The key is data area about RAID0 and RAID5.

This paper proposes two data-storing mechanisms: disk clustering and disk de-clustering. For disk clustering placement, the next disk can store the subsequent data stripe units only after the last disk fills with the N-2 data stripe units and this process continues cyclically on the G disks in each node. For disk-de-clustering placement, the consecutive data can be stored one stripe unit after another cyclically on G disks in each node. For the generalized RAID5x, G node-periodical stripes forms a global periodical stripe of (2N-3)×N×G stripe units.



	Node0			Node1			Node2			Node3			Node4		
0	D_0	D_5	D_{10}	D_1	D_6	D_{11}	D_2	D_7	D_{12}	D_3	D_8	D_{13}	D_4	D_9	D_{14}
⋮	D_{15}	D_{20}	D_{25}	D_{16}	D_{21}	D_{26}	D_{17}	D_{22}	D_{27}	D_{18}	D_{23}	D_{28}	D_{19}	D_{24}	D_{29}
M_0-1		⋮			⋮			⋮			⋮			⋮	
M_0	D_0	D_5	D_{10}	D_1	D_6	D_{11}	D_2	D_7	D_{12}	D_3	D_8	D_{13}	D_4	D_9	D_{14}
⋮	D_{15}	D_{20}	D_{25}	D_{16}	D_{21}	D_{26}	D_{17}	D_{22}	D_{27}	D_{18}	D_{23}	D_{28}	D_{19}	D_{24}	D_{29}
	D_{30}	D_{35}	D_{40}	D_{31}	D_{36}	D_{41}	D_{32}	D_{37}	D_{42}	D_{33}	D_{38}	D_{43}	D_{34}	D_{39}	D_{44}
M_0+M_1-1		⋮			⋮			⋮			⋮			⋮	
M_0+M_1	D_0	D_5	D_{10}	D_1	D_6	D_{11}	D_2	D_7	D_{12}	D_3	D_8	D_{13}	D_4	D_9	D_{14}
	D_{15}	D_{20}	D_{25}	D_{16}	D_{21}	D_{26}	D_{17}	D_{22}	D_{27}	D_{18}	D_{23}	D_{28}	D_{19}	D_{24}	D_{29}
	D_{30}	D_{35}	D_{40}	D_{31}	D_{36}	D_{41}	D_{32}	D_{37}	D_{42}	D_{33}	D_{38}	D_{43}	D_{34}	D_{39}	D_{44}
		⋮			⋮			⋮			⋮			⋮	
	P_{12-14}	P_{27-29}	P_{42-44}	P_{3-5}	P_{18-20}	P_{33-35}	P_{9-11}	P_{24-26}	P_{39-41}	P_{0-2}	P_{15-17}	P_{30-32}	P_{6-8}	P_{21-23}	P_{36-38}
		⋮			⋮			⋮			⋮			⋮	
	D_6	D_{21}	D_{36}	D_{12}	D_{27}	D_{42}	D_3	D_{18}	D_{33}	D_9	D_{24}	D_{39}	D_0	D_{15}	D_{30}
	D_7	D_{22}	D_{37}	D_{13}	D_{28}	D_{43}	D_4	D_{19}	D_{34}	D_{10}	D_{25}	D_{40}	D_1	D_{16}	D_{31}
	D_8	D_{23}	D_{38}	D_{14}	D_{29}	D_{44}	D_5	D_{20}	D_{35}	D_{11}	D_{26}	D_{41}	D_2	D_{17}	D_{32}
$M-1$		⋮			⋮			⋮			⋮			⋮	

Fig.1 Node layout with N=5, G=3, and disk de-clustering for RAID0 and RAID5x



	Node0			Node1			Node2			Node3			Node4		
0	D_0	D_5	D_{10}	D_1	D_6	D_{11}	D_2	D_7	D_{12}	D_3	D_8	D_{13}	D_4	D_9	D_{14}
⋮	D_{15}	D_{20}	D_{25}	D_{16}	D_{21}	D_{26}	D_{17}	D_{22}	D_{27}	D_{18}	D_{23}	D_{28}	D_{19}	D_{24}	D_{29}
M_0-1		⋮			⋮			⋮			⋮			⋮	
M_0	D_0	D_{15}	D_{30}	D_1	D_{16}	D_{31}	D_2	D_{17}	D_{32}	D_3	D_{18}	D_{33}	D_4	D_{19}	D_{34}
⋮	D_5	D_{20}	D_{35}	D_6	D_{21}	D_{36}	D_7	D_{22}	D_{37}	D_8	D_{23}	D_{38}	D_9	D_{24}	D_{39}
	D_{10}	D_{25}	D_{40}	D_{11}	D_{26}	D_{41}	D_{12}	D_{27}	D_{42}	D_{13}	D_{28}	D_{43}	D_{14}	D_{29}	D_{44}
M_0+M_1-1		⋮			⋮			⋮			⋮			⋮	
M_0+M_1	D_0	D_{15}	D_{30}	D_1	D_{16}	D_{31}	D_2	D_{17}	D_{32}	D_3	D_{18}	D_{33}	D_4	D_{19}	D_{34}
	D_5	D_{20}	D_{35}	D_6	D_{21}	D_{36}	D_7	D_{22}	D_{37}	D_8	D_{23}	D_{38}	D_9	D_{24}	D_{39}
	D_{10}	D_{25}	D_{40}	D_{11}	D_{26}	D_{41}	D_{12}	D_{27}	D_{42}	D_{13}	D_{28}	D_{43}	D_{14}	D_{29}	D_{44}
		⋮			⋮			⋮			⋮			⋮	
	P_{12-14}	P_{27-29}	P_{42-44}	P_{3-5}	P_{18-20}	P_{33-35}	P_{9-11}	P_{24-26}	P_{39-41}	P_{0-2}	P_{15-17}	P_{30-32}	P_{6-8}	P_{21-23}	P_{36-38}
		⋮			⋮			⋮			⋮			⋮	
	D_6	D_{21}	D_{36}	D_{12}	D_{27}	D_{42}	D_3	D_{18}	D_{33}	D_9	D_{24}	D_{39}	D_0	D_{15}	D_{30}
	D_7	D_{22}	D_{37}	D_{13}	D_{28}	D_{43}	D_4	D_{19}	D_{34}	D_{10}	D_{25}	D_{40}	D_1	D_{16}	D_{31}
	D_8	D_{23}	D_{38}	D_{14}	D_{29}	D_{44}	D_5	D_{20}	D_{35}	D_{11}	D_{26}	D_{41}	D_2	D_{17}	D_{32}
$M-1$		⋮			⋮			⋮			⋮			⋮	

Fignre2 Node layout with N=5, G=3, and disk clustering for RAID0 and RAID5x

The physical layout of storage areas for the above storage node array can be obtained by the following computation. For the system information, the address mapping for Simple RAID0 is f_{Sdd}: $A_0 \rightarrow S(i,j,k)$ where $i=A_0 \mathrm{mod} N$, $j=[A_0/N)]$ modG, and $k=[A_0/(N \times G)]$.

For the generalized RAID0, the address mapping for disk de-clustering is f_{Hdd}: $A_1 \rightarrow S(i,j,k)$ where $i=A_1 \mathrm{mod} N$, $j=[A_1/N]$ modG, and $k=L+[A_1/(N \times G)]$ and the address mapping for disk clustering is f_{Hdc}: $A_1 \rightarrow S(i,j,k)$ where $i=A_1 \mathrm{mod} N$, $j=[A_1/(N \times (N-2))]$modG, and $k=L+[A_1/N]\mathrm{mod}\ (N-2) +[A_1/(N \times G \times (N-2))] \times (N-2)$.

For the generalized RAID5x, the address mapping of data area for disk de-clustering is f_{Rdd}: $A_2 \rightarrow S(i,j,k)$ where $i= A_2 \mathrm{mod} N$, $j=[A_2/N]$modG, and $k=L+H \times (N-2) +[A_2/(N \times G)]$ and the address for disk clustering is f_{Rdc}: $A_2 \rightarrow S(i,j,k)$ where $i= A_2 \mathrm{mod} N$, $j=[A_2/(N \times (N-2))]$modG, and $k=L+ H \times (N-2)+ [A_2/N]\mathrm{mod}(N-2)+[A_2/(N \times G \times (N-2))] \times (N-2)$.

Whether the generalized RAID5x applies disk de-clustering or clustering, the address mapping of parity area or mirrored area is the same. The logical address A_2 can be used to computing the physical address $S(i_p,j_p,k_p)$ of the parity information and the physical address $S(i_C,j_C,k_C)$ of the mirrored copy as follows:

For the parity information, f_p: $A_2 \rightarrow S(i_p,j_p,k_p)$ where

$$i_p = \begin{cases} \{([A_2/(N-2)] \mathrm{mod}\ N+1) \times (N-2)\} \mathrm{mod}\ N \\ \{([A_2/(N-2)] \mathrm{mod}\ N+1) \times (N-2)\} \mathrm{mod}\ N+1, \\ N \textit{ is even and } [A_2/(N-2)] \mathrm{modN} \geq N/2 \end{cases}$$

$j_p=[A_2/(N \times (N-2))]$modG, and $k_p= L+H \times (N-2)+R \times (N-2) + [A_2/(N \times G \times (N-2))]$.

For the mirrored copy, f_C: $A_2 \rightarrow S(i_C,j_C,k_C)$ where

$$i_C = \begin{cases} \{([A_2/(N-2)] \mathrm{mod}\ N+1) \times (N-2)\} \mathrm{mod}\ N+1 \\ \{([A_2/(N-2)] \mathrm{mod}\ N+1) \times (N-2)\} \mathrm{mod}\ N, \\ N \textit{ is even and } [A_2/(N-2)] \mathrm{modN} \geq N/2 \end{cases}$$

$j_c=[A_2/(N \times (N-2))]$modG, and $k_p= L+H \times (N-2)+R \times (N-1)+ [A_2/(N \times G \times (N-2))] \times (N-2)+ A_2 \mathrm{mod}(N-2)$.

Fig.1 and Fig.2 illustrate the storage scheme with the three areas respectively to use disk de-clustering and disk clustering for both RAID0 and RAID5x.

3 The Dependable Virtual Scheme

The scheme described above is a homogeneous storage system with a fixed single-level address mapping. It is not convenient to configure, upgrade, and scale up a storage system. Therefore, this paper proposes a virtual storage scheme with two-level address mapping and request decomposing. Even if there are no disks of the same type to displace the failed disks, the homogeneous storage system with the virtual storage scheme can reconfigure into a heterogeneous storage. It is helpful to upgrade and scale up the storage system. Such a storage system needs three-level data-accessing mechanism: clients, data servers, and storage nodes. Data servers are responsible to manage, configure, and protect the storage system.

To implement two-level address mapping, the above mapping has to be divided into the node address mapping and disk address mapping. In what follows, we describe them respectively. For the system area, the node mapping is S_N: $A_0 \rightarrow S(i,m_0)$ where $i=A_0 \mathrm{mod} N$, $m_0=[A_0/N)]$, and $0 \leqq m_0 \leqq L \times G-1$. The disk mapping is S_d: $m_0 \rightarrow S(j,k)$ where $j=m_0$ modG and $k=[m_0/G]$. For the generalized RAID0 area, the node mapping is H_N: $A_1 \rightarrow S(i,m_1)$ where $i=A_1 \mathrm{mod} N$, $m_1=[A_1/N)]$, and $0 \leqq m_1 \leqq H \times G \times (N-2)-1$. The disk mapping for disk de-clustering is H_{dd}: $m_1 \rightarrow S(j,k)$ where $j=m_1$ modG and $k= L+[m_1/G]$ and the disk mapping for disk clustering is H_{dc}: $m_1 \rightarrow S(j,k)$ where $j=[m_1/(N-2)]$modG and $k=L+m_1 \mathrm{mod}\ (N-2)+[m_1/(G \times (N-2))] \times (N-2)$. For the generalized RAID5x area, the node mapping of the data sub-area is R_N: $A_2 \rightarrow S(i,m_2)$ where $i=A_2 \mathrm{mod} N$, $m_2=[A_2/N)]$, and $0 \leqq m_2 \leqq R \times G \times (N-2)-1$. The node mapping of the parity sub-area is p_N: A_2→ $S(i_p,m_p)$ where

$$i_p = \begin{cases} \{([A_2/(N-2)] \bmod N+1)\times(N-2)\} \bmod N \\ \{([A_2/(N-2)] \bmod N+1)\times(N-2)\} \bmod N+1, \\ N \; \textit{is even and} \; [A_2/(N-2)] \bmod N \geq N/2 \end{cases}$$

$m_p=[A_2/(N\times(N-2))]$, and $0 \leqq m_p \leqq R\times G-1$. The node mapping of the mirrored sub-area is C_N: A_2→ $S(i_C,m_C)$ where

$$i_C = \begin{cases} \{([A_2/(N-2)] \bmod N+1)\times(N-2)\} \bmod N+1 \\ \{([A_2/(N-2)] \bmod N+1)\times(N-2)\} \bmod N, \\ N \; \textit{is even and} \; [A_2/(N-2)] \bmod N \geq N/2 \end{cases}$$

$m_C=[A_2/(N\times(N-2))]\times(N-2)+ A_2 \bmod (N-2)$, and $0 \leqq m_C \leqq R\times G\times(N-2)-1$. The disk mapping of the data sub-area for disk de-clustering is R_{dd}: m_2→S(j,k) where $j=m_2$ modG and $k= L+H\times(N-2)+[m_2/G]$ and the disk mapping for disk clustering is R_{dc}: m_2→S(j,k) where $j=[m_2/(N-2)]$modG and $k=L+H\times(N-2)+m_2 \bmod(N-2)+[m_2/(G\times(N-2))]\times(N-2)$. The disk mapping of the parity sub-area is p_d: m_p→$S(j_p,k_p)$ where $j_p=m_p$modG and $k_p=L+H\times(N-2)+R\times(N-2)+[m_p/G]$. The disk mapping of the mirrored sub-area is C_d: m_C→$S(j_C,k_C)$ where $j_C=[m_C/(N-2)]$modG and $k_C=L+H\times(N-2)+R\times(N-1)+ m_C \bmod(N-2)+ [m_C/(G\times(N-2))]\times (N-2)$.

For the above mapping, they form the two-level address access mechanism. For the system area, the composition of S_N: A_0→$S(i,m_0)$ and S_d: m_0→S(j,k) forms the mapping f_{Sdd}: A_0→S(i,j,k). For the generalized RAID0 area, the composition of H_N: A_1→$S(i,m_1)$ and H_{dd}: m_1→S(j,k) forms f_{Hdd}: A_1→S(i,j,k) while the composition of H_N: A_1→$S(i,m_1)$ and H_{dc}: m_1→S(j,k) forms f_{Hdc}: A_1→S(i,j,k). For the generalized RAID5x area, the composition pairs of (R_N: A_2→$S(i,m_2)$, R_{dd}: m_2→S(j,k)), (p_N: A_2→$S(i_p,m_p)$, p_d: m_p→$S(j_p,k_p)$), and (C_N: A_2→$S(i_C,m_C)$, C_d: m_C→$S(j_C,k_C)$) form the mapping triple (f_{Rdd}: A_2→S(i,j,k), f_p: A_2→ $S(i_p,j_p,k_p)$, f_C: A_2→ $S(i_C,j_C,k_C)$) . Likely, the composition pairs of (R_N: A_2→$S(i,m_2)$, R_{dc}: m_2→S(j,k)), (p_N: A_2→$S(i_p,m_p)$, p_d: m_p→$S(j_p,k_p)$), and (C_N: A_2→ $S(i_C,m_C)$, C_d: m_C→$S(j_C,k_C)$) form the mapping triple (f_{Rdc}: A_2→S(i,j,k), f_p: A_2→ $S(i_p,j_p,k_p)$, f_C: A_2→ $S(i_C,j_C,k_C)$) .

To take the heterogeneous storage node into account, all the node mappings are the same as in the homogeneous storage system. However, in contrast with the other storage nodes, the disk mapping in the i-th node such as S^i_d: m_0→S(j,k), H^i_{dx}: m_1→S(j,k), R^i_{dx}: m2→S(j,k), p^i_d: m_p→S(j,k), and C^i_d: m_C→S(j,k) can be implemented in the distinct ways where the subscript x denotes the letter d for disk de-clustering or the letter c for disk clustering. The composition pairs of (S_N: A_0→$S(i,m_0)$, S^i_d: m_0→S(j,k)), (H_N: A_1→$S(i,m_1)$, H^i_{dx}: m_1→S(j,k)), (R_N: A_2→$S(i,m_2)$, R^i_{dx}: m_2→S(j,k)), (p_N: A_2→ $S(i_p,m_p)$, p^i_d: m_p→S(j,k)), and (C_N: A_2→ $S(i_C,m_C)$, C^i_d: m_C→S(j,k)) form the mapping sequence {S^i_{Nd}: A_0→S(i,j,k), H^i_{Ndx}: A_1→S(i,j,k), (R^i_{Ndx}: A_2→S(i,j,k), p^i_{Nd}: A_2→ $S(i_p,j_p,k_p)$, C^i_{Nd}: A_2→ $S(i_C,j_C,k_C)$)}. They support the data access of three areas: system, generalized RAID0, and generalized RAID5x. For clients, the storage system has only two areas: one for the generalized RAID0 and another for the generalized RAID5x. The system area serves the configuration and management of the storage system and is not visible to storage users.

To complete data-accessing request, a large request has to be divided into multiple smaller requests to multiple nodes and each request in a node can be subdivided into smaller sub-requests to multiple disks. Therefore, it is necessary to implement two-level request decompositions: one for storage node and another for disk. The request decomposition for storage node is completed by data servers while the request decomposition for disk is completed by the correspondent storage node. In the normal state, the read request decomposition for the generalized RAID5x is the same as for the generalized RAID0. The write request decomposition of the generalized RAID5x is more complicated than that of the generalized RAID0 and is involved with multiple operations of redundant information in the different storage nodes. Two-level address mapping provides the two-level request decomposition with address information. For example, the write request of the generalized RAID5x needs the address sequence {i, i_p, i_C} respectively associated with the data, parity, and mirrored area in the first level request decomposition.

In the networked environment, except for the failed nodes and failed disks, the storage system is possible to confront the various manual attacks anytime. The data for tolerating double node to fail is very important. Once their request decomposing is destroyed manually, the layout against double node fault won't work. To implement the dependable virtual storage mechanism, the role of the data servers is the key. The data servers can be used not only as the controller of the storage node array, but also as the roots of the storage system trusted chains. They support the secure access between clients and data servers, between clients and storage nodes, and between storage nodes. To protect the integrity of the storage system behaviors, the key is the authentication between physical devices, the data of the system area, and the controlling commands related to storage-accessing requests. The system area includes the global information of the storage system and the local information in each node. The global information is involved with the configuration information and the parameters such as N, L, H, R and the like. The local information in each node is associated with the command-executing log in the node in order to learn about the information of command completion. When the information related to the system area is transmitted between data servers and storage nodes or between storage nodes, they need to be encrypted.

For the data of the generalized RAID0 or RAID5x area, the storage system is unnecessary to promise the confidentiality for the data. Therefore, such data are not encrypted during the transmission. However, the controlling commands related to the node mapping and request decomposition for some storage node should be encrypted and at the same time append to its massage authentication code during the transmission. If the encrypted command is corrupted by manual attacks, the receiver can find the errors. There are five types of command associated with three areas and are related to node mappings such as S_N: $A_0 \rightarrow S(i,m_0)$, H_N: $A_1 \rightarrow S(i,m_1)$, R_N: $A_2 \rightarrow S(i,m_2)$, p_N: $A_2 \rightarrow S(i_p,m_p)$, C_N: $A_2 \rightarrow S(i_C,m_C)$. When each node receive a correct command, it check the range of m_x associated with node mapping by using the information of the system area in the node where x can denote one of the labels *0*, *1*, *2*, *p*, and *C*. Once it passed the examination, the controlling command executes and the node provides the integrity of operations by reading from or writing to the system area.

4 Conclusions

The dependable virtual storage scheme proposed above is the generalization of the distributed storage node array, hides the complexity of storage nodes by applying the two-level address mapping and request decomposition, and protect the integrity of storage system behaviors by the access authentication between devices, the encryption of system information in transmission, and the encryption and authentication of controlling commands in transmission. Each node has its own disk mapping, so it has the correspondent strategy to treat the failed disk. It provides two areas: the generalized RAID0 area and the generalized RAID5x area for user to choose. The security for users' data is solved by the negotiation between clients and data servers. As a result, the storage system is able to survive the manual attacks and has the low impact on the storage-accessing performance.

For the generalized RAID5x area, the storage efficiency is (N-2)/(2N-3), where N is the number of nodes. It is higher than the storage efficiency (N-1)/(2N) of RAID5+1 and is less than the storage efficient (N-2)/N of RAID6-like layout. However, the storage efficiency of the virtual storage scheme is 1-R(N-1)/M, where R(2N-3)<M is related to the storage configuration. With the increase of disk volume, the storage efficiency of the virtual storage scheme increases when the requirements of the generalized RAID5x is fixed.

Under the normal state, the read request accessing performance in the generalized RAID5x is the same as in the generalized RAID0 because of disk-zoning methods and the small requests have the better performance because of the mirrored copies of the generalized RAID5x. For the generalized RAID0 and the generalized RAID5x, they have at the least the parallel degree N. Under the configuration of homogeneous storage, disk clustering within a global periodical stripe has the maximum N of data-accessing parallel degree in a single read request, has the complex mapping mechanism and the higher request-coalescing capability to implement the concurrent processing of multiple requests under condition of maximal parallel degree N. Disk de-clustering with a global periodical stripe has the maximum N×G of data-accessing parallel degree in a single read request like RAID0 and has a simple mapping mechanism. Under the configuration of heterogeneous storage, each storage node changes the inner layout according to the characteristics of the different disks to optimize the disk-accessing performance.

Table 1 Double fault based read operations for the generalized RAID5x with N=5

PFN	NPC	NNS
$Node_0$, $Node_1$	3	18 data stripe units on the two failed nodes in a global periodical stripe.
$Node_0$, $Node_2$	6	
$Node_0$, $Node_3$	6	
$Node_0$, $Node_4$	3	
$Node_1$, $Node_2$	3	
$Node_1$, $Node_3$	6	
$Node_1$, $Node_4$	6	
$Node_2$, $Node_3$	3	
$Node_2$, $Node_4$	6	
$Node_3$, $Node_4$	3	

The complexity of write requests for the generalized RAID5x is like RAID5 but increase the operations of mirrored copies, so the data-recovering performance in the failed node is higher than RAID5 in case of single-node fault because it only needs mirroring operations. Even if the double nodes fails, the data-recovering performance is still higher than RAID5 and at most 1/(N-2) of all the data stripe units in failed nodes for N×(N-1)/2 patterns are involved in XOR-based parity computation as is illustrated in Table.1, where PFN is for the pair of the failed nodes, NPC for the number of the parity to compute, and NNS for the number of the non-available stripe units. RAID6 and most of XOR-based code with double fault tolerance have the much more complicated parity computation than the RAID5.

References

[1] Thomasian A, “Clustered RAID Arrays and Their Access Costs”,The Computer Journal, 48(6), 2005, pp.702-713

[2] Surugucchi K, Hafner J L and Golding R A, “Reliability for Networked Storage Nodes”, In Proceedings of 2006 International Conference on Dependable Systems and Networks - Dependable Computing and Communications Symposium(DSN-DCCS 2006), Philadelphia: Sheraton Society Hill, 2006, pp.589-596

[3] J. S. Plank, Xu Lihao, “Optimizing Cauchy Reed-Solomon Codes for Fault-Tolerant Network Storage Applications”, Proceedings of the Fifth IEEE International Symposium on Network Computing and Applications NCA '06, IEEE Computer Society, 2006, pp.173-180

[4] J. L. Hafner, “WEAVER Codes: Highly Fault Tolerant Erasure Codes for Storage Systems”, Proceedings of the 4th USENIX Conference on File and Storage Technologies (FST'05), USENIX Association, 2005, pp.211-224

[5] J. L.Hafner, “HoVer Erasure Codes For Disk Arrays”, Proceedings of the International Conference on Dependable Systems and Networks (DSN'06) - Volume 00 DSN '06, IEEE Computer Society, 2006, pp.217-226

[6] A.Thomasian, “Multi-level RAID for very large disk arrays”, ACM SIGMETRICS Performance Evaluation Review, 33(4), 2006, pp.17-22

[7] Ming Hu, Minghua Jiang, “RAID5x: A Performance-optimizing Scheme against Double Disk Failures”, Proceedings of 2006 International Symposium on Distributed Computing and Applications for Business, Engineering, and Science (DCABES 2006), Shanghai University Press, Volume II, 2006, pp.1060-1063

[8] R.Anderson, Security Engineer: A Guide to Building Dependable Distributed Systems, John Wiley & Sons, 2001

[9] Naor D, Factor M, Nagle D, Riedel E and Satran J. The OSD Security Protocol. Proceedings of the Third IEEE International Security in Storage Workshop (SISW'05). IEEE Computer Society, 2005, Volume 00**:** 29 - 39

[10] Ming Hu, Minghua Jiang, “Simulation of Attacks on Network-based Error Detection”, Workshop on Intelligent Information Technology Application’2007(IITA’2007), IEEE Computer Society, Volume II, 2007, pp.1060-1063

A Framework of Logistics Outsourcing Partnership Governance in a Closed-Loop Supply Chain

Ning Zhang Baowen Sun

School of Information, Central University of Finance and Economics 39 South College Road, Haidian District, Beijing, 100081, China

Email: zhangning75@gmail.com

Abstract

From implementing a closed-loop supply chain, companies can generate more value and profits. However, closed-loop supply chains add complexity to overall supply chain management. It is a feasible way for companies to outsource all or part of logistics to third-party providers. Creating a partnership with a provider is essential to the success of outsourcing. This paper focuses on governance of this kind of partnership to achieve a "win-win" goal for both partners. Based on the scope dimension analysis, the partnership is classified into three different types, each of which is analyzed based on different theoretical perspectives. Partnership governance involves the service recipient, the service provider, and the partnership itself. Governance factors include outsourcing strategy of the recipient, the market position of the provider, and profit, responsibility, trust of the partnership. A framework of logistics outsourcing partnership governance is finally presented in the whole process of a closed-loop supply chain.

Keywords：Closed-loop Supply Chain; Reverse Logistics, Forward Logistics; Outsourcing; Partnership Governance

1 Introduction

Increasing attention is being given nowadays to reverse logistics, which can be integrated with forward logistics to form closed-loop supply chain. Reverse logistics is the process of moving goods from the point of consumption to the point of origin for the purpose of recapturing value or proper disposal [1]. Closed-loop supply chain has become increasingly important as a profitable and sustainable business strategy. In the new economy, consumers are empowered. They can easily compare products and buy them from competitors. As a result, companies have to be proactive in order to satisfy the consumers. From implementing a closed-loop supply chain, companies can generate more value and profits by maintaining customer support and improving customer satisfaction. On the other hand, with the increasing consumer awareness about environment protection, closed-loop supply chain can also create a green image for a company, which can also improve both the sales and the value of the company.

However, closed-loop supply chains add complexity to overall supply chain management: key issues are product design for recovery, reengineering, product data management, installed base support, and evaluating (end-of) life scenarios, etc.[2]. Implementing a closed-loop supply chain also implies a significant initial investment. Therefore, although many companies realize great chance of implementing closed-loop supply chain, there is great challenge for them to maintain both forward logistics and reverse logistics functions since they are often not core functions and continue to distract companies' activities from a main focus. Allowing a third party who specializes in logistics to maintain the function solves the difficulty. Hence companies are increasingly outsourcing all or part of their logistics efforts to third-party providers. Creating a partnership with a provider is essential to the success of logistics outsourcing. Although a mature and seamless partnership would most likely enhance the benefits of outsourcing, failure in the partnership can lead to negative and potentially irreparable consequences. This paper focuses on logistics outsourcing partnership governance to achieve a "win-win" goal for both partners.

Section 2 presents literature review of logistics outsourcing partnership. Based on the scope dimension analysis, the logistics outsourcing partnership is classified into three different types in Section 3, each of which is analyzed based on different theoretical perspectives. Section 4 details partnership governance and its factors. Integrated with all above, a framework of logistics outsourcing partnership governance is finally presented in a closed-loop supply chain in Section 5.

2 Literature Review

Partnership is a relationship based on agreed cooperation and coordination. Lambert, Emmelnainz, and Gardner [3] defined a partnership as "a tailored business relationship based upon mutual trust, openness, shared risk, and shared rewards that yield a competitive advantage, resulting in business performance greater than would be achieved by the firms individually. "

Literature of logistic outsourcing relationship is mostly focused on the selection of a partner from the perspective of client. The selection criteria include compatibility with the users, cost of service, quality of service, reputation of the company, long-term relationships, performance measurement, willingness to use logistics manpower, quality of management, information sharing and mutual trust, operational performance, information technology capability, size and quality of fixed assets, experience in similar products, employee satisfaction level, financial performance, geographical spread and range of services provided, risk management, etc. [4-7]. Andersson and Norrman [8] have suggested an eight-point plan for the selection and implementation of logistics outsourcing services. These points include (i) define or specify the service, (ii) understand the volume bought, (iii) simplify and standardize, (iv) market survey, (v) request for information, (vi) request for proposal, (vii) negotiations, and (viii) contracting. Jharkharia and Shankar [9] introduce a comprehensive decision methodology for the selection of a provider which allows for evaluation of alternative providers in two steps: (i) initial screening of the providers, and (ii) ANP-based final selection. Kumar and Malegeant [10] point out the benefits of strategic alliances between manufacturers and eco-non-profit organizations for the closed-loop supply chain.

3 Types Of Logistics Outsourcing Partnership

The nature of logistics outsourcing partnership determines the third-party providers' roles, just cost reducers or value enhancers, opportunistic agents or helpful stewards, suppliers of services or active collaborators [11]. Here we classify the logistics outsourcing partnership into three types based on the scope dimension, as shown in Figure1.

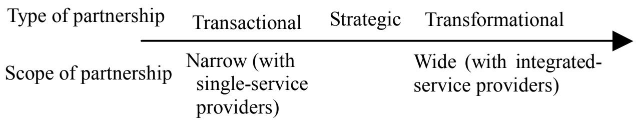

Type of partnership
Transactional
Strategic
Transformational
Scope of partnership
Narrow (with single-service providers)
Wide (with integrated-service providers)

Figure1 Types and scope of logistics outsourcing partnership

The scope of logistics outsourcing partnership depends on whether service recipients outsource different logistics functions to different providers or to one or a couple of providers. Single-service providers usually have high level of professional expertise and experience and can offer high level of service. But outsourcing different logistics functions to different single-service providers is a challenging work to a service recipient and will lead to the increase of the total outsourcing costs because the service recipient has to keep good relationship with each provider. Multi-service or even integrated-service providers can share more processes, information and knowledge with service recipients, and then have more understandings of the overall processes and strategies of the recipients. But keeping relationship with only one or a couple of providers will bring more risks to the recipient.

3.1 Transactional Partnership

In transactional partnership, service recipients aim at generating cost savings, preventing future investments or reducing staffing burden. Transactional partnership does not foster any kind of strategic relationship with the providers as they are generally short term in nature and essentially task based.

The essence of this type of partnership being transaction aimed at cost reduction, transaction cost economics (TCE) [12] seems to be the dominant theoretical perspective to conceptualize its nature. According to this perspective, recipients engage in transactional partnership based on the realization that the transaction costs associated with partnering is relatively lower than internalizing certain activities into their own hierarchical structures. Providers may be opportunistic in their behavior meaning they might resort to cheating, distortion of information, shirking of responsibility or other forms of dishonest behavior. Therefore, so long the recipients' needs for cost reduction and higher work quality are satisfied there will be no propensity to engage into partnerships that are deeper relationship-oriented.

3.2 Strategic Partnership

Strategic partnership is primarily driven by the growing need to concentrate more on business core in order to develop sources of current and future competencies. Value creation in strategic partnership occurs through building long-term relationships with a few best-in-class multi-service providers. Cumulative experience and learning scope are the two attributes of providers that recipients chiefly rely on.

Cumulative experience and learning scope constitute useful organizational resources and can be potential sources of competitive advantage. Therefore, the perspective of resource-based theory [13] may be used to conceptualize strategic partnership. According to this theory, a firm may be viewed as a collection of imperfectly imitable resources and capabilities that forms the basis of its successful competition against other firms. Through strategic partnership, recipients benefit from the useful resources of their providers without having to invest in possessing them. Tapping into the wide experience of providers enables their clients to fill in the resource-voids of their businesses. Further, organizational learning that results from this partnership renders the recipients to new ways of doing business by focusing attention and resources more narrowly on the business functions they do best.

3.3 Transformational Partnership

Transformational partnership implies a rapid, step-change improvement in enterprise-level performance of recipients. The motivation is to use outsourcing for the purpose of redefining existing businesses. The partnership nature may be viewed as powerful force for change for the recipients, and the providers may be considered as allies in the battle for market share and competitive advantage.

Allies possess valuable resources and capabilities. Therefore a useful perspective to conceptualize this form of partnership is resource-dependence theory [14]. According to resource-dependence theory, firms are actively involved in their broader environments and are dependent on other organizations for supply of critical resources. Recipients are dependent on their providers to supply critical resources which enable them to benefit from their providers' high-end skill base, cumulative domain expertise, and industry-specific knowledge that result in integrated, innovative solutions. Such resources help them to initiate rapid improvising changes or bring turnarounds of failing businesses. These transformations result in radical business models meant for achieving competitive edge through growth without investing for capacity enhancement.

4 Factors of Logistics Outsourcing Partnership Governance

Corporate governance may be defined as the organizational expression of the company's business objectives, a structure that therefore includes the means of attaining those objectives as well as guidelines for performance monitoring. Partnership governance has to result in realizing the mutually set goals of the partnership. It is complex since there is no common hierarchy (the companies are legally and economically independent of each other) and their respective goals may not be aligned (for example, the cost-saving goal of the service recipient versus the return-on-investment goal of the service provider).

Partnership governance involves the service recipient, the service provider, and the partnership itself.

4.1 Logistics Outsourcing Strategy of the Recipient

The logistics outsourcing strategy, which is derived from the company's general business strategy and must be aligned with it, details which logistics functions may be outsourced and which must be taken care of by the recipient itself, how many providers should be contracted, whether providers should be allowed to subcontract some of the services, and some other matters. The strategy must be shared with the company's providers so as to achieve complementary and shared goals. Providers can direct their efforts towards the realization of this strategy for both their offer and their delivery of outsourced logistics functions.

4.2 Market Position of the Provider

For providers it is important to know the market, what their vision for the future is and which logistics services to have in their portfolio. They themselves must also know the sectors and segments of which their market is composed. This includes the geographical scope within which they can deliver their services. Providers must maintain good relationships with consulting firms and keep them informed of their strategy and the services they can deliver, because these consulting firms often help recipients make selection list, only providers on which are sent a request for information or a proposal.

4.3 Profit in the Partnership

Outsourcing is designed to deliver financial benefits to the recipient. It must be kept in mind, however, that the provider is also a business and must maintain a profitable operation to survive and excel. The partnership cannot be driven by cost reduction above all other considerations. In order for the provider to continue to be motivated to provide high-quality services, there must be profit in the partnership. The profit and reward that goes along with outstanding work motivates the provider to commit resources, ensure quality and service levels, identify new opportunities, address the recipient's business issues in a timely and proactive manner, and innovate.

4.4 Responsibility in the Partnership

In all cases, the responsibilities of the providers need to be clear at all times. This is even more important if more than one provider is involved in delivering the services that their client needs. Then these providers have responsibilities to one another as well as to the recipient. Clear responsibilities prevent providers from blaming one another or their client should anything go wrong. The most important matter is to have all parties work together smoothly. On the other hand, the responsibilities of the recipients that are often ignored or minimized are also very important. Sufficient internal management resources required to effectively manage a provider partnership should be devoted.

4.5 Trust in the Partnership

Mutual trust between the recipient and their providers is important – not only during the selection process but also during the contract period, when the services agreed are delivered. A trusting relationship may lead to inter-organizational transactions and to new, unexpected revenue opportunities that may not be included in the scope of the original contract. Such trust has to be generated and maintained on a group level as well as individuals. To do so, people should communicate openly. Should problems arise, they are then immediately discussed with a focus on finding solutions.

5 A Framework of Logistics Outsourcing Partnership Governance

Figure2 illustrates the framework of logistics outsourcing partnership governance in a closed-loop supply chain.

A closed-loop supply chain consists of forward logistics and reverse logistics. The forward logistics represents the normal flow of material from raw material to finished goods to the ultimate consumer. Third-party providers can perform this transportation process. The reverse logistics is more complicated as the product being dispositioned can be handled in many different ways. The processes include collection, inspection, reprocessing or disposal, and redistribution, which can also be performed by third-party providers.

(1) Collection: all activities rendering used items (product, component, or material) available and physically moving them to some point for further treatment.

(2) Inspection: results in splitting the flow for various recovery and disposal options.

(3) Reprocessing: reusable flows undergo the actual transformation of a used item into a reusable item of some kind. The product recovery operations aimed at recapturing value include repair, reuse, remanufacturing, refurbishing, and recycling.

(4) Disposal: the non-reusable flows are disposed of to incinerators and landfills.

(5) Redistribution: directing reusable items to be marketed to new markets, and physically moving them to potential new users.

Acknowledgment

This research is partially supported by the National Social Science Foundation of China under Grant 07CJY044.

References

[1] Rogers, D. S., and Tibben-Lembke R. S., "Going backwards: reverse logistics trends and practices, " The University of Nevada, Reno, Center for Logistics Management, Reverse Logistics Council, 1998

[2] Van Wassenhove, L., and Geyer, R., "The impact of constraints in closed-loop supply chains: the case of reusing components in durable goods," Proceedings of the 10th LCA Case Studies Symposium on Recycling, Closed-Loop Economy and Secondary Resources, Barcelona, Spain, December 2-3, 2002

[3] Lambert, D. M., Emmelhainz, M. A., and Gardner J. T., "Building Successful Logistics Partnerships," Journal of Business Logistics, Vol.20, No.1, 1999, 165-181

[4] Boyson, S., Corsi, T., Dresner, M., and Rabinovich, E., "Managing third party logistics relationships: what does it take?" Journal of Business Logistics, 20(1): 73-100, 1999

[5] Langley, C. J., Allen, G.. R., Tyndall, G. R., "Third-party logistics study 2002: results and findings of the seventh annual study,". Council of Logistics Management, Illinois, USA, 2002

[6] Lynch, C. F., "Logistics outsourcing: a management guide," Council of Logistics Management Publications, Illinois, USA, 2000

[7] Razzaque M. A., Sheng C. C., "Outsourcing of logistics functions: a literature survey," International Journal of Physical Distribution and Logistics Management, 28(2), 1998, 89-107

[8] Andersson, D., and Norrman, A., "Procurement of logistics services-a minutes work or a multi-year project?" European Journal of Purchasing and Supply Management, 8, 2002, 3-14

[9] Jharkharia, S., and Shankar, R., "Selection of logistics service provider: An analytic network process (ANP) approach," Omega, 35, 2007, 274-289

[10] Kumar, S., and Malegeant, P., "Strategic alliance in a closed-loop supply chain, a case of manufacturer and eco-non-profit organization," Technovation, 26, 2006, 1127-1135

[11] Kedia, B. L., and Lahiri, S., "International outsourcing of services: A partnership model," Journal of International Management, 13, 2007, 22-37

[12] Williamson, O. E., "The economics of organization: the transaction cost approach," American Journal of Sociology, 87, 1981, 548-577

[13] Barney, J. B., "The resource-based theory of the firm," Organization Science, 7 (5), 1996, 469

[14] Pfeffer, J., and Salancik, J. R., "The External Control of Organizations: A Resource Dependence Perspective," Harper and Row, New York, 1978

Decision Method of E-Business Project Based on Rough Sets

Lihua Ma[1] Huizhe Yan[2] Wanqing Li[3]

School of Economics and Management, Hebei University of Engineering, Handan, Hebei, 056038, P.R.China

Email:1 malihua2004@126.com; 2 yanhuizhe@163.com; 3 liwanqing@263.net

Abstract

In order to solve the problem of E-Business project decision, a new decision method is proposed by analyzing project including uncertain factors. Firstly, the set of factor is established including condition attribute and decision attribute. Secondly, experts qualitatively describe risk factors and establish a decision database, called decision table. Thirdly, the attribute reduction algorithm based on Rough Sets is used to eliminate the redundant risk factor and its value of decision table. Finally, the minimum decision rules are abstracted based on data mining technology. The method is more convenient and practical compared with the traditional one.

Keywords：Rough Sets; data mining; attribute reduction; E-Business; decision method

1 Introduction

With the development of computer and communication technology, more and more enterprises use E-Business to conduct their business activities through Internet. So E-Business project is popular to many enterprises. But it is well known that E-Business is a high-risk project because of many uncertain causes including complex technology, specialized equipment, special environment personnel disposition and so on [1,2,3]. So the risk decision of E-Business is a very important task. Then in order to solve the problem of risk decision of E-Business project, a new decision method is proposed. The rest of the paper is organized as follows. In section 2, the theory and model of Rough Sets algorithm are introduced. In section 3, first, the set of factor is established, including condition attribute and decision attribute. Secondly, experts qualitatively describe risk factors and establish a decision knowledge database, called decision-making table. Thirdly, the algorithm based on Rough Sets is used to eliminate the redundant risk factor and its value from the decision table. Finally, the minimum decision rules are created based on data mining technology. And merits of this method are more than traditional method in many aspects, like convenience, objectivity and feasibility.

2 Preparation of manuscripts

Rough Sets theory proposed by Z.Plawlak in 1980s is a novel mathematic method to study uncertain data, deficiency of data, incomplete data, or even inconsistent data [4,5,6].

2.1 Indiscernibility relation

In Rough Sets, the relation is close between knowledge and classification. And knowledge is defined as an ability to classify. Suppose $K=(U,R)$ is a knowledge base, where U is a nonempty finite set called domain, R is the equivalence relations of U , U/R is all the equivalence classes$R$. $[X]_R$ is an equivalence class of R including element $x \in U$. If $P \subseteq R$ and $P \neq \Phi$, then all intersection of equivalence relations are an equivalence relation in P , called indiscernibility relation about P, as in $ind(P), [x]_{ind\{R\}} = \bigcap_{R \in P} [x]_R$, $P \subseteq R$.

2.2 Upper approximation, lower approximation and boundary of Rough Sets

In Rough Sets, accuracy concepts are signified by two accuracy sets including upper approximation and lower approximation. In a knowledge base $K=(U,R)$, for each subset and an equivalence relation $R \in ind(K)$, suppose two subsets are as follows.

$$R(K)=\{x|[x]_R \subset X, x \in U|\}$$

$$R^{-}(X)=\{x|[x]_R \subset X \neq \phi, x \in U\}$$

Then $R_{-}(X)$ and $R^{-}(X)$ are upper and lower approximation sets of X about R. Suppose boundary domain of X about R as. $bn_R(X)=R^{-}(X)-R_{-}(X)$ And suppose $pos(X)=R\_(X)$ as positive region of X about R, $negR(X)=U-R\_(X)$ as negative region of X about$R$.

2.3 Information system and decision table

In Rough Sets, information system is takes the form of relation table. Knowledge system with condition attribute and decision attribute is a decision table. A decision table is a kind of critical knowledge system. Suppose $S=(U,A,V,f)$ is a knowledge system, where $S=(x_1,x_2,\cdots,x_n)$ is a finite set of object, $A=\{a_1,a_2,\cdots,a_n\}$ $A=\{a_1,a_2,\cdots,a_n\}$ is a finite set of attribute, here in V is field composed of attribute A, f: $U \times A \rightarrow V$ is a information function, each element of U with a unique value that is a about V, $A=C \cup D$, C is the condition attribute set, D is the decision attribute set.

2.4 Reduction algorithm based on Rough Sets

Simplified decision table reduction is the result of simplifying condition attribute reduction, and the classification function remains to be. And simplified decision table contains less complicated condition attributes. We know a simplified condition is necessary in making decisions [7, 8, 9,10,11].

1) Attribute reduction

For an information system $S=(U,A,V,f)$, $A=C \cup D$, $B \subseteq C$,if $\gamma_C(D)=\gamma_B(D)$ and B is individual in relation to D, then B is the simplification of attribute D in relation to C, as in $RED_D(C)$. The calculation is shown as follows.

```
Input: C,D,U.
Output: simplification of attribute C in relation to D
Step 1  $s \leftarrow 0, RED(s) \leftarrow \phi$;
Step 2  $i \leftarrow 1$;
Step 3  $j \leftarrow 1, m \leftarrow 0$;
Step 4 For subset $C(i,j)$ of $C$ covering, $j$ subset of element $i$
    (1) $t \leftarrow 0$
    (2) If $(RED(t) \neq \phi) \wedge (RED(t) \subseteq C(i,j)), m \leftarrow m+1$, if $m=C^i_{|C|}$, turn to Step 7,else turn to Step 5
    (3) If $t \geq s$ turn to (5)
    (4) $t \leftarrow t+1$, turn to (2)
    (5) If $\gamma_C(D)=\gamma_{C(i,j)}(D)$ turn to (6), else turn to Step 5
    (6) $s \leftarrow s+1, RED(s)=C(i,j)$
Step 5 If $j \geq C^i_{|C|}$ turn to Step 6, else $j \leftarrow j+1$, turn to Step 4
Step 6 If $i \geq |C|$ ends, else $i \leftarrow i+1$, turn to Step 3
Step 7 Outputs $RED(s)$
```

2)Attribute value reduction

For an information system $S=(U,RED_D(C) \cup D,V,f)$, the calculation is shown as follows.

```
Input: $S=(U,RED_D(C) \cup D,V,f)$ , $RED(C)=\{C_1,C_2,\cdots,C_n\}$
Output: core value table $S'$ of $S$
Step 1  $S'=(U,C \cup D,V' \leftarrow Null,f')$
Step 2  For each condition attribute $C_k$ (repeat as follows)
    For each $x_i \in U$ and $C'_k(x_i)=Null$ (repeat as follows)
        If $\exists x_i((x_j \neq x_i) \wedge \forall C_l(C_l \neq C_k \wedge C_l(x_j)=C_l(x_i))) \wedge (D(x_j) \neq D(x_i))$
        Then $C'_k(x_j)=C_k(x_j)$, $C'_k(x_i)=C_k(x_j)$
Step 3 Output $S'$
```

3 Case Study

3.1 Condition attribute sets and decision attribute sets

The risk factor sets are called condition attribute sets, which reflect E-Business project risks, including technical feasibility *a*, amount of investment *b*, capital-raising ability of project *c*, and market expectation *d*. Decision attribute sets include enterprise scale *e* and risk process methods *f*. We can get information to make decision from a decision table, called risk decision-making table. The table is composed of some rows and arrays to represent the attributes and the objects. We study an E-Business project as an example to abstract Table 1 as follows.

Table 1 Risk decision table of E-Business project

U A	a	b	c	d	e	f
2	2	2	1	2	2	3
3	2	2	1	1	2	3
4	1	1	1	1	1	3
5	1	2	1	2	1	3
6	2	2	1	3	2	2
7	2	3	1	3	2	2
8	3	3	1	3	2	2
9	3	3	2	3	2	1
10	3	4	2	3	2	1
11	3	4	2	2	2	1
12	3	3	2	2	2	1
13	3	2	2	2	2	1

3.2 Dispersing attribute and establishing expert knowledge base

Dispersing condition and decision attributes are used to establish the knowledge base. Firstly, using condition attribute sets from above; we disperse the results of expert evaluation as follows. Technical feasibility is divided into 3 grades $\{1,2,3\}$ to represent $\{low, average, high\}$. Similarly, the amount of investment is also divided into 4 grades $\{1,2,3,4\}$ to represent $\{lower, low, high, higher\}$. Capital-raising ability of project is divided into 2 grades $\{1,2\}$ to represent $\{bad, good\}$. Market expectation is divided into 3 grades $\{1,2,3\}$ to represent $\{bad, average, good\}$. Secondly, we use decision attribute sets above, and the results of expert evaluation are represented as follows. Enterprise scale is divided into 2 grades $\{1,2\}$ to represent $\{small, big\}$. Risk process methods are divided into 3 kinds $\{1,2,3\}$, including risk bearing, risk sharing and risk avoiding.

3.3 Attribute reduction

From Table 1, the redundant attributes are eliminated and core attributes are preserved. The minimum decision rules are composed of core attributes without redundant attributes, the new table is called simplified attribute table as follows Table 2.

Table 2 Simplified attribute table

U A	a	c	d	e	f
1	2	1	2	2	3
2	2	1	1	2	3
3	1	1	1	1	3
4	1	1	2	1	3
5	2	1	3	2	2
6	3	1	3	2	2
7	3	2	3	2	1
8	3	2	2	2	1

3.4 Attribute value reduction

From Table 2, we get simplified attribute value table as follows.

3.5 Interpretation analysis

From Table 3, we know decision rules as follows.

$a_2d_2 \to (2,3)$ or $a_2d_1 \to (2,3)$, $a_1 \to (1,3)$, C $d_3 \to (2,2)$ or $c_1 \to (2,2)$, $c_2 \to (2,1)$.

From above, we know 4 decision rules as follows.

(1) $a_2d_2 \vee a_2d_1 \to (2,3)$.

(2) $a_1 \to (1,3)$.

(3) $d_3 \vee c_1 \to (2,2)$.

(4) $c_2 \to (2,1)$.

From 4decision rules above, we get strategy of risk decision as follows.

(1) When the technical feasibility is “average”, and market expectation is “average” or “bad”, the strategy of big enterprise is risk avoiding. That is to say, that will give up project or modify it.

(2) When the technical feasibility is “low”, the strategy of small enterprise is risk avoiding. That is to say, that will give up the project or modify it.

(3) When market expectation is “good” or capital raising ability of project is “bad”, the strategy of big enterprise is risk sharing. That is to say, that will share risks and profits with cooperators.

Table 3 Simplified attribute value table

U A	a	c	d	e	f
1	2	-	2	2	3
2	2	-	1	2	3
3	1	-	-	1	3
4	1	-	-	1	3
5	-	-	3	2	2
6	-	1	-	2	2
7	-	2	-	2	1
8	-	-	-	2	1

(4) When capital-raising ability of project is “good”, the strategy of big enterprise is risk bearing. That is to say, that will solely bear the risks in full.

4 Conclusions

The risk decision of E-Business project is sort of multiple attribute decision-making processes. It is certain that we have to deal with massive data. And data mining technology is a method of auxiliary decision, which can abstract implicit regularity from massive data. In this paper, the reduction algorithm based on Rough Sets is discussed as a practical data mining technology. A new decision method is proposed in order to solve the risk decision problem of an E-Business project. The factor set is established including condition attribute and decision attribute. Then experts qualitatively describe risk factors and create a decision table, and the attribute reduction algorithm based on Rough Sets is used to eliminate the redundant risk factor and its value of decision table. Finally, the minimum decision rules are created based on data mining results. We can make proper decision from the rules are to improve precision and explanatory ability in practice.

References

[1] Stephen Mason. Approaching Contract Risk in E-Commerce [J]. Risk Management Buulletin, 2000, (5): 8-12

[2] Bennet P.Lientz,Kathryn P.Rea,Shen T. E-Commerce Project Performance Management[M].Beijing: Electronic Industry Press,2003(in Chinese)

[3] Guan Shuming,Lv Xuhua,Guo Hong. An evolutionary game analysis on trust problem in B-C electronic commerce [J]. Journal of Hebei University of Engineering (Natural Science Edition),2007,24(1):99-102(in Chinese)

[4] Z. Pawlak. Rough Sets-Theoretical Aspects of Reasoning about Data [M]. Klystron Academic Publisher, 1994

[5] Liu Q. Rough Sets and Rough Inference [M]. Beijing: Science Press, 2001 (in Chinese)

[6] Zhu H. Research of Minimum Decision Algorithm Based on Rough Set [J]. Computer Application, 2002,22(9): 19-21 (in Chinese)

[7] Lv A M, Lin Z J, Li CH M. Approaching of Data Mining and Knowledge Discovery [J]. Survey Science, 2000,12(4): 36-39 (in Chinese)

[8] Zeng Huanglin. Rough Sets and Application [M].Chongqing University Press,1996 (in Chinese)

[9] Yang Shanlin. Intelligent Decision Method and Intelligent Decision Support System [M]. Science Press, 2005 (in Chinese)

[10] Wu Bing, Li Weixiang, Zhao Lindu. Least Decision-Making Method Based on Rough Set Theory [J]. Information Technique, 2002, (4): 4-5 (in Chinese)

[11] Zheng Xiuli, Wang Lening,Chen Zhongzhu. Data Mining-based Project for Electronic Business Client Potential Exploitation [J]. Computer Engineering and Application, 2002, (5): 194-195 (in Chinese)

Application and Research of a New Multi-party Digital Signature Scheme in E-Commerce and E-Government

Yang Xia[1] Hongwei Zhang[1] Ling Zhang[1] Zhao Xu[2]

1 School of computer science, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China
xia-y@163.com

2 School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou, Jiangsu 221116, China

Abstract

Currently e-commerce and e-government have a rapid development. But they are restricted by various choke points. One of the most emergent is the security of secret information. Given the demands of multi-party transmission in e-commerce and e-government, this paper, based on present signature technology, puts forward the concept of united-signature and designs a new multi-party digital signature scheme suitable for both e-commerce and e-government. This can ensure the validation, confidentiality, integrality, and non-denial of the secret information in multi-party transmission. By applying the scheme to the practice, the paper used some instance to illustrate its feasibility and sum up its characteristic, therefore proved its operation significance and use value.

Keywords：e-commerce; e-government; digital signature; united-signature; secret information

1 Introduction

With the rapid development of e-commerce and e-government, more and more people surf and shop on line. Therefore, it becomes more and more common to attract investment, apply for enterprise enrollment, report personal tax returns, and to do other administrative activities. In China, the government portals become an important channel for public information and more than 50% of the administrative licensing items can be done on line by 2010. However, they are restricted by various choke points. One of the most emergent things is the security of the secret information in multi-party transmission. For example, in the online payments, security is a priority in the multi-party communications among customers, businesses, enterprises, banks and the issuer of electronic currency and so on. In e-government construction, there are government agencies and the public involved in the transmission and transaction of many documents, applications, tables. So there is an urgent need for a new multi-party digital signature scheme to ensure the validation, confidentiality, integrality, and non-denial of the secret information in multi-party transmission.

There is a variety of digital signatures, such as multi-signature, blind signature, proxy signature, threshold signatures, whose application are closely related to the specific environment. And on multi-party signature, a lot of new researches have been done. For example, there appear proxy multi-signature, multi-proxy signature and multi-proxy multi-signature. They allow many-to-one, one-to-many, many-to-many relationship between the original signature and the agents[1].ElGamal type digital multi-signature can detect and prevent the fraudulent practices of signature [2]. A forward secure multi-signature makes an adversary unable to forge multi-signature pertaining to the past even if he or she has gotten the current all signers' keys and the previous generated multi-signature remains valid [3]. Multi-signature based on discrete logarithm can efficiently avoid the attackers who deny that they had taken part in process of signing some messages with others [4]. Multi-signature and group signature based on Bilinear Pairing makes bilinear mapping a new tool in building a password system to construction of a new structure based on the multi-signature model and visit signature model, etc[5]. Based on the combination of e-commerce and e-government applications, we present a more adaptable solution of multi-party digital signature and give an example of e-government to describe our design as follows.

2 Multi-Party Signature Requirements

E-commerce and e-government not only need the individual signature, but also, on occasion, need the multi-party signature that allows a number of signatories sign and authenticate the same message. Meanwhile, the signatories can only sign the messages which belong to themselves according to their hierarchy. In addition, the authority of each department or the staff should be different. For example, the same document concerning a given matter, may be visible to department A but not to department B. Therefore, we must take strict security measures to ensure the different confidentiality according to the hierarchy and authority of each staff in the documents transmission.

In e-government examination and approval, such situation often appears: A business of approval needs many signatories and they may sign without a certain order. But because of their different hierarchy and authority, each signatory is limited to a specific part, having no idea of other secret messages. Then, signatories transmit their signatures to the synthesizer to verify the signatures and complete the signature of the approval. It is as Fig. 1 shows:

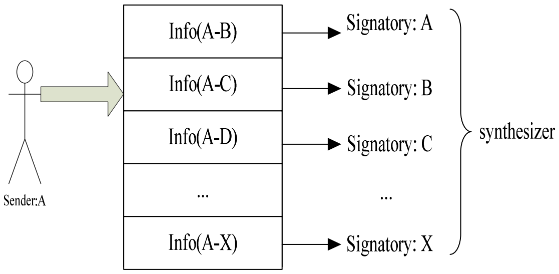


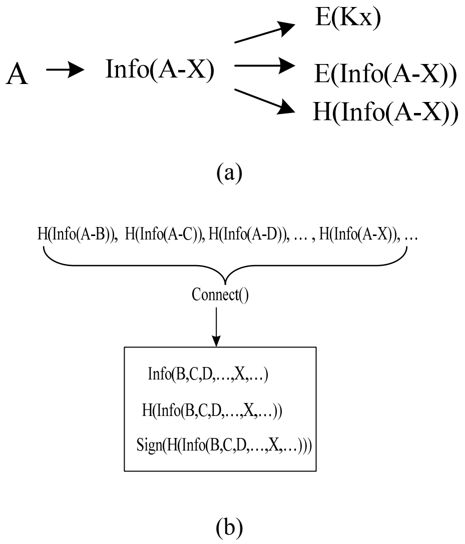
Figure 1 Scheme design

3 Design of Multi-party Digital Signature

We use letter A to stand for the first sender(such as the applicant of approval business) and use letters B, C, D, E,... to stand for the signatory respectively. Info(A-B) denotes the information A passes to B. Info(A-C) denotes the information A passes to C. Similarly , Info(A-X) denotes the information A passes to X , etc. The working principle of our scheme is as follows.

3.1 Generation of united-signature

A generates Info(A-B), Info(A-C) ,… Info(A-X) …. and A deals with them as Fig. 2 follows[6]:


A
Info(A-X)
E(Kx)
E(Info(A-X))
H(Info(A-X))
H(Info(A-B)), H(Info(A-C)), H(Info(A-D)), … , H(Info(A-X)), …
Connect()
Info(B,C,D,…,X,…)
H(Info(B,C,D,…,X,…))
Sign(H(Info(B,C,D,…,X,…)))

(a)

(b)

Figure 2 A’s secure disposal of secret information

1) A uses digital envelope technology to encrypt the secret information

A generates symmetric key Kb and uses Kb to encrypt Info(A-B), which results in cryptograph E(Info(A-B)) .Then, A uses B’s RSA public Key to encrypt Kb, which results in E(Kb);

According to the same operation, E(Info(A-C)), E(Kc), E(Info(A-D)), E(Kd), E(Info(A-E))，E(Ke)… E(Info(A-X)), and E(Kx)...are resulted .

2) A generates united-signature of all secret information to be sent

A uses SHA-1 Hash Algorithm to respectively generate digest H (Info（A-B）), H (Info（A-C）), H (Info（A-D）), H (Info（A-E）)… H (Info（A-X）)…; A unites all the digests of these secret information and uses SHA-1 Hash Algorithm to generate H(Info(BCDE…X…)). Then A uses its own RSA private Key to encrypt H(Info(BCDE…X…)) and thus generates the united-signature Sign[H(Info (BCDE…X…))];

3.2 Process of application

Each signatory gets the secret information directly sent by A or transmitted by other signatories. The secret information each obtained include (Let's take D as an example):

Cryptograph E(Info(A-D));

E(Kd);

Others' digests: H (Info(A-B))、H (Info(A-C))、H (Info(A-E))…H (Info(A-X))…;

United-signature: Sign[H(Info(BCDE…X…))];

D's process to decrypt and validate the secret information (other signatories undergo similar process):

Decrypt E(Kd) with its RSA private key and get symmetric key Kd;

Decrypt E(Info(A-D)) with Kd and get Info(A-D);

Encrypt the Info(A-D) with SHA-1 Hash Algorithm and get digest H(Info(A-D));

Connect H(Info(A-D)) with others' digests received like H (Info(A-B))、H (Info(A-C)) 、H (Info(A-E))…H (Info(A-X))…and get Info(BCDE…X…);

Encrypt Info(BCDE…X…) with SHA-1 Hash Algorithm and get digest H'(Info(BCDE…X…));

Use public key of A to decrypt Sign[H(Info (BCDE…X…))] and get digest H(Info(BCDE…X…));

Comparing H'(Info(BCDE…X…)) with H(Info (BCDE…X…)), if they are the same , it indicates that Info(A-D) was provided by A without any tampering.

Having decrypted cryptograph and validated united-signature, D encrypts digest H(Info(A-D)) with its RSA private key and Sign[H(Info(A-D))] is resulted. Then, D transmit its signature Sign[H(Info(A-D))] to the synthesizer . The process of each signatory goes on like this until the last one. Finally, the synthesizer verifies the signatures and identifies who has signed, then complete the signature of the approval.

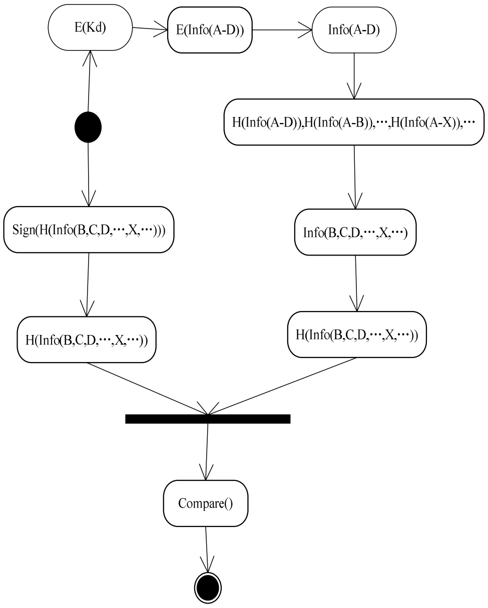


Figure3 The process of D's decrypting cryptograph and validating united-signature

There is another figure to show the process of D's decrypting cryptograph and validating united-signature. As follows:

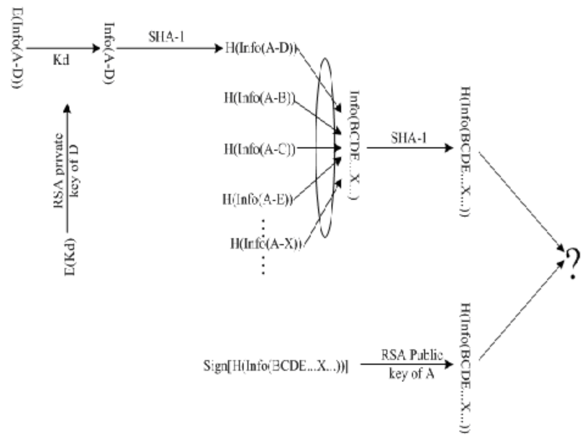


Figure 4 Another way to show the process of D's decrypting cryptograph and validating united-signature

Of course, replay attacks pose a great threat in the process of communication. Time-stamp technology strategies can avoid it. This paper does not propose to discuss the details of such problems here.

4 Advantages of the schema

4.1 United-signature , part-validation

Although each receiver can only get a part of secret information belonging to themselves and can not get the others', they can use united-signature to validate the authenticity, integrality, non-denial of the part of secret information sent to themselves.

4.2 Simultaneous signature, mislocation avoiding

The sender makes a united-signature of all parts of secret information. This can connect different parts of secret information sent to different receivers and the connection indicates that they belong to the same approval business and that they are signed contemporary. Each receiver can judge which business item the information they received belong to, thus avoiding dislocation of secret information involved with many business items and optimizing the signature speed at the same time.

4.3 Double encryption, efficient security

The sender uses digital envelop technology to encrypt every part of secret information. We use public key in outer layer and use symmetric key inside which is usually very short. This can avoid the difficulty of symmetric key distributing and the long time of Asymmetric Key Encryption. Therefore, we obtain the agility of Asymmetric Key Encryption and the high efficiency of Symmetric Key Encryption [7]. In addition, we use different symmetric key for each receiver. This makes each receiver unable to see the others' secret information and achieve high security.

4.4 Agility and simplicity, good adaptability

Compared with the whole complex arithmetic in the process of multi-party signing, our same signature scheme is simpler. Moreover, each signatory can sign respectively and the way of digital signature may change according to their preferences. Therefore, our scheme is more flexible in e-government business and in distributing electronic commerce[8].

5 Conclusion

We can discuss its security in many areas[9] particularly the transmission security of the secret information. Given the demands of multi-party transmission in e-commerce and e-government, this paper puts forward the concept of united-signature and designs a new multi-party digital signature scheme. In addition to the security of the secret information available to each signatory, the signatory cannot see others' secret information in the transmission. In this way, the more flexible, secure, and efficient e-government will be made true. This scheme is of great operation significance and use value.

References

[1] JI Jia-Hui, LI Da-Xing, WANG Ming-Qiang, New Proxy Multi-Signature, Multi-Proxy Signature and Multi-Proxy Multi-Signature Schemes from Bilinear Pairings[J], Chinese Journal of Computers, 2004 Vol.27 No.10 P.1429-1435

[2] LU Jian Zhu, CHEN Huo Yan, LIN Fei, Elgamal Type Digital Multisignature Schemes And Its Security [J], Journal Of Computer Research And Development, 2000 Vol.37 No.11 P.1335-1339

[3] WANG Xiao-Ming , FU Fang-Wei , ZHANG Zhen, A Forward Secure Multisignature Scheme [J], Chinese Journal of Computers, 2004 Vol.27 No.9 P.1177-1181

[4] LU Lang-Ru, ZENG Jun-Jie , KUANG You-Hua, NAN Xiang-Hao, A New Multisignature Scheme Based on Discrete Logarithm Problem and its Distributed Computation[J], Chinese Journal of Computers, 2002 Vol.25 No.12 P.1417-1420

[5] MA Chun-Bo, AO Jun, HE Da-Ke, Multi-Signature and Group Signature Based on Bilinear Pairing[J], Chinese Journal of Computers, 2005 Vol.28 No.9 P.1558-1563

[6] Zheng Xiao-Mei, The research and implementation of XML security technology[D], China university of Mining technology, 2004

[7] Gary P.Schneider, James T.Perry, Electronic Commerce [M], China Machine Press, 2002

[8] Xia Yang, Zhang Qiang, The Design and Implementation of Distributed E-Business System [J], Microelectronics & Computer, 2006 Vol.23 No.10 P.100-103

[9] XIA Yang , TANG Liang, ZHANG Qiang, The Safety and Resolving Method about E-Government[J], Computer Engineering and Design, 2005 Vol.26 No.1 P.110-113

[10] MO Le-qun, WANG Xiao-ming, YAO Guo-xiang.Security analysis of a digital multisignature scheme[J], Computer Applications, 2005 Vol.25 No.10 P.2294-2298

[11] Tzer-Shyong Chen, Digital Multi-Signature Scheme Based on the Elliptic Curve Cryptosystem[J], Computer Science and Technology, 2004

Biography

Yang Xia: male, born in 1962, Xuzhou, Jiangsu China, Ph.D. Candidate, Associate Professor, Master mentor, Study: E-Commerce and E-government, Internet Technology, Network and Database Technology

Hongwei Zhang , male, born in 1982, Xuzhou, Jiangsu China, graduate student. Study: Web Services Technology, Network and Database Technology, E-Commerce and E-govenment

Ling Zhang , female, born in 1982, Qingdao, Shandong China, graduate student. Study: E-business and E-government, Network and Database Technology.

Zhao Xu , Male, 1955.1, Shuyang, Jiangsu China, Professor, Study: Communication and Information System, Broad Band Access Network Technology, Fiber Optical Communication Technology.

The Effect of E-business on the International Trade

Qi Wei   Yazhuo Liu

School of Economic and Management, Lanzhou University of Technology, Lanzhou, 730050, P.R.China

Email：weiq@lut.cn

Abstract

In this paper, we highlight the pervasive influence of e-business on the international trade. E-business benefits range from employment, to productivity gains, consumer surplus, and improvement in product quality, etc. E-business increases the international trade because of the above benefits. In this paper , five aspects that e-business influences the international trade are studied, such as e-business reshaping the way companies go to the international market, e-business and export performance reinforcing each other and so on.

Keywords：E-business; International Trade; Export; WTO

1 Introduction

The evidence that the e-business has affected international trade can be found everywhere. Women in a remote part of Guyana are selling hand-woven hammocks to people around the world for as much as $1000 a piece by e-business (NYT 3/28/2000). An importer in the Dominican Republic found a Bolivian supplier of Soya oil and a Chinese supplier of sewing machines by e-business. A producer of draperies and other goods from Hicksville, Long Island is negotiating deals by e-business in Turkey, Saudi Arabia, and South Africa, after years of serving only the domestic market. Of greater importance for sheer trade volumes, global business-to-business web sites have already been set up in a number of industries. Some examples include SciQuest, a global marketplace for laboratory and scientific materials; Commerx, a global exchange for plastics, metals, and packaging materials; and e-Steel, which links buyers and sellers of steel products around the world. In one well-known example, Daimler-Chrysler, GM, and Ford founded COVISINT, an Internet-based market for car parts that aggregates thousands of suppliers worldwide. Forrester Research, a leading consulting company, showed that global e-business crossborder trade was $44 billion in 2000 and growed to $1.4 trillion in 2004, accounting for 18percent of total exports.

With the growth of the global e-business crossborder trade, the e-business is changing the rules of the international trade .The main purpose of this paper is to analyse the effect that e-business has had on international trade in recent years. We show that e-business reshapes the way companies go to the international market, and e-business and export performance reinforce each other, and e-business increases export, and e-business affects the legal determinants, especially WTO rules is changing with the growth of e-business.

2 E-business: a new way of doing business

E-business is the use of electronic means to exchange information and to carry out activities and transactions. From this definition, proposed by a number of national and international organizations and accepted by the vast majority of businesses (OECD, 1999), we can see that e-business covers many and varied areas of economic activity.

Singapore is the first country where the entire trade transaction process was based on information technology and EDI based information exchange.

TradeNet, a value-added network linking the trading community (traders, freight forwarders, and cargo and shipping agents) to more than twenty government agencies involved in the import/export process, was launched in Singapore in 1989. Instead of submitting documents to and obtaining permits separately from each government agency, a single electronic document is routed through the network and returned with the necessary approvals within 15-30 minutes. This compares to 2-3 days before the introduction of TradeNet. Today, more than 98% of all trade declarations in Singapore are processed through this system, allowing companies to move cargo at short notice and reducing costs by as much as 50%.Now electronic submission of trade documentation has become the rule in a number of other countries as well: in the United States, Canada, and some member states of the European Union, more than 90% of customs declarations are submitted electronically. In the future, the Internet is likely to facilitate electronic customs clearance further, as new packages are developed to facilitate information flows.

E-business has a broad range of financial and other applications, such as the dissemination and exchange of digital data, electronic funds transfers, electronic stock exchange activities, commercial auctions, co-operative design and engineering, electronic bidding, direct consumer sales, and after-sale services and so on. E-business has the potential of transforming existing trade networks and of reducing the handicap of geographical isolation. Firms around the world can now experiment with new ways to contest international markets and countries can further benefit from international specialization. Moreover, in the case of services and digitized goods, e-business allows bypassing conventional distribution channels, fostering international market integration. Not only the tangible goods but also intangible goods can be traded by e-business in the international market.

E-business has considerable advantages for the consumers and enterprises. These advantages range from employment, to productivity gains, consumer surplus, and improvement in product quality, etc. The consumer can enjoy a wider choice of products and services at lower prices, as well as certain convenience (no unnecessary trips, no restricted business hours). For enterprises, the adoption of e-business allows a reduction in co-ordination costs and leads to efficient electronic markets. The opportunities it offers are so great that it would appear there is no going back. A money figure can easily be placed on some of the competitive advantages it confers in Table 1: in particular, the cost reductions it allows and the gains in accuracy and speed it affords [1]. Other competitive advantages are equally substantial though it is harder to put a dollar value on them.

Table 1 Competitive advantages of e-business

Example of reduction of the costs of commercial transactions					
	Booking an Airplane ticket[a]	Banking transaction[a]	Bill paying[a]	Software distribution[a]	Stockbroking[b]
Traditional system	$8.00	$1.08	US$2.22-3.32	US$15.00	US$150 to 60
Via Internet	$1.00	$0.13	US$0.65-1.10	US$0.20-0.50	US$10
Cost reduction	87%	89%	71 to 67%	97-99%	93 to 83%
Example of Competitive advantages for manufacturing firms					
	Development costs[c]	Number of Engineering changes[c]	Rejects and adjustments[c]	Document distribution[c]	Reduction of new product design time[c]
Reduction relative to traditional system	25-35%	50-90%	75-95%	80%	40-60%

[a] OEDC(1999)

[b] Arthur Andersen(1999)

[c] NGM(1997,vol.2,p.27).

3 The Effect of E-business on the International Trade

3.1 E-business reshapes the way companies go to the international market

E-business has become an indispensable means of doing business. E-business is meant to reshape the way companies go to the international markets, the way customers buy products and services. E-business technologies are meant to help adopters to reach new customers more efficiently and effectively. E-business transforms the exchange of goods, services, information, and knowledge through the use of information and communication technologies (ICTs). There are several models of e-business, namely, (i) business to business (B2B), (ii) business to consumer (B2C), (iii) consumer to consumer (C2C), (iv) business to government (B2G), and government to business (G2B). Although there are several models of e-business, only two models, namely, B2B and B2C have experienced the highest growth. Within these two, B2B has grown from 3.5 billion US$ in 1995 to 34.0 billion US$ in 1998 while the growth of B2C has been 1.0 billion US$ in 1995 to 4.0 billion US$ in 1998. The share of B2B has increased from 77.78% in 1995 to 89.47% in 1998. These data suggest that B2B has grown faster than B2Cworldwide.

Broadly defined, there are three modes of e-business transactions. These are offline, online, and e-business using shared or individual portals. Offline e-business is normally done through e-mail systems while online e-business transactions take place with Uniform Resource Locators (URLs) of companies. The most effective way of doing e-business is through portals. Portals are the essential additions in network technologies. They fulfill an important role of aggregating contents, services, and information on the net. Broadly speaking, their position on the net is between users (buyers) and web contents. This unique position enables portals to leverage marketing and referrals as they are intermediaries between web users and companies. E-business has emerged as the fastest growing technology in recent times.

Because e-business is reshaping the way the company go to the international market, many developing countries is enforcing the use of e-business. In India, the government had taken initiative since late-1980s in the adoption of e-business by allowing companies to file their annual returns through floppies. Having taken cognizance of the UN General Assembly resolution and the recommendations of Indian industry associations, the Government of India drafted an IT bill to boost e-business in the country in 1999. Passed by the Parliament on 16 May 2000, the IT Bill 99 deals with the matters such as digital signature, electronic governance, electronic acknowledgement, acquisition and despatch of electronic records, regulation of certifying authorities, penalties and adjudication, the cyber regulations appellate tribunal, offences, and laws related to network service providers[2]. With the passage of the IT Bill 99 by the Parliament, e-business and e-governance are likely to grow as fast.

3.2 E-business and export performance reinforce each other

Interaction between the export performance and the adoption of e-business is depicted in Figure 1. It can be seen from Fig.1 that international orientation factors—which are represented by imports, exports, technological and financial collaboration and the adoption of e-business —mutually reinforce each other[3]. For instance, the adoption of e-business may be imperative to interact with foreign buyers for augmentation of exports and larger exports might provide the additional resources needed to use e-business more extensively. Similarly, sales turnover and the adoption of e-business are expected to influence each other. This holds true not only for adoption of e-business but also for any innovative activity carried out on the shop floor. One of the major prerequisites for the success of e-business is the existence of a very strong and reliable communication network. Access to a higher bandwidth is not within the control of individual firms. It forms part of the institutional infrastructure. Realizing the important role that availability of higher bandwidth plays in the success of web-enabled services, greater emphasis has been laid on the privatization and deregulation of telecommunication services in developing and developed countries. Conduct variables such as skill intensity of firms, investment on R&D, and wage rates are important factors that are expected to have bearings on the adoption of e-business. Higher remuneration is always a major incentive for the workforce to create and adopt innovations effectively and efficiently.

Firms might adopt e-business because of its perceived impact in reducing coordination costs. But as depicted in Figure 1, many more benefits exist. Due to the inexpensive access to global markets and information that the Internet enables, it is fast becoming the world's largest and most versatile marketplace for services, products, and information. Web-enabled services are likely to strengthen the competitiveness of firms as these technologies may change the relationship with customers by creating a stronger link between firms and its clients.

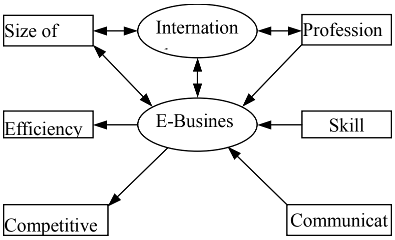


Figure 1 Export performance and e-business linkages

3.3 E-business increases export

Two of the most dramatic changes over the past fifteen years have been 'globalization', the massive rise in foreign capital flows and international trade, and the revolution in Information Technology (IT). In real terms (US$ 1995), world exports rose from $4.1 trillion (19% of gross world product) to $8.7 trillion dollars (24% of gross world product) between 1990 and 2003 (World Bank, 2006). Inward foreign direct investment rose from 1.0% of gross world product in 1990 to 4.9% of gross world product by 2000, before falling back to 1.7% in 2004. The revolution in Information Technology has been even more dramatic.

Despite the ubiquity of the e-business today, the first e-mail was sent over the predecessor to the internet only in 1972 and Netscape released the first commercially successful web browser, the Netscape Navigator, only in 1994 (Friedman, 1999). Even in the United States, where more than 185 million people had internet access by 2004, only 3 million had access in 1990 (International Telecommunication Union, 2005b). Because the e-business has lowered communication costs, observers have suggested that it is one of the primary reasons for globalization. For example, Friedman (1999, p. xviii) writes: "The new information technologies are able to weave the world together even tighter. These technologies mean that developing countries don't just have to trade their raw materials to the West and get finished products in return: they mean that developing countries can become big-time producers as well. These technologies also allow companies to locate different parts of their production, research and marketing in different countries, but still tie them together through computers and teleconferencing as though they were in one place[4]."

There is considerable anecdotal evidence that the internet has improved export opportunities in developing countries. For example, Friedman (1999) notes that America Online (AOL) employed 600 college-educated customer service representatives in the Philippines in 1999 to answer e-mails containing technical and billing inquiries from customers in the United States. Perhaps the most cited example of an industry in a developing country that has thrived because of e-business is the software industry in India, centered in Bangalore, which was estimated to have exports of about $6.2 billion in 2000/2001. Several papers using aggregate country-level data on trade flows have found that exports are higher in countries with greater e-business use. In a panel growth regression, Freund and Weinhold (2004) find that (lagged) growth of e-business is significantly correlated with the growth of trade between 1997 and 1999[5]. In cross-sectional levels regressions, they also find a statistically significant correlation between lagged e-business use and exports in 1998 and 1999, although the relationship is not statistically significant before 1998. In a second paper, Freund and Weinhold (2002) find that exports of services to the United States grew more quickly for countries with greater e-business penetration in a sample of 31 middle- and high-income countries. In a later study, Clarke and Wallsten (2006)find that aggregate exports from developing countries to developed, but not other developing countries is higher when internet penetration is higher. Clarke and Wallsten (2006) instrument for e-business penetration using variables representing government regulation of the telecommunications sector to control for the possibility of reverse causation[6].

3.4 International E-business influences legal rules

Most countries have by now introduced domestic legislation devoted to foster an adequate “digital” environment with special emphasis on the rules and regulations relevant for e-business. The quality of legal rules and enforcement, creditor rights, shareholder rights, and technology integration are positively related to e-business revenues[7]. The growth of Internet-based commerce is likely to depend on, among other things, the quality of legal rules and enforcement in ensuring that transaction data are protected and secured from electronic hackers worldwide. This concern, if not adequately addressed, is likely to increase transaction costs and reduce the incentive to transact on the web. Consistent with this prediction, Diamond and Verrecchia (1991) and Bartov and Bodnar (1996) show that high transaction costs reduce trading liquidity in US equity markets. In related studies, La Porta et al. (1997) and Dolven et al. (2000) find that countries with a poor tradition of law and order have smaller equity and debt markets relative to countries that have a high-quality law and order.

There can be little doubt that open markets are essential for the development of electronic commerce. A liberal regime encourages technical progress and the development of efficient practices. Market forces, however, may need to be complemented by industry self-regulation and/or government intervention to secure the following:(i) standards for the emerging global telecommunications infrastructure; (ii) adequate investment in the infrastructure; (iii) user-friendly and broad-based access; (iv) a predictable legal and regulatory environment which enforces contracts and property rights, (v) the security and privacy of data; (vi) rules for dealing with what constitutes unacceptable or conditionally acceptable content; (vii) a predictable framework of taxation and financial regulation; and (viii) equal opportunity through better education for users in industrialized and developing countries.

Realising the potential of ICTs in electronic business (e-business), the United Nations Commission on International Trade Law (UNCITL) adopted a Model Law in 1996. The UN General Assembly recommended to its members in January 1997 that they give due consideration to this Model Law, when they enact their laws related to e-business. Despite several advantages, the growth of e-business has been dismal particularly in developing countries. The major impediments in the adoption of such technologies have been the validity and authenticity of information. Lack of proper cyber laws could also be held responsible for the slow changeover to e-business. Given rapid changes in the area of electronic commerce, care needs to be taken to ensure that regulation does not fall behind or unnecessarily interfere with new developments.

3.5 E-business in the WTO

The pursuit of negotiations on e-business in the WTO is rationalized as a way to ensure ‘fair trade’ or an equality of competitive opportunities for foreign and domestic firms in this new area of economic activity[8]. Negotiations could, in principle, cover both “deep” integration issues (i.e., those dealing with differences in national regulatory regimes) and the “shallow integration” agenda (i.e., the elimination of discrimination at the border).

Starting in 1997, WTO members began to discuss if and how e-business should be dealt with in the WTO. The discussion of WTO-related issues is divided into seven parts. The first part examines questions relating to the infrastructure required for electronic commerce. It focuses on the outcome of recent negotiations aimed at liberalizing trade in information technology products and basic telecommunication services, and also considers the coverage of Internet access services in Members’ commitments under the GATS. The second part deals with market access issues regarding electronically transmitted products, including the implications of the recent United States and European Union agreement on duties affecting transactions on the Internet, the recent U.S. proposal to the WTO General Council on the customs treatment of electronic transmissions, and questions regarding the categorization of electronic transactions in the WTO framework. The third part looks at what has been achieved by way of trade liberalization commitments under the General Agreement on Trade in Services (GATS) in areas seemingly of most significance for electronic commerce (see Box 9 on the role of GATS).The four part discusses the role of the WTO in trade facilitation with emphasis on the ways that the Internet and Electronic Data Interchange can simplify trade and customs administration. The fifth part looks at the way in which electronic commerce could transform the traditional approach to government purchases through the use of electronic technologies. The sixth part deals with trade-related aspects of intellectual property rights and discusses the importance of protecting copyrights and related rights, and trademarks and domain names for the future development of electronic commerce. The final part examines regulatory issues relating to electronic commerce from a WTO perspective.

4 Conclusion

This paper has presented e-business has many advantages, such as that fixed costs associated with trade are reduced, that benefit the international trade. As the number of firms with better information increases, the effect of distance on trade is actually likely to increase.

With the growth of e-business, many aspects of international trade have been influenced. E-business is reshaping the way companies go to the international market , and e-business and export performance is reinforcing each other, and E-business increase export, and international e-business influences legal rules, and how e-business should be dealt with is negotiated in the WTO. From the analyses in the paper, it can be forecast that the degree of the impact of e-business on international trade will be deepened.

References

[1] Louis A. Lefebvre, E´ lisabeth Lefebvre, (2002). E-commerce and virtual enterprises: issues and challenges for transition economies, Technovation, 22:313–323

[2] K. Lal., (2002). E-business and manufacturing sector: a study of small and medium-sized enterprises in India, Research Policy ,31:1199–1211

[3] K. Lal. (2004).E-Business and Export Behavior: Evidence from Indian Firms, World Development, 32:505–517.

[4] Friedman, T.L., (1999).The Lexus and the Olive Tree. Anchor Books, New York, NY

[5] Freund, C., Weinhold, D., (2004).The effect of the internet on international trade, Journal of International Economics ,62:171–189

[6] George R.G. Clarke, (2007).Has the internet increased exports for firms from low and middle-income countries? Information Economics and Policy,36: 1-22

[7] Gordian Ndubizu, Bay Arinze, (2002). Legal determinants of the global spread of e-commerce , International Journal of Information Management,22: 181–194

[8] Carlos A.,Primo Braga., (2005) .E-commerce regulation: New game, new rules?, The Quarterly Review of Economics and Finance ,45:541–558

Design and Implementation of EIC Algorithm in Mobile Business

Qin Wang Runtong Zhang Zhongyue Sun

Institute of Information Systems, Beijing Jiaotong University, Beijing, 100044, China

Email：melonzn@sina.com

Abstract

Faced with these problems about the limited storage space and security risks of file downloading in mobile phone, a new algorithm named by EIC (Encryption in Compression) Algorithm is designed in this paper, which is based on a famous compression algorithm–Huffman, to simultaneously realize both compression and encryption once through adding key in the process of compression. It could achieve two functions of encryption and compression through one computing, not only to save the storage space of user's mobile phone but to avoid malice transmission of file downloading. Through emulated experimentation, it has desired effect which is suitable for file downloading service in mobile phone.

Keywords：Mobile Phone; Huffman Algorithm; EIC Algorithm; File Downloading; Compression; Encryption

1 Introduction

File downloading service in mobile business is becoming more and more popular with the development of mobile technology. The service usually consumes more storage space while the storage ability of mobile phone is relatively limited. At the same time, when enjoying the service, mobile users also inevitably face some security threats, such as data being intercepted or copyright-protected files being used by unauthorized users. So, here the problem coming, for one thing, encryption should be done to protect the security of downloading file, for another, because of the limited storage space of mobile phone, the downloading file needs to be compressed to reduce resource consumption. Although the two goals could be achieved by running two algorithms respectively, it is a waste of time and resource. One single algorithm that can simultaneously accomplish both compression and encryption is required urgently.

The method of solving this problem is to find a appropriate compression algorithm and add the key in the process of compression, which means improving a compression algorithm to make it have the function of encryption. Therefore, compression and encryption are realized through one computing. In this paper, a new algorithm named by EIC (Encryption In Compression) algorithm is presented through improving Huffman compression algorithm to realize encryption and compression with one computing.

2 Huffman algorithm summary

Huffman algorithm was presented at the beginning of last century. Accoring the frequency of symbol, it could generate the shortest encoding in the average length. It is a variable encoding. Huffman Algorithm is selected in this paper mainly basing on:

1) Huffman algorithm is the best compression algorithm to processing text files, which is specially good at processing binary text files [9]. Storage files types in mobile phone mainly include .jpg, .mp3, .avi and binary text etc. The binary text file's processing will be studied in this paper.

2) Huffman algorithm is more suitable for mobile phone with relatively low processing capability [5]. In

this paper compression and encryption to binary text file in mobile phone is presented.

Huffman encoding is the best method among all the lossless compression algorithms. It uses different 0-1 strings to replace each symbol, and the length of the 0-1 string is decided by the frequency of the given symbol. Common symbols need few bits to be substituted, while uncommon ones need many bits [3]. Usually, it uses a binary tree to help to generate the codes. In the end, the total length of the files will be much shorter than before. In this way, it achieves a good performance in compression.

3 EIC algorithm design

3.1 Design Thought

EIC algorithm is based on Huffman algorithm which is a famous compression algorithm. The only alteration is, in Huffman it is settled to mark the left node "0" and the right node "1", but in the EIC algorithm "0" or "1" is decided whether this operation is in the "key seat" [8]. That is to say, the array of integer key[ ] decide where "key seats" are. When in a normal seat, the marking operation is as the same as Huffman, but if in a key seat, the 0-1 will be marked opposite. In this way, the characteristics of Huffman and the compression efficiency will not be changed, but the only difference is that EIC involves the key array. So without the key, the illegitimate users cannot decompress the cryptograph back into the original text. This can be regarded as a sort of encryption. Thus, the EIC has realized two functions, compression and encryption.

3.2 Design Process

There are several steps to be implemented in this algorithm as follows. And before that, It is assumed that an array of integer has gotten as the key, which named by int key[ ].

Step1, Figure out the frequency of each symbol.

Step2, Arrange the frequencies from small to big.

Step3, Set up the Binary Tree.

Each time pick up the two minimum frequencies and make them to be leaf nodes of the binary tree while the summation to be the root node. After this, the two leaves are no longer involved in comparison, but the new root node prepares to be compared.

Marked rule of 0-1 is as follows:

1) In this marking operation, there is a counter n. Every time we mark 0-1 for two leaves, n++.

2) Before marking, check whether the n is equal with key[i]. If n is equal with a value in the key [],the left node is marked by "1" while right node by "0", Otherwise, left node "0" and right node "1".

That is to say that the key marks the location of left and right node reversal. This change will retain characteristics of prefix encoding in the original algorithm [6]. It will not affect the depth calculation and encoding endowment, in which only adding the key so that users without having the key don't correctly decode to achieve encryption purposes in the process of encoding.

Step4, Repeat step3, until a root node with a frequency of 1 is gotten.

Step5, By stringing the 0, 1 of each leaf node encountered in the path from the top root node to the bottom leaf node, each symbol's code is generated.

For an example: A group of data is {32,22,22,43,49,22,22,17,48,43}. And int key[ ]={2,3} is assumed. The result of statistics frequency is shown as in Table1.

Table 1 The result of statistics frequency

Symbol	Number	Frequency
22	4	2/5
43	2	1/5
17	1	1/10
32	1	1/10
48	1	1/10
49	1	1/10

The result of node arrangement is shown in Figure1.

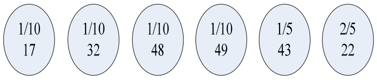
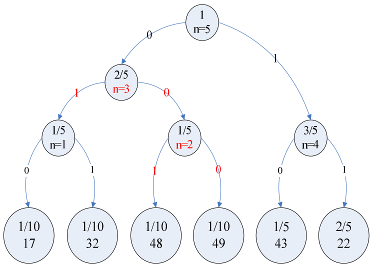


Figure1 The result of node arrangement

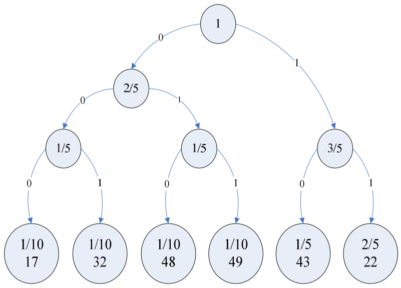
A binary tree is found as shown in Figure2 (Top) and the result of encoding is shown in Fig.2 (Bottom).



Symbol	Frequency	Code
22	2/5	11
43	1/5	10
17	1/10	010
32	1/10	011
48	1/10	001
49	1/10	000

Figure2 The Binary Tree and Encoding of EIC

But, the result of Huffman algorithm is shown in Figure3.



Symbol	Frequency	Code
22	2/5	11
43	1/5	10
17	1/10	000
32	1/10	001
48	1/10	010
49	1/10	011

Figure3 The binary tree and encoding of Huffman

Through comparison, we could know:

1) The original 80 bit is compressed to 24 bit and compression rate is no difference between Huffman and EIC algorithm.

2) If the key [] is not gotten, decoding isn't correctly implemented and so have the function of encryption.

3) EIC algorithm not only realizes the protection of copyright, but also lessens the storage space of data.

3.3 Derivation of Encryption Key

Usually a mobile user enjoys file downloading service through login in service provider's website. The mobile phone number of the user is the only, and so his login password on the website. Therefore, encryption key is generated by the user's mobile phone number and login password.

The process of key derivation is shown as follows:

Step1, A string is generated by connection of the mobile phone number and password.

Step2, The string will be converted to an array of bytes.

Step3, The byte number of the array is computed and then the array is grouped by a group of 8 bytes. (8 may be changed into the other numbers, because the encryption algorithm - RC4 to be selected don't limit to key length). If not enough 8, appropriate amount of zero bytes is added behind the string to make it 8 integer times. For example, if a byte array contains 53 bytes, 3 bytes is added behind the byte array to make it to have 56 bytes.

Step4, Every even-number-array byte is to be overturned, meaning that the first group is unchanged, the second group overturned, and the third group unchanged, the fourth group overturned and so on….

```
for( i = 0; i < b.length / 8; i++)
    if (i / 2 == 0){
        for (j = 0; j <8; j++){
            t = b[i][j];
            b[i][j] = b[i][8-j-1];
            b[i][16-j-1] = t;
                         }
}
```

Step5, XOR is operated between the first group and the second group, and then XOR is operated between the result of XOR and the third group, and so on, until the last group finishs XOR operation. When operations of all groups are finished, the final result is a array containing 8 bytes and so it is an initial key.

```
for (i = 0; i < b.length / 8; i++)
    b[i+1] = b[i] ^ b[i+1];
```

Step6, Parity of key is changed and the last byte is selected as parity checked byte. In each byte, totals amount of binary bit "1", if amount is even, checked byte will be home to "1", if not, checked byte for "0."

After the aboved six steps, key of symmetric encryption algorithm is eventually generated.

4 EIC algorithm implementation

Emulated Environment

Inter(R) 1.70GHz 504 MB Memories

Eclipse 3.2 WTK 2.2

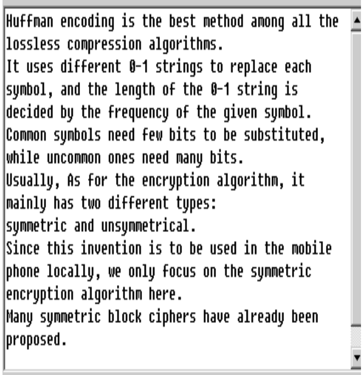
Emulated Result

The simulation experiment has main two parts:

1) Without EIC algorithm, directly read and save.

2) Adopt EIC algorithm, compress and encrypt before saving.

Firstly, the original text which will be used in this demonstration is as follows in Figure4.



Huffman encoding is the best method among all the
lossless compression algorithms.
It uses different 0-1 strings to replace each
symbol, and the length of the 0-1 string is
decided by the frequency of the given symbol.
Common symbols need few bits to be substituted,
while uncommon ones need many bits.
Usually, As for the encryption algorithm, it
mainly has two different types:
symmetric and unsymmetrical.
Since this invention is to be used in the mobile
phone locally, we only focus on the symmetric
encryption algorithm here.
Many symmetric block ciphers have already been
proposed.

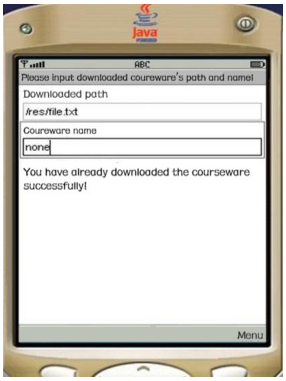
Figure 4 The interface of original text

Here, it is assumed that the user's password is "bamboo", and the number of phone is "13811197416".

1) Without EIC Algorithm

When mobile users download courseware, path and name of the courseware should be input and denominated. If the courseware has been downloaded successfully, "You have already downloaded the courseware successfully" will be displayed on the screen. Otherwise it would be "The courseware is downloaded unsuccessfully". See in Figure5.

When this file is read, the user will see the original text because no algorithm is adopted, as shown in Fig.6.



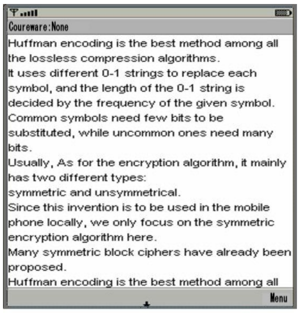
Please input downloaded coureware's path and name!

Downloaded path: /res/file.txt

Coureware name: none

You have already downloaded the courseware successfully!

Figure 5 The interface of courseware downloading



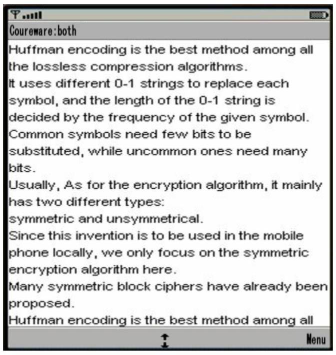
Coureware:None

Huffman encoding is the best method among all
the lossless compression algorithms.
It uses different 0-1 strings to replace each
symbol, and the length of the 0-1 string is
decided by the frequency of the given symbol.
Common symbols need few bits to be
substituted, while uncommon ones need many
bits.
Usually, As for the encryption algorithm, it mainly
has two different types:
symmetric and unsymmetrical.
Since this invention is to be used in the mobile
phone locally, we only focus on the symmetric
encryption algorithm here.
Many symmetric block ciphers have already been
proposed.
Huffman encoding is the best method among all

Figure6 The interface of text without algorithm

2) Adopt EIC algorithm (compress and encrypt before storage)

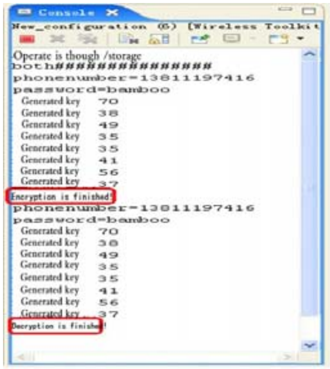
Now, the same file is downloaded with adopting EIC algorithm, which means the file will be compressed and encrypted before storage and decompressed and decrypted before reading. The file will be named by "both" this time. After downloading this courseware successfully, the right original text can be viewed in Fig.7. In Eclipse's console, the information of generated key and the successful information of encryption and decrypiton are shown in Figure8.



Coureware:both

Huffman encoding is the best method among all
the lossless compression algorithms.
It uses different 0-1 strings to replace each
symbol, and the length of the 0-1 string is
decided by the frequency of the given symbol.
Common symbols need few bits to be
substituted, while uncommon ones need many
bits.
Usually, As for the encryption algorithm, it mainly
has two different types:
symmetric and unsymmetrical.
Since this invention is to be used in the mobile
phone locally, we only focus on the symmetric
encryption algorithm here.
Many symmetric block ciphers have already been
proposed.
Huffman encoding is the best method among all

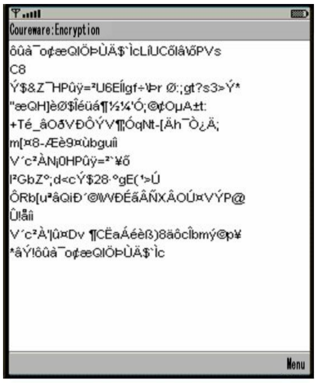
Figure7 The interface of text with EIC algorithm



```
Operate is though /storage
both####################
phonenumber=13811197416
password=bamboo
 Generated key   70
 Generated key   38
 Generated key   49
 Generated key   35
 Generated key   35
 Generated key   41
 Generated key   56
 Generated key   37
Encryption is finished!
phonenumber=13811197416
password=bamboo
 Generated key   70
 Generated key   38
 Generated key   49
 Generated key   35
 Generated key   35
 Generated key   41
 Generated key   56
 Generated key   37
Decryption is finished!
```

Figure8 The interface of encryption and decryption

For a legal user, when he is downloading the courseware, it will be encrypted and compressed, and before the courseware is read, it will be firstly decrypted and decompressed. But the file saved in the mobile phone is the one having been encrypted and compressed. See in Figure9. Only the legitimate users could decrypt and decompress the file and so protecting file copyright.



```
ôûà¯o¢æQIÖÞÙÄ$`ÌcLÌUCõlâ\õPVs
C8
Ý$&Z¯HPûÿ=²U6EÍlgf÷\Þr Ø:;gt?s3>Ý*
"æQH]èØ$Îéüá¶½¼'Ó;©¢OµA±t:
+Té_âOðVÐÔÝV¶ÓqNt-[Äh¯Ò¿Ä;
m[¤8-Æè9¤ùbguiî
V´c²ÀN¡0HPûÿ=²`¥õ
l²GbZ°;d<cÝ$28·°gE(¹>Ú
ÔRb[u²âQiÐ´©WÐÉãÂÑXÂOÚ¤VÝP@
Û!åiì
V´c²À'|û¤Dv ¶CËaÁéèß)8äôcÎbmý©p¥
*âÝ!ôûà¯o¢æQIÖÞÙÄ$`Ìc
```

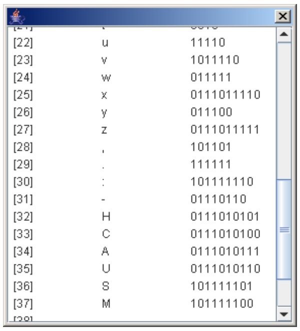
Figure9 The interface of encryption text

5 EIC algorithm analysis

5.1 Encryption Intensity Analysis of EIC Algorithm

Firstly, if there are N nodes in the original text, then there would be（N-1）+（N-2）+……+1 evaluation points of 0-1. And there are two possibilities of 0-1 and 1-0 in every point when the key and it's length isn't known. So if the exhaustion method is adopted to decode, then $2^{N(N-1)/2}$ times is experimented. And when there are more than 15 nodes, this frequency would be larger than 2^{100}. And this number is as large as $1.3*10^{30}$. So the expense of doing this is lager than that of payment.

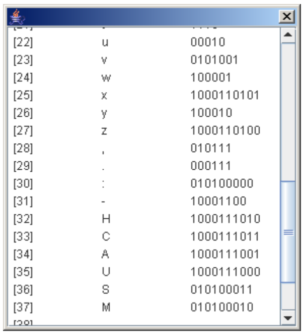
Consequently, from the view of theory, this algorithm is strong enough. On the other hand, from the view of application, the result of special encoding is shown in the following text. If Huffman is adopted, the result of encoding is shown in Figure10.



```
[22]    u    11110
[23]    v    1011110
[24]    w    011111
[25]    x    0111011110
[26]    y    011100
[27]    z    0111011111
[28]    ,    101101
[29]    .    111111
[30]    :    101111110
[31]    -    01110110
[32]    H    0111010101
[33]    C    0111010100
[34]    A    0111010111
[35]    U    0111010110
[36]    S    101111101
[37]    M    101111100
```

Figure10 The encoding result of Huffman

If EIC is adopted, the coding result is shown in Figure11.



```
[22]    u    00010
[23]    v    0101001
[24]    w    100001
[25]    x    1000110101
[26]    y    100010
[27]    z    1000110100
[28]    ,    010111
[29]    .    000111
[30]    :    010100000
[31]    -    10001100
[32]    H    1000111010
[33]    C    1000111011
[34]    A    1000111001
[35]    U    1000111000
[36]    S    010100011
[37]    M    010100010
```

Figure11 The encoding result of EIC

Apparently, the coding result of EIC is different from that of Huffman and so EIC algorithm has an effective encryption function. Without key in stored code in the mobile terminal, files aren't correctly gotten through Huffman tree.

6 Compression Efficiency Analysis of EIC Algorithm

Limited storage space is a limiting factor of mobile terminal. So the efficiency of compression in EIC algorithm is examined.

An experiment result is as follows:

1) The size of the local file which is used to be downloaded is 1.26KB.

2) When EIC algorithm is not used, the size of this file which is stored in mobile phone is 3.21KB in RMS. (The size of this file is lager than that of the original file because this record contains not only the context of original file but some other useful information)

3) When EIC algorithm is used, the size of this file is 1.15KB in RMS.

Another experiment result is 9.98KB, 30.2KB and 8.18KB respectively.

By the aboved data comparison, we could see that EIC can compress data effectively. And we could see that the length of EIC encoding is the same as that of Huffman encoding from the Fig.10 and Fig.11. In other words, compression efficiency of EIC is the same as that of Huffman.

7 Conclusion

A new algorithm named by EIC algorithm is designed in this paper, which realizes compression and encryption of downloading file once. EIC algorithm is developed on the basis of a famous compression algorithm—Huffman and could realize encryption through adding a key in the process of compression. Because of endowed value exchange between Huffman's special left and right child through the key (Key is computed through a certain algorithm according to user's mobile phone number and username), EIC algorithm could achieve two function of encryption and compression through one computing.

References

[1] Kawamura, Muling Guo, "Encoding of Multi-alphabet Sources by Binary Arithmetic Coding", SPIE Proceedings Vol. 3653, pp.1041-1049, 1998

[2] Xu Hui, "Cipher Analysis about Arithmetic Encryption Mechanism", Shanghai Jiaotong University [D], 1999

[3] Wang Liguang, Wang Minling, "Many Crossed Trees in Huffman Algorithm", Journal of Nanhua University (Natural Science Version), Oct.2004

[4] Wang Tong, "Design and Application of a Dynamic Huffman Optimized Algorithm", Journal of the Air Force Engineering University (Natural Science Version) ,No.2, Jun. 2005, pp.91-93

[5] Wu Lenan, Data Comprssion (The second edition), Beijing: The Electronic Industry Pub., Oct. 2005

[6] Li Huang, "A Random Public Key Encryption Algorithm and a Compression Algorithm", Journal of the Science Information Development and Economy, Vol.16, No.1, 2006, pp.225-226

[7] Zhan Jianfei, J2ME Devlopment Guide, Beijing: The Electronic Industry Pub., Jan. 2006

[8] Li Juan, "Improvement and Application of Huffman Algorithm", Journal of Computer Knowledge and Technoloy, Feb. 2006, pp.59-61

[9] Yongkang Peng, "Implementation of Huffman Algorithm", Journal of Zhengzhou Railway Occupation Technical College, Jun. 2006

[10] Xiong Xuan, Fu Yingfang, "Research of Data Compression in Optimized Huffman Algorithm", Journal of Computer Knowledge and Technology, Jun.2006

The Constructing of Virtual Enterprise Based on Multi-Agent System

Xuefeng Wang

School of statistic, Jiangxi Finance and Economics University, Nanchang, Jiangxi 330013, China
Email: wxfyf@163.com

Abstract

In the rapidly changing competitive global environment, a dynamic collaboration among companies is required to enable processes of the partners to be combined in a timely and cost effective manner to combine a complementary set of competencies to gain new market opportunities. Advanced manufacturing systems need to be developed for enterprises to cooperate with each other in order to survive in increasingly competitive global market. Based on the analysis of the feasibility and effectiveness of multi agent technology for virtual enterprises, we try to identify the feasibility and effectiveness of multi-agent system (MAS) and Web services technology for virtual enterprise (VE) and put forward a multi-agent based reference architecture of VE in this paper. An agent generic framework for wrapping the autonomous enterprises of VE is presented and the running mechanism of virtual enterprise in whole life cycle of VE is discussed.

Keywords: Multi-agent, virtual enterprise, web service, whole life cycle

1 Introduction

In recent years, the markets present continually unpredictable changes, and the speed of product innovation and technology innovation is much quicker. The quick response to the environment changes becomes the main problem for the manufacturing enterprises. As a result, enterprise cooperation vertically along a supply chain and horizontally among peers becomes more and more significant for these enterprises to survive in the increasing competitive global market. A VE is a temporary alliance of enterprises to shall skills of core competencies and resources in order to respond to business opportunities, and the cooperation among the enterprises is supported by computer networks [1]. In a VE, manufactures no longer produce complete products in isolated facilities. They act as nodes in a network of suppliers, customers, engineers, and other specialized service function [2]. VE permits bridging the gap between big enterprises and small firms in terms of external behaviors through the support of a common VE working environment. The concept of the VE gives rise to an entirely new type of entrepreneurs, termed as agents, that co-ordinate the activities of the independent partners [3]. Agent-based manufacturing has become a new paradigm for next generation manufacturing systems, together with other recently emerging manufacturing paradigms such as Holonic Manufacturing Systems, Agile manufacturing, and Reconfigurable manufacturing [4]. Multi-agent based architectures for manufacturing systems appear to provide adequate responses to such requirements since their distributed nature provides flexibility and reactivity to changing situations. Several intelligent-based architectures for manufacturing systems have been proposed in the literature: examples are the NIST Real-Time Control System (RCS), the MetaMorph Architecture, and the AARIA project [5]. In view of the fact that these distributed organization are generally managed by heterogeneous software systems running on heterogeneous computing environments, the recently emerged Web services technology provides a higher level interoperability for connecting business activities across the Web both between enterprise and within enterprises. In this research, we try to identify the feasibility and effectiveness of multi-agents system and Web service technology for modeling and control of VE, and put forwards a multi-agent based reference architecture of virtual enterprise. An agent generic framework for wrapping the autonomous enterprises of virtual enterprise is presented and the running mechanism of virtual enterprise in whole life cycle of VE is discussed.

2 Agent Architecture and Multi-Agent System

An agent is supposed to act spontaneously, executing pre-emptive and independent actions that eventually benefit the user through accomplishment of the assigned goals [6]. Different kinds of agents can be characterized and other sophisticated features are usually needed to make then act intelligently in their respective environment. It is usually assumed that the core of an agent includes the three characteristics: it perceives the world in which it is situated; it has the capability of interacting with other agents; it is pro-active in the sense that is may take the initiative and persistently pursues its own goals. Agent architecture is a map of the internals of an agent and it data structure, the operations that may be performed on these data structures, and the control flow between these data structures. In order to utilize MAS to carry out VE reference architecture and improve the universality of agent, each autonomous enterprise has to be encapsulated as an agent and its general architecture can be represent as three element form: <Agent kernel, interface, Agent shell> . The general architecture for agent in my paper can be derived as Figure 1.

The kernel of agent represents the practical operation complexion in enterprise and can be realized by workflow management system software. The shell of agent is responsible for the communication and collaboration with other agent. The interface is used to communicate the information between the kernel and shell. For instance, the kernel of agent can transfer product status, product capacity and performance index to the shell of agent by transforming these information into understandable format; at the same time, the shell of agent can transforms the information derived from the results of communicating and negotiating with other agents into understandable format and instructs the action of the kernel by transferring information to the kernel of the agent. The shell of the agent composes of working storage, communication module and collaboration module. In order to facilitate the augment and modification of functions, each module in the shell adopts object-oriented module design.

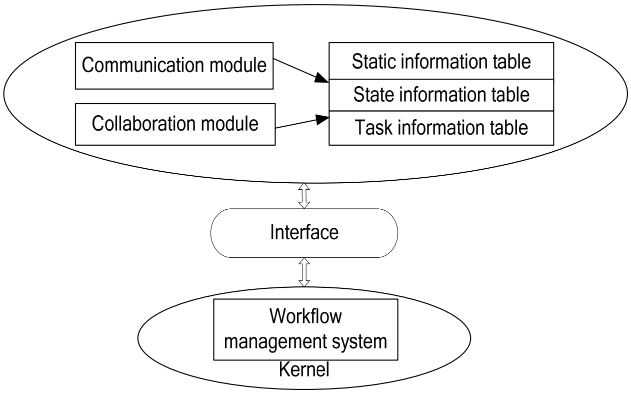


Figure 1 The general architecture of the Agent

Multi-agent system emerged, as a scientific area, from the previous research efforts in distributed artificial intelligence stated in early eighties. The concept of an agency is now being broadly used not only as a model for computer programming units display certain kind of characteristics but also in more abstract and general way, as a new metaphor for the analysis, specification, and implementation of complex software systems[7]. A MAS can be defined as a collection, possibly heterogeneous, computational entities, having their own problem-solving capabilities and which are able to interact in order to reach on overall goal. The basic application principle of MAS is pursuing the global objective and synchronously maximizing the autonomy of each VE member. Each enterprise member is autonomous entity in VE, therefore, each enterprise entity and VE can be naturally represented by agent and MAS respectively. The agent representing enterprise possess of stronger autonomous ability than the agent representing manufacture resources. The traditional information system cannot fit for new environmental characteristic of VE, such as distributed, dynamic and heterogeneous, whereas, multi-agent system entirely meets those special requirement, especially in hereinafter aspect: guarantee the independence and autonomy of enterprise member; facilitate the harmony between enterprises through applying new communication and information mode; scheme the enterprise behavior based on the negotiation and cooperation between enterprises.

3 Virtual Enterprise Infrastructure and Agent Functions

3.1 Virtual enterprise infrastructure

VE can be considered as a temporary alliance of globally distributed independent enterprises that participate in the different phases of the life cycle of the product or service, and work to share resources, skills, and costs, supported by Information and Communication Technologies, in order to better attend market opportunities and successfully fulfill a responsible corporate strategy[6]. Participating to a VE project may help companies to increase its market share and benefits. A large variety of organizations can be set up according to the industrial option taken by the VE partners. It is possible to identify three groups of characteristics, which would mainly influence the kind of VE organization: market characteristics; production process; strategic objectives of the association. The characteristics of these groups are tightly linked and each group can be seen an aggregation of several parameters. But the key point of a VE project is customer demand, so VE structures are highly dynamic and its life cycle can be very short, reactivity and flexibility, which are the major benefits of VE, are a source of problems to solve.

The main objective of VE infrastructure is to link different organizations to make the work together in a collaborative and reactive manner. The entire VE project can be decomposed in elementary activities for which a management system has to be developed. To manage all these activities, procedures and interfaces have to be precisely defined, and a large investment in time and work is needed. This will end up with a large and complex system, which has many chances to be non-flexible. This is contrary to the main objective of VE creation: reactivity. A compromise must be found between the management structure accuracy and its size. The size of the structure will be adapted to the length of the project life cycle. Taking those into account, we put forward an infrastructure of VE based on MAS and Web services technique is as Figure 2

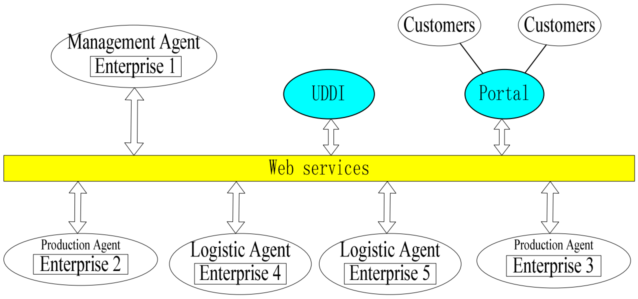


Figure 2 The infrastructure of VE based on MAS and Web services technique

The proposed architecture is composed of three sub-types of agents, namely management agent, production agent and logistics agent, UDDI (Universal Description Discovery Integration) and a common Web portal. Web services and multi-agent technologies are combined to provide an integrative solution for enterprise collaboration. We can see from this figure that all the components are connected to the Internet through Web services. In this enterprise infrastructure, UDDI becomes the enterprise register center and any service provider can enable its business to be programmatically accessible on the internet and registers their service to UDDI distributed, and heterogeneous partner can search dynamically accessible service by UDDI. There are some inherent advantages that make web services the ideal platform for realizing VE. This is important for VE and both within and between organizations there will be a plethora of Platforms and languages [8]. Agents communicate semantically using Agent Communication Language encapsulated in message of SOAP (Simple Object Access Protocol) over HTTP. Web services operate using a set of open standards, such as the SOAP, WSDL (Web Service Description Language) and UDDI based on the well-known platform and vendor independent standard, XML (Extensible Markup Language). With Web services integration of heterogeneous systems becomes seamless. Confining communications to SOAP/XML over HTTP, many Web applications can reach the point of true interoperation.

3.2 Agent functions

Although Web-based technologies provide effective implementation-level means to integrate distributed heterogeneous Web applications, they provide no mechanisms to facilitate the problem-solving capabilities of distributed components which usually need coordination, collaboration and negotiation processes. The MAS approach can more easily cope with the heterogeneity and autonomy of participating components and therefore be more suitable for the integration of legacy systems [9,10]. The management agent in this system is the heart of MAS and acts as an organizer and manager of the VE. It is responsible for accepting the product orders from customers, creating a VE, decomposition of orders, coordinating the global operation of VE and taking on certain production and logistics task. Production agents represent various enterprises composing of VE and are responsible for the execution of such assigned tasks. Production agents accept the production task through negotiating with management agent and fulfill the final production by cooperating with other agent. Production agents comprise two sub-types of agents, namely planner agents and control agents. Planner agents are responsible for creating and optimizing execution plans for the task assigned by higher-level commands. Control agents are responsible for the execution of such assigned tasks. Control agents receive and process information from planner agents, environment sensors, mediator agents, and higher level agents. Logistics agent represents the enterprise which is charged with the transmission of the raw material and production between enterprises and corresponds to a given logistics service. Note that the role of enterprise in VE is not changeless all the time, for example, an enterprise acts as management agent in a VE for a time, whereas possibly takes on production Agent or logistics Agent in other time.

4 The Life Cycle of Virtual Enterprise and System Operation

4.1 The life cycle of virtual enterprise

The life cycle of a VE can be represented as a sequence of steps organized in four phases: creation/configuration, design, execution, and dissolution. The VE is set up with the objective of making one particular type of product or delivering one particular type of service [11]. When the market of that product/service declines then the VE smoothly dissolves, allowing partners to find new missions to pursue some new opportunities. Taking account of this, we can describe the total life cycle of the VE in Figure 3.

4.2 System operation

During the first phase of the development of a VE, the creation phase, management agent might search and recognize the market opportunities, planning a formation of a new VE. Analyzing the new market requirement, this agent must plan and draft the value chain that will eventually comprise the VE, and estimating costs and revenues. Through querying public UDDI or private UDDI, the management agent acquires the published information about corresponding service providers. The management agent has to estimate all necessary skills and competencies, satisfying the imposed requirements and to plan the associated partnership organizational architecture of the potential new VE through negotiation with service providers based on the relation information described by WSDL.

During the design phase, based on the evaluating results in the creation phase, all enterprises belonged to the VE can register in the private UDDI which is constructed by the VE. The Web services technologies along are believed to provide a loosely coupled integration

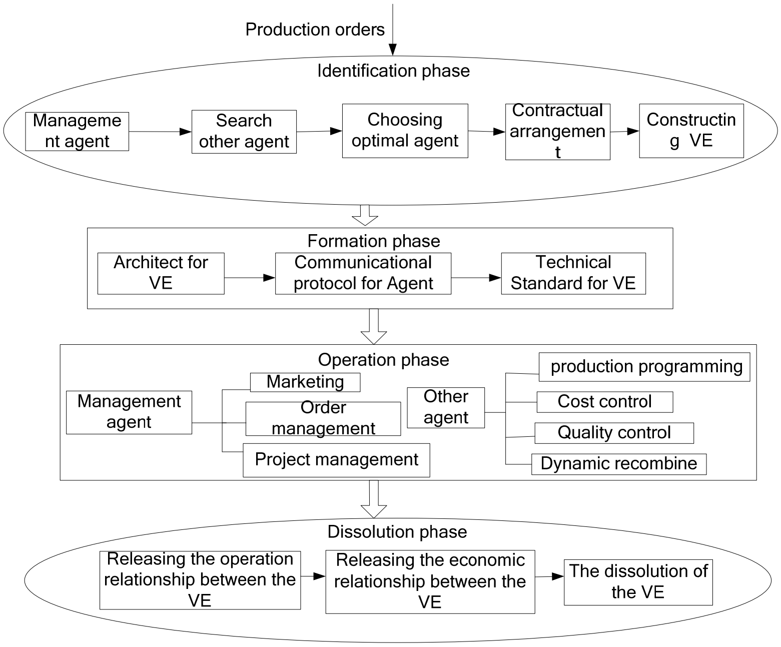


Figure 3 The whole life cycle of the VE

solution for enterprise integration. All members committed to the VE can communication with each other through SOAP/XML over HTTP and realize the seamless integration of heterogeneous systems. Using Web services, all the business logics inside an enterprise can be encapsulated as different modular components, so the complexity of internal system is totally transparent to outside applications. In the phase, the management agent and other agents specify detailed procedures for carrying out the mission. These procedures include the design of the new products, development of all material and information flows and the corresponding control and information systems, as well as business transactions between the VE partners, involving negotiations, commitments, contracts, shipping and logistics, payment instruments.

During the operation phase, management agent schedules and synchronizes the partners' operational plans. The agent is not directly involved in the day-day running of enterprise but rather stands ready to be called upon for advice and to settle disputes among the VE members. Each enterprise member connects to Internet and it protected by a firewall for security and privacy reasons. In the VE environment based on Web service, each enterprise adopts Web services manner to realize it function, this is to say, Web services can be combined to accomplish a given project.

Finally, during the dissolution phase the main task of the management agent is to archive the mission documentation, to distribute all data and related information to partners, and to arrange after-sale services and customer support. When the whole assigned task is accomplished, all members logout their function from the private UDDI and the private UDDI is deleted subsequently.

5 Conclusions

The development of an advanced manufacturing paradigm is being strongly driven by the economic factors worldwide and facilitated by the recent developments in the information and communication technologies. Cooperation among enterprises is very important both for the vertical supply chains and horizontal virtual enterprises to survive in the global competitive market. Multi-agent technology is considered as the most serious method for the research on distributed intelligent system.

On the basis of analysis on the feasibility and effectiveness of multi agent technology for virtual enterprises, an agent generic framework for wrapping the autonomous enterprise of VE was presented in this paper. Each autonomous enterprise can be encapsulated as an agent, realizing the interior functions separated from of the VE and the essential function mapping of the original systems.

Shorter lifetime, loose coupling, dynamically changing partners, better scalability become a new trends of VE. In this environment it is necessary to study methods for implement dynamic virtual enterprise. A multi-agent and Web service based reference architecture of virtual enterprises is proposed in this paper. In addition, the running mechanism of virtual enterprise in whole life cycle of VE is discussed. In this enterprise infrastructure, UDDI become the enterprise register center. Any service provider can enable its business to be programmatically accessible on the Internet and registers their service to UDDI distributed. Heterogeneous partner can search dynamically accessible service by UDDI and interoperate with other partner using standard data exchange format XML and message interchange protocol SOAP. Based on the general infrastructure, enterprise member interoperates with each other only through the communication modules and ensure the private section—the kernel of enterprise is not influenced by the outside factor. So the infrastructure not only can implement the interoperation and information exchange among enterprise, but also preserve the autonomy ability of enterprise. The infrastructure facilitates the reuse, reconstruction and extension of VE and increases the feasibility of a enterprise joining in several VEs at one time. In addition, Enterprise can dynamically drop out or join VE to gain the maximum global benefit.

References

[1] N. Wu, and P. Su, "Selection of partners in virtual enterprise paradigm," Robotics and computer integrated manufacturing, Vol. 21, 2005, pp. 119-131

[2] M.T. Martinez, P. Fouletier, K.H. Park, and J. Favrel, "Virtual enterprise-organization, evolution and control," International of Journal of Production Economics, Vol. 74, 2001, pp. 225-238

[3] Jgdev, H., and Browne, J., "The extended enterprise- a context for manufacturing," Production planning and Control, Vol. 9, 1998, pp. 216-229

[4] Q. Hao, W. Shen, and L. Wang, "Towards a cooperative distributed manufacturing management framework," Computers in Industry, Vol. 56, 2005, pp. 71-84

[5] N. G. Odrey, and G. Mejia, "A re-configurable multi-agent system architecture for error recovery in production system," Robotics and Integrated Manufacturing, Vol. 19, 2003, pp. 35-43

[6] R. Chalmeta, and R. Grangel, "ARDIN extension for virtual enterprise integration," The journal of systems and software, Vol. 67, 2003, pp. 141-152

[7] L. Mihailov, "Fuzzy analytical approach to partnership selection in formation of virtual enterprises," Omega, Vol. 30, 2002, pp. 393-401

[8] E. Oliveira, K. Fischer, and O. Stepankova, "Multi-agent systems: which research for which applications," Robotics and Autonomous System, Vol.27, 1999, pp. 91-106

[9] Y. Peng et al. , "A multi-agent system for enterprise integration," International Journal of Agile Manufacturing, Vol. 1, 1998, pp. 213-229

[10] L. Song, R. Nagi, "Design and Implementation of a Virtual Information System for Agile Manufacturing," IIE Transactions, Vol.29, 1997, pp. 839-857

[11] M. R. Stytz, T. Adams, B. Garcia, and B. Zuritz, "Rapid Prototyping for Distributed Virtual Environments," IEEE SoftWare, Vol.14, 1997, pp. 83-92

Research on the Value of Blog Marketing

Licheng Ren[1] Jing Li[2]

1 School of Economics and Management, Taiyuan University of Science and Technology, Taiyuan, Shanxi, China
Email: rlc2000@sina.com

2 Department of Management, Xinhua College of Sun Yat-sen University, Guangzhou, Guangdong, China
Email: littlebench117@163.com

Abstract

With human being entering into the 21st century, Network Economy emerges and nearly all companies are working online by using electronic commerce to do their web marketing. As a new network application, Blog has caught the eye of many people and multitudinous enterprises, which has also become one kind of new marketing model. This paper has constructed the value model of Blog Marketing through the comparative analysis between Blog Marketing and other models of marketing. Finally, some Blog operation tactics has been proposed.

Keywords: Blog, Blog Marketing, Blog value, Electronic Commerce, Network marketing

1 Introduction

Since 1990s, computer network technology has made a rapid development. With the popularization of the Internet, information processing and transmission break through the traditional constraints of time and space. Networking and globalization become an irresistible trend, and the Internet has become a corporate trend. Enterprises can take advantage of the characteristics of the low-threshold network which is also easily developed and applied to a lower cost, highly efficient information dissemination and web marketing.

With the development of information technology, Blog, as a new form of network application, has come into being and also caught multitudinous enterprises' and people's attention. The year of 2005 is a significant milestone for the development of Global Internet and Blog, and in this year the number of Blog broke through 100 million around the Global, and which reached 16 million in China. Chinalabs commented that nowadays the birthrate of Blog is so surprised, and about only in 74 seconds a Blog established. On December 26, 2007, the China Internet Network Information Center (CNNIC) shows that the number of Chinese Blog author amounts to 46.982 million with the number of Blog space 72.822 million, that is to say, the average level is 1.55 spaces one person in *the 2007 China Blog Marketing research reports*.

At present, Blog has become a new network era term and its value has become increasingly emerged. Now Blog Marketing has also become the general public topic among scholars. The exploration of the value of Blog Marketing is of a positive significance to enterprise's network marketing.

2 Research Scope and Methods

From 2002, Blog in China has entered its fifth year. Five years ago, Blog birthed and it started to enter China. Five years later, Blog is so popular in China and has gradually moved toward the community, which has been synchronized with the world. In 2005, Blog exploded and continued to grow in the next two years, and until 2007 Blog industry has continuously matured, and the ecological Blog industry began to be mature. At the same time, as the portal into Blog market, Blog market pattern in China began to structurally change. How does Blog change? What's the future of Blog? Where is the value of Blog? This is of concern to experts and scholars in recent years, and also of the focus of this paper.

Jeremy Wright, an author and business consultant with a passion for Blogging, communication, time management, and anything else that makes people's lives easier, shows us why we use Blog and how company can use Blog to raise its visibility and transform internal communications in his book *BLOG MARKETING*[1]. John Battelle, on the 2006 Webmaster World Forum conference, said that "Blogs will soon become a staple in the information diet of every serious businessperson . . . Blogs offer an accelerated and efficient approach to acquiring and understanding the kind of information all of us need to make business decisions". Professor Hugh Hewitt of Chapman University, in its "Blog", described how Blog changed the United States and the changes and response strategies which was brought about by Blog[2]. Debbie Weil, whose Blog "Blog WriterForCEOs.com" is called the authority of the business Blog area, proposed the enterprise tactical dimension issues in her new book *The Corporate Blogging Book* [3].

YingJian Feng, the founder of new observation of marketing online, began to pay attention to the network marketing application from 2002, and he firstly raised the definition of Blog Marketing which is widely used in the world.

Based on the theoretical literature and previous studies, this paper constructed the value model of Blog Marketing through the comparative analysis between Blog Marketing and other models of marketing. Finally, some Blog operation tactics has been proposed.

3 Theory Research

3.1 What's blog and blog marketing

There are many views on Blog and different people have different opinions. Simply, Blog, also known as Web Blogs, refers to a particular network publication and an article published manner, advocate the exchange and ideas sharing. One Blog is a web page which is usually constructed by simple articles and frequently updated compositions. These articles are arranged by year and date, and real-time information is communicated through the network. Blog is not purely a technical innovation, but a gradual evolution of network application.

Blog Marketing, simply saying, is to carry out network marketing by using Blog network application form. We regard all the marketing which use the Blog for a variety of products, brand, Image, etc. as Blog Marketing. Blog is open, informative, initiative and sharing, Its basic features make Blog Marketing be a information transmission form which is based on personal knowledge resources(including thinking, experience and other manifestations)[4]. Thus, Blog Marketing based on the learning, mastering, and effective use of the domain knowledge uses knowledge transmission to make the purpose of information marketing.

3.2 What's the feature of blog

As a new network application form, Blog has also become a new media except for traditional media and portal website which has some characteristics as follows:

Firstly, the content and publication manner of Blog article are more flexible [5]. Because of its flexibility, Blog is more popular. The content is personal or enterprise's hobbies, and unrestricted. In addition, professional Blog sites have large quantities of Blog users, so many valuable articles are usually more easily concerned by users, and can be easily searched. From the promotion efficiency, it's higher than a general enterprise website.

Secondly, the amount of information is bigger and the expression form is flexible. The amount of Blog article information completely depends on the needs of the problem description. And under the circumstances warrant, we can adopt text, images, sounds, flash, and other forms.

Thirdly, Blog article appeared to be more formal, more credible. Blog article is not simply used for advertising, its writing is different from the commodities information, and it's operated by the enterprises themselves.

Fourthly, Blog spread is greater autonomy, and without direct costs. It's the lowest cost way to promote.

3.3 Comparison between blog and other tools of marketing

At present, the common network marketing tools and information dissemination media enterprise commonly used are mainly as follows: company's Web site; portal advertising, portal news; industry Web site, professional site which is a supply and demand information platform; and network communities, BBS.

Blog and other media are similar in the dissemination of information, for example, Blog also played the role of network marketing information transformation. But as a new network application form, Blog also has its own new features, as shown in table 1.

Table 1 The comparison of Blog Marketing and other marketing models

	content themes and published form	information capacity	credibility	costs	difficulty level to be captured and searched
Blog	open and flexible	easy more or less	high & objective	low & no direct costs	easy & no costs
company website	official	more	high & objective	high	medium &ranking costs
advertising and news on website	official	less	medium commerce nature	high	hard
supply and demand information exchange platform	official	less	medium	high	hard
community and forum online	open and flexible	easy more or less	high & subjective	low & no direct costs	easy & no costs

4 The Value Model of Blog Marketing

New technology brings enterprises a new way to transmit information. Based on the characteristics analysis of Blog Marketing, a value model of Blog Marketing is constructed on the basis of the content analysis by comparison of Blog and other medias of marketing. As shown in Figure 1.

- Blog content can increase the search engine visibility, so as to bring website visiting. Baidu, Google, Yahoo and other search engines have strong Blog content search function, so we can use Blog to increase the amount of website induced by search engine which can improve the visibility of the search engine. Moreover, the publishing of Blog articles are all free.

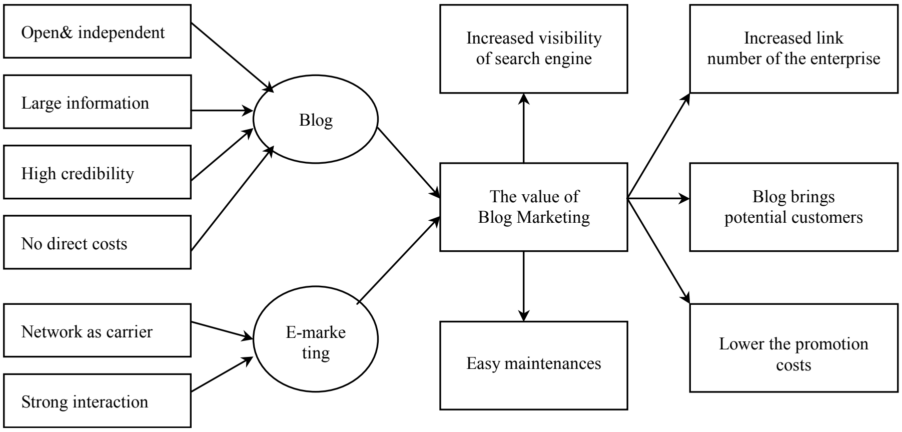

Open& independent
Large information
High credibility
No direct costs
Network as carrier
Strong interaction
Blog
E-marke ting
Increased visibility of search engine
The value of Blog Marketing
Easy maintenances
Increased link number of the enterprise
Blog brings potential customers
Lower the promotion costs

Figure 1 The value model of Blog Marketing

- Blog articles can easily increase the number of links. Though it's difficult for an unpopular site to build a link with a valuable one, they can use their own Blog to do it. So this way can not only bring new sites visiting, also the web site can increase its order in the search engine ranking.
- Blog can bring potential users. Blog content is posted on Blog hosting site, such as Blog site (www.bokee.com), Blogger site which is owned to Google (www.blogger.com), and other portal site platform. These Blog sites often have a large number of user groups, and those valuable Blogs will attract a large number of potential users to browse in order to achieve the purpose of information transferring. This kind of marketing is the basic form of Blog Marketing, and is also the most direct value performance of Blog Marketing.
- Blog Marketing can reduce promotion costs. Through Blog, enterprise can add some appropriate information to the Blog content (such as a hot product links, online coupon download links) in order to achieve the purpose of website promotion. This kind of Blog Marketing has reduced promotion cost which does not increase the website promotion cost, but upgrade site visiting.
- Blog can lower the cost of research of readers' behavior and also reduce the cost of maintaining users. Blog is also a platform for a lot of people to question, answer, comment and exchange information, so customer relationships can be better maintained. At the same time, it will increase the interaction and improve the effectiveness of online survey, which means to lower the research costs.

5 Operation Tactics of Blog Marketing

Although Blog Marketing has no uniform model for different areas and enterprises, its basic idea is same and can be formulated as a basic mode reference. Based on Blog Marketing research, there are three basic forms as follows [6]: use the third-party Blog platform to publish article and do marketing activities; self-built Blog channels to encourage company's employees to write enterprise Blog; A person who is capable of the operation and maintenance can establish his own Blog site for himself to do network marketing.

In view of the characteristics of the enterprise and its marketing, this paper, taking the first form as a case, explores the specific steps which can be summed up to five-step.

Firstly, choose a Blog hosting site and create a Blog account. Select a Blog Marketing platform which is suitable for the enterprise, and get the authority to publish information. Generally speaking, we should choose Blog hosting site whose visiting is large and owned high visibility. For a higher influence site, the credibility of its content is also higher. For small businesses, they can choose a number of Blog hosting sites to registrant at the same time.

Secondly, from the enterprise itself, it should develop a long-term plan to do its Blog Marketing. This plan includes the main elements of the writing staff plan, content field, publishing cycle. Blog writing has its own flexibility and randomness, and there is a need for a longer period to evaluate Blog Marketing working.

Thirdly, we should make Blog Marketing integrate with the whole marketing strategy. Whether a person or a Blog team, if you want to keep a longer value of Blog Marketing, it's necessary to persevere in writing, and enterprises should adopt a reasonable incentive mechanism too.

Fourthly, with comprehensive utilization of Blog Marketing and other marketing resources, enterprises should make sure that Blog Marketing will fully integrate with other marketing. Blog Marketing is not independent but an integral part of the whole marketing activities. In order to make more efficient use of all kinds of resources, combining Blog article with other media has become an indispensable work.

Fifthly, assess the effectiveness of the Blog

Marketing and do the further improvements. Like other marketing strategies, it's necessary to evaluate the effect of Blog Marketing. Continuously improve Blog Marketing plan according to the problems emerged. There are many ways to evaluate its effect, you can refer to other evaluated methods about network marketing, or learn from some successful enterprises to get their experiences.

6 Conclusions

As a typical application of Internet 2.0, Blog is a hot topic in the world. The concept of enterprise Blog and Blog Marketing and its application has been widely concerned in the global context, and Blog Marketing has also entered a new field. Because of its incomparable advantages compared with other network marketing media, its value impact allows people to believe that Blog can bring about tremendous business value. Therefore, Blog Marketing will become a commercial application which enterprises should not be overlooked.

With many enterprises continuously committing to Blog Marketing studying and application practice, Blog application is established from passive acceptance to active participation. Blog Marketing will break through the traditional means of marketing and will open up a new application field of public relation marketing. So, Blog Marketing will become a symbol of E-commerce marketing mode in the near future.

References

[1] J.Wright, Blog Marketing, McGraw–Hill Education–Europe, 2005

[2] H.Hewitt, Z.S.Yang, and H.Pan, Blog: The forefront positioning of the information revolution, China Railway publisher, 2007

[3] D.Weil, The Corporate Blogging Book, Portfolio Hardcover, 2006

[4] Y.J.Feng, "The concept of Blog Marketing". EB/OL. http://www.marketingman.net/zhuanti/blog/5201.htm , 2005

[5] Y.J.Feng, "The concept of Blog Marketing". EB/OL. http://www.marketingman.net/zhuanti/blog/5202.htm , 2005

[6] S.L.Zhan, "The advantage of Blog marketing and its operational model", Business Times, No.14, May 2006, pp.22-23

[7] C.Zhou, "The new weapon of network marketing", E-commerce World, No.7, July 2006, pp.58-59

[8] R.X.Zhang, and J.Xu, "The applied research of Blog marketing", Productivity Research, No.7, July 2006, pp.261-262

[9] "Blog Marketing: with the power of network", Business operators, No.8, April 2006, pp: 4

[10] CNNIC, "2007 China Blog Marketing research reports", China

Inventory Optimization in Cluster Supply Chains Based on SRCS

Chunling Liu[1] Peilin Guo[2] zhijun Wu[2]

1 School of Electronic & Information, Wuhan University of Science and Engineering, Wuhan, 430073, China
Email: Chunringliu5@yahoo.com.cn

2 School of Economics & Management, Wuhan University of Science and Engineering, Wuhan, 430073, China

Abstract

This paper describes a robust inventory control strategy to find the optimal decision variables to weaken the bullwhip effect in cluster supply chains under environment of demand uncertainty. The model is established to combine real-time optimization and simulation with regarding to the across-chain inventory cooperation. The robust inventory control based on switched system (SRCS) incorporates online inventory decision-making system to overcome model uncertainties and external disturbance. It is proved that this method is more effective and efficient to tune the controller in face of changing environment.

Keywords: inventory management; robust inventory; online switched system; cluster supply chains

1 Introduction

One of the most important aspects affecting the performance of a supply chain is the management of inventories, since the decisions taken in this respect have a significant impact on material flow time, throughput and availability of products. Particularly the current research interesting is the problem of coordination in across-chain inventory control in cluster supply chain. Cluster supply chain (CSC), different from traditional single chain supply chain, is located in the industrial cluster region, with the relation of "supply-client", through the link of formal or informal contract of 'trust and commitment', formed by organizations containing different firms of the same industry such as research organizations, supplier, manufacturer, wholesaler, retailer, and even end users. Cluster supply chain system is made of a couple of paralleled single supply chains in the agglomeration location, not only do all enterprises in one single supply chain cooperate one another internally, but cooperation and coordination exist across different single supply chains externally as well[1].

Nowadays, numerous literatures have been devoted to study of inventory control models in supply chain, and made fruitful progress. However, few of these involve across-chain inventory management which objectively exists among cluster supply chain in many industrial cluster locations[4]. In addition, dynamic uncertain environments and integrated control for complex systems require developing a model of combining inventory management and manufacturing process[2][3]. So, the switched robust control (SRC), one of hybrid controls, is introduced to this paper for establishing the model, therefore, we focus on across-chain inventory in cluster supply chains to analyze its bullwhip effect and approach of implementing it. The following parts are arranged in such way: section 2 briefly describes switched robust control. The section 3 and 4 explore the hybrid optimization and decision approaches. At last, an example is used to show the solving procedure in section 5.

2 Switched Robust Control

Switched Robust Control (SRC) is a robust control method controlling system uncertainty effectively which includes not only parameters uncertainty but also structure uncertainty. The SRC switched controllers considered in this paper are based on the state-space model below:

$$\hat{x}(k+1) = A_i\hat{x}(k) + B_1\omega(k) + B_2\hat{u}(k) \qquad (1)$$

Where, $i \in \{1,2,\cdots,N\}$ is switched rules, $\hat{x}(\Box) \in R^n$ is state vector, $\hat{u}(\Box) \in R^n$ is control vector, $\omega(\Box) \in R^n$ is external disturbance, $A_i \in R^{n\times n}$, $B_1 \in R^n$, $B_2 \in R^n$.

Given the general definition of quadratic cost function

$$J = \sum_{k=0}^{\infty}[\hat{x}^T(k)Q\hat{x}(k) + \hat{u}^T(k)R\hat{u}(k)] \qquad (2)$$

Where, $0 < Q = Q^T \in R^{n\times n}$ and $0 \le R = R^T \in R^{n\times n}$ are respectively state and control weight matrix.

Define the subsidiary output sign as the following

$$z(k) = C\hat{x}(k) + D\hat{u}(k) \qquad (3)$$

Where, $C = \begin{bmatrix} Q^{1/2} & 0 \end{bmatrix}^T$, $D = \begin{bmatrix} 0 & R^{1/2} \end{bmatrix}^T$.

H∞ control problem is to design a controller according to the certain disturbance restraining level γ , so as to satisfy

$$\|z\|_2^2 < \gamma^2 \|\omega\|_2^2 \qquad (4)$$

For the uncertain discrete system (1) and (4), introducing the state feedback control as

$$\hat{u}(k) = K\hat{x}(k) \qquad (5)$$

In above, the SRC controller selects the input $u(k)$ by solving the following optimization problem

$$J = \min\left(\sum_{k=0}^{\infty} z^T(k)z(k) - \gamma^2\|\omega\|_2^2\right) \qquad (6)$$

Theory 1[7] If the system (1) and (4) is stable and satisfying $\|T_{wz}(\zeta)\|_\infty < \eta$, then there existing positive dignity matrix $P_i > 0, \forall i \in I$, matrix W_i and $\forall(i,j) \in I \times I$, $(1 \le m \le s)$ satisfying then the system (1) is quadratic stable, and satisfy condition (6), and accordingly, control law is $\hat{u} = K_i\hat{x}$, $K_i = W_iP_i^{-1}$.

$$\begin{bmatrix} P_i & A_iP_i + B_2W_i & B_1 & 0 \\ P_iA_i^T + W_i^TB_2^T & P_i & 0 & P_iC^T + W_i^TD^T \\ B_1^T & 0 & I & 0 \\ 0 & P_iC + DW_i & 0 & \eta^2 I \end{bmatrix} > 0 \, (7)$$

Theory 2[5] The following stations are equal:

i) Existing difference passive dignity Lyapunov function as formula (8),

$$\begin{aligned} V(k,x_k) &= x^T(k)P(\mu(k))x(k) \\ &= x^T(k)\left(\sum_{i=1}^{N}\mu_i(k)P_i\right)x(k) \end{aligned} \qquad (8)$$

Sustains the system (1) asymptotic stable;

ii) Existing N system matrixes $P_1,\cdots,P_N$ satisfy

$$\begin{bmatrix} P_i & A_i^TP_j \\ P_jA_i & P_j \end{bmatrix} > 0, \forall(i,j) \in I \times I, I = \{1,\cdots,N\} \qquad (9)$$

The Lyapunov function is set by the following formula

$$V(k,x_k) = x^T(k)\left(\sum_{i=1}^{N}\mu_i(k)P_i\right)x(k)$$

iii) existing N subsystem $S_1,\cdots,S_N$ and N matrixes $G_1,\cdots,G_N$,safisfy

$$\begin{bmatrix} G_i + G_i^T - S_i & G_i^TA_i^T \\ A_iG_j & S_j \end{bmatrix} > 0, \forall(i,j) \in I \times I \qquad (10)$$

The Lyapunov function is set by the following formula

$$V(k,x_k) = x^T(k)\left(\sum_{i=1}^{N}\mu_i(k)S_i^{-1}\right)x(k)$$

3 Inventory Control in Cluster Supply Chains

The inventory system of cluster supply chain in this section is composed of two single-chain supply chains which encompass one manufacturer and one retailer (showed in Figure 1) and manufacture character-equal substitutable product.

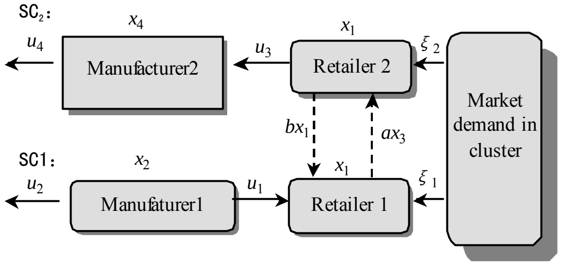


Figure 1 structure of across-chain inventory cooperation in CSC

Suppose x_1, x_2 represent inventory level of retailer and manufacturer in SC1 respectively, x_3, x_4 represent inventory level of retailer and manufacturer in SC2 respectively. Suppose u_1, u_2 represent order of retailer and manufacturer in SC1 respectively, whereas, u_3, u_4 represent order in SC2 respectively. ξ_1, ξ_2 represent the market demand of SC1 and SC2.

In practice, the two supply chains maintain long-term cooperation so as to enlarge the whole demand. When the demand from customers of retailer 1 increases sharply and suddenly, then the retailer 2 may transship inventory to retailer 1 for its emergent need, the supply quantity as $a\hat{x}_3 (0 < a \le 1)$; when the uncertain demand from customer of retailer 2 increases sharply and suddenly, vise visa, the supply quantity as $b\hat{x}_1 (0 < b \le 1)$.

Thus, regarding the inventory state as the state variable, the inventory model may be defined as[4]

$$\boldsymbol{x}(k+1) = \boldsymbol{A}_i \boldsymbol{x}(k) + \boldsymbol{B}_1 \boldsymbol{\xi}(k) + \boldsymbol{B}_2 \boldsymbol{u}(k) \tag{11}$$

Where,

$$A_i = \begin{bmatrix} 1-b & 0 & a & 0 \\ 0 & 1 & 0 & 0 \\ b & 0 & 1-a & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, B_1 = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$B_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \xi(k) = \begin{bmatrix} 0 \\ \xi_1(k) \\ 0 \\ \xi_2(k) \end{bmatrix}$$

The demand from the customer is divided into two parts of being certain and uncertain ones:

$$\xi_i(k) = d_i(k) + \omega_i(k) \tag{12}$$

Thus, in seek of demand disturbance, the inventory variable (state variable) and order (control variable) are all disturbed in the cluster supply chain system (11).

Suppose the standard values of inventory vector, order are $\boldsymbol{x_s}, \boldsymbol{u_s}$, then the error system of cluster supply chain is

$$\hat{\boldsymbol{x}}(k+1) = \boldsymbol{A_i}\hat{x}(k) + \boldsymbol{B_1}\boldsymbol{\omega}(k) + \boldsymbol{B_2}\hat{\boldsymbol{u}}(k) \tag{13}$$

The switched rules are set as

$$\begin{cases} 0 < a \le 1 \text{且} b = 0 \ \omega_{1,k} > 2S_{11}, \omega_{2,k} \le 0, \\ \qquad \text{且} \hat{x}_{11,k} < S_{11}, \hat{x}_{21,k} > S_{21} \\ 0 < b \le 1 \text{且a} = 0 \ \omega_{2,k} > 2S_{21}, \omega_{1,k} \le 0, \\ \qquad \text{且} \hat{x}_{21,k} < S_{21}, \hat{x}_{11,k} > S_{11} \\ a = b = 0 \qquad \text{others} \end{cases} \tag{14}$$

Where, S_{11}, S_{21} are respectively secure inventory level of retailers in SC1 and SC2. The system satisfies $a\Box b = 0$ at any time, that is to say, the transshipment between the retailers may not happen, but they cannot mutually replenish at the same time.

In practice, there is $a = b = 0$, when the routine order channel can meet the real demand. But when terminal demand in supply chain increases sharply so that the demand cannot be satisfied by routine order or if it can be satisfied, the system will face high cost and large risk. With the advantage of cluster supply chain, the system may provide the emergent need by the across-chain inventory coordination to optimize the maximal profit. The system structures are uncertain from (EQ. 13). The uncertainty of structure is not generally disturbance but multi-model switched control. How to determine values of a, b is often related to the subjective factor of decision-maker and practical dynamic demand and inventory when across-chain transshipping occurs because of the complex elements in supply chain management.

The bullwhip effect is described as the proportion of the sum of inventory and order fluctuation to the terminal demand fluctuation, the definition may be showed in [4] namely

$$r_k = [(\hat{\boldsymbol{x}}_{\boldsymbol{k}})^T \boldsymbol{Q}\hat{\boldsymbol{x}}_{\boldsymbol{k}} + (\hat{\boldsymbol{u}}_{\boldsymbol{k}})^T \boldsymbol{R}\hat{\boldsymbol{u}}_{\boldsymbol{k}}]/(\boldsymbol{\omega}_{\boldsymbol{\kappa}})^T \boldsymbol{S}\boldsymbol{\omega}_{\boldsymbol{k}} \tag{15}$$

Where Q, R, S are set symmetrical positively dignity weighted matrix. The parameter r_k describes bullwhip effect in cluster supply chains. The bullwhip effect becomes stronger with increase in r_k, while weaker with decrease in r_k.

For some external disturbance $\omega(k)$, if the controlled output $z(k)$ always maintains small level in the system, then the system with such index present "better" performance. In this case, the controlled output is less influenced by both external disturbance, and the capacity of restraining disturbance in the system appears stronger.

Thereby, the solver satisfying performance index (EQ. 8) surely guarantees minimizing bullwhip effect. The solving process of robust controller may be obtained with given model in section 2. However, how to switch among multiple modes must depend on one online decision-making system which will be introduced in section 4.

4 Online Inventory Control System

Due to complex and ever-changing elements in implementing inventory management in supply chain, the decision procedures are usually complicated. Therefore it is necessary to establish an online inventory decision system. The online system will act as real-time simulator of inventory optimization to provide decision guidance for decision-maker in cluster supply chain. The online inventory optimization and simulation system not only stores a series of switched rules about the system but provides way of setting online switched rules by artificial interacting interface.

This system (shown in Figure 2) consists of optimizer, game module of inventory cooperation, routine replenishment module, replenishment module under inventory cooperation, real system or simulation system. Among them, optimizer encompasses parameters entry, parameter tuning module, demand predicting module, artificial interface and data simulation engine. The game module of inventory cooperation is optimization module that decision-makers of supply chains analyze cooperation decision online. The parameters interpretation in Figure 2 is shown in Table 1.

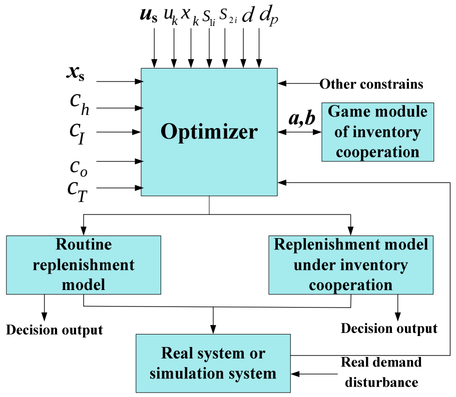

u_s u_k x_k S_{1i} S_{2i} d d_p
x_s
c_h
c_I
c_O
c_T
Optimizer
Other constrains
a,b
Game module of inventory cooperation
Routine replenishment model
Replenishment model under inventory cooperation
Decision output
Decision output
Real system or simulation system
Real demand disturbance

Figure 2 online inventory optimization simulation system

Data simulation engine consists of KIB (Knowledge Interchange Broker), switched rules, data transformation definition shown in Figure 3. In these parts, KIB is the core component which is written by JAVA language. JNI (Java Native Interface) is the main interface that Java language calls for non-Java codes. JMatlink is computing engine of utilizing Matlab's in Java application, Applets and Servlets. JNI in Figure 3 is mainly realized by JMatlink[6].

Table 1 parameter of stock optimization simulation system

Parameters	Interpretation
x_s	standard inventory(n dimension vector)
u_s	standard order(n dimension vector)
x_k	Real inventory fluctuation (n dimension vector)
u_k	Real order fluctuation (n dimension vector)
$S_{1i}(i=1,\cdots,m)$	Safety inventory level of i_{th} tier in SC1
$S_{2i}(i=1,\cdots,m)$	Safety inventory level of i_{th} tier in SC2
d	Real certain demand
d_p	Predicted demand
a	Replenishing ratio providing for SC1 by SC2
b	Replenishing ratio providing for SC2 by SC1
c_h	Average inventory hold cost/unit
c_I	Average inventory cost/unit
c_O	Average order cost/time
c_T	Average transshipment hold cost/unit

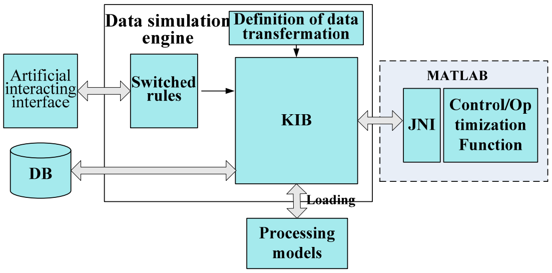

Data simulation engine
Definition of data transformation
Artificial interacting interface
Switched rules
KIB
MATLAB
JNI
Control/Optimization Function
DB
Loading
Processing models

Figure 3 simulation structure of inventory optimization in CSC

Two classical application cases by JMatlink are shown in the following:

i) Return Java image from figure image windows of Matlab by engGetFigure(), for example,

```
Image im;
JMatLink engine = new JMatLink(Inv32fig_1);
engine.engOpen(Inv32_1);
engine.engEvalString("surf(peaks)");
im = engine.engGetFigure(1, 300, 250);
engine.engClose(Inv32_1);
```

ii) Add two new methods to restore Matlab figure as Java image, for example,

Image engGetFigure(long epI, int figure, int dx, int dy)
Image engGetFigure(int figure, int dx, int dy)

5 Simulation Test Case

Two-echelon cluster supply chains are taken as instance in Figure 1 and the switched robust inventory model (EQ. 1 and EQ.4) is given in this section.

Suppose the standard system of cluster supply chains are

$$x_{1s} = [450, 500, 700, 760]'unit,$$

$$u_{1s} = [500, 550, 720, 840]'unit$$

And suppose the initial values of order error are zero, but the initial values of inventory error are

$$\hat{x}_0 = [0, 10, 60, 40]'unit$$

Suppose the system face demand disturbance at k=0 ($\omega_1 > 2S_{11}$). The retailer in SC1 has larger need at period of k =0 (namely, $\omega_1 > 2S_{11}$). The other supply chain will provide emergent inventory supply for the retailer in SC1 by contract when meeting the condition of inventory transshipment. Suppose the supply ratio is a =0.8.

The subsystem 1 is set as system without across-chain inventory cooperation, The subsystem 2 is set as system with across-chain inventory cooperation (the retailer of SC2 supplies inventory for retailer of SC1), the according parameters are

$$Q = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$S = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$A_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 1 & 0 & 0.8 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The system is switched to mode 2 (subsystem 2 is determined) in periods of k=1、k=16, while mode 1 in other periods by results of online decision-making module. Set $\gamma = 3$, according to the condition that the system has certain performance of restraining disturbance and must satisfy robust stability (EQ. 8, 10). The according solves are

$$P_1 = \begin{bmatrix} 1.3654 & -0.4509 & 0 & 0 \\ -0.4509 & 1.7756 & 0 & 0 \\ 0 & 0 & 1.3654 & -0.4509 \\ 0 & 0 & -0.4509 & 1.7756 \end{bmatrix},$$

$$W_1 = \begin{bmatrix} -1.2927 & 0.3727 & 0 & 0 \\ -0.7126 & -0.9839 & 0 & 0 \\ 0 & 0 & -1.2927 & 0.3727 \\ 0 & 0 & -0.7126 & -0.9839 \end{bmatrix}$$

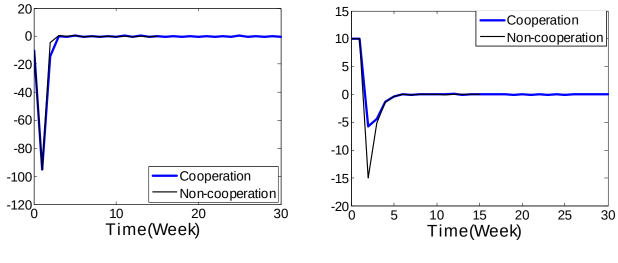
$$P_2 = \begin{bmatrix} 1.9778 & -0.1873 & -1.0157 & 0.0024 \\ -0.1873 & 1.9704 & -0.4724 & 0.0174 \\ -1.0157 & -0.4724 & 2.3412 & -0.2728 \\ 0.0024 & 0.0174 & -0.2728 & 1.7256 \end{bmatrix},$$

$$W_2 = \begin{bmatrix} -0.8879 & 0.5715 & -0.7915 & 0.0066 \\ -0.5227 & -0.9225 & -0.3078 & -0.0729 \\ -0.1702 & -0.0797 & -0.5501 & 0.2427 \\ -0.1287 & -0.0694 & -0.2247 & -1.0361 \end{bmatrix}$$

In supply chain, the inventory fluctuation and order fluctuation at every tier can only maintain stable within a small scope in seek of demand disturbance. The simulation and computing are carried out by emulating online decision in this section.

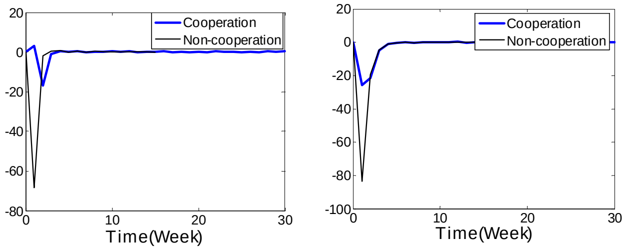
It is shown from Figure 4~6 that the system may tend to be stable within a small scope by exerting H∞ control, while the effect is better when existing across-chain inventory cooperation. One evident reason for this result is that the order fluctuation is largely dwindled by across-chain inventory cooperation, thus the bullwhip effect can be better weakened.

Additionally, it is proved by simulation in this section that the online decision-making system of cluster supply chain can optimize the whole system in multiple periods, and then take on quicker response to complex market.



(a) retailer (b)manufacturer

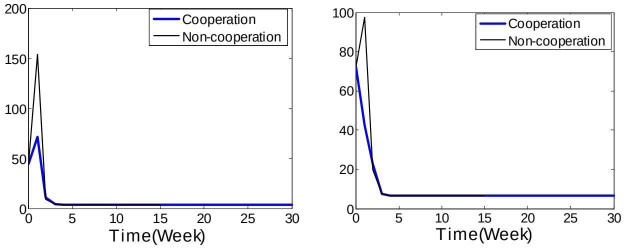
Figure 5 Variation tendency of inventory in SC1



(a) Retailer

(b) Manufacturer

Figure 6 Variation tendency of order in SC2



(a) Supply chain 1

(b) Supply chain 2

Figure 7 Variation tendency of bullwhip effect

Acknowledgements

This paper is supported by Hubei EFC (2007Y071, B200717001).

References

[1] J. Z. Li. Inventory Management in Cluster Supply Chain. Press of Chinese Economy, 2006

[2] E. Lopez, B. Ydstie, I. E. Grossmann. A model predictive control strategy for supply chain optimization. Computers & Chemical Engineering, 27(2003), PP.1201-1218

[3] H. S. Sarjoughian, D. P. Huang. Hybrid Discrete Event Simulation With Model Predictive Control For Semiconductor Supply-chain Manufacturing. Proceedings of the 2005 Winter Simulation Conference, PP.256-265

[4] C. L. Liu, J. Z. Li. Research on H∞ Control of Bullwhip Effect in Cluster Supply Chains Based on Cooperation between Two single Chains. Chinese Management Science, 2007, 15(1), PP. 41-46

[5] P. Daafouz, Riedinger, and C. Lung. Stability analysis and control synthesis for switched systems: a switched Lyapunov function approach. IEEE Trans. Automat. Control, 2002, 47(11), pp.1883-1887

[6] S. Muller. http://www.held-mueller.de/JMatLink/

[7] L. Yu. Robust control——linear matrix inequality method. Press of TsingHua University, 2002

Research on RFID Based E-ticket in China Railway

Liyi Zhang[1] Yongwang Zhao[2]

1 School of Information Management, Wuhan University Wuhan, Hubei 430072, China
2 Email: 1 lyzhang@whu.edu.cn ;2 yinzhao0509@163.com

Abstract

Now, e-ticket is widely used in transportation industry in China, such as airline industry, bus industry, etc. Moreover, there are sufficient advantages to use RFID technology in e-ticket. Nevertheless, traditional paper-ticket is still used in China railway, which is a most important transportation industry in China. In this paper, a kind of RFID based e-ticket technology is brought in China railway industry. Also, a new ticket system model is redesigned to satisfy the new e-ticket. The new e-ticket in the railway will greatly reduce labor and time costs for the station and facilitate China travelers who usually travel by train.

Keywords: E-ticket, Electronic Ticket, RFID, Radio Frequency Identification, China Railway

1 Introduction

Information technology is gradually changing many industries of the world, especially in the transportation industry. E-ticket (electronic ticket) [1] is becoming a more and more popular choice for many travelers. Using e-ticket, a traveler can no longer use and save traditional paper ticket. He just needs to go to the station to check in with e-ticket and his personal identification information. This will greatly facilitate travelers and save their lots of time and procedures.

E-ticket is a kind of ticket which is a paperless document used for ticketing passengers[1]. Now, it is widely used in airline industry and bus industry in China. It reduces a lot of costs for both the station and passengers. Nevertheless, its use is very limited in China railway, which is also an important transportation industry.

Although there have been several proposals of e-ticket in railway. Cuanfu Meng and Fuzhang Wang [2] analyzed the feasibility of RFID based e-ticket and proposed some difficulties in developing e-ticket. Yong Lv[3] presented some proposals for integrating e-commerce with China railway. Neither of them went on to propose an e-ticket system model and some actual operating methods.

In this paper, e-ticket application model in airline industry is analyzed. Because there are a sufficient number of unique characteristics in China railway, only some of its workflows and merits are adaptive. According to the actual situation of China railway, a new kind of e-ticket is designed. At the same time, a new ticket system is presented.

2 Background

In this section, we introduce e-ticket and RFID technology. Then, we discuss the current e-ticket application in transportation industry.

2.1 E-ticket

E-ticket is a paperless electronic document used for ticketing passengers, particularly in the commercial airline industry [1]. In fact, it is an electronic image of common paper ticket. [2] Usually, it at least contains all the information which is included in old paper ticket. Some other special information is included additionally, such as personal identification, ticket identification codes, etc.

There are mainly two kinds of e-ticket now. One is just a series of personal identification information and some sequence numbers or letters got by accessing to the Internet or telephoning to the related company to reserve a seat, just like most e-ticket used in China airline industry now. Another is a card, or a kind of coin, which is used to pay for the bus or underground. This kind is often RFID based e-ticket. When using it, data stored in the e-ticket can be read out by the reader of ticket system easily with the absence of contact. This means that it can be easily used for ticketing mass passengers.

2.2 RFID technology

RFID stands for Radio-Frequency Identification. It is a suite of technologies that include “tags” which get applied to items that need to be tracked, “readers” or “interrogators” that scan the tags nearby for their data, and a series of integration technologies that link the readers back to central databases and systems that track the data being scanned [4, 6].

An RFID system typically consists of a radio-enabled device that communicates with or "interrogates" a tag or label, which is embedded with a single chip processor and an antenna. The "interrogator" or RFID reader may be a fixed antenna or a portable one. RFID systems can be largely automated, reducing the need for manual scanning. [4]

An RFID tag is based on a chip or integrated circuit (IC) usually composed of silicon. A tag insert or inlay is the IC attached to an antenna, which is usually printed or etched on a substrate material. The tag itself is the inlay plus its encapsulated protective packaging. The packaging can be flexible or stiff, as the application warrants. [4, 5]

In the back end of the system, a host computer stores all collected data within a database. Since RFID tags can also carry data, tags can serve as data transfer agents, synchronizing disparate information systems. The tags can either be Read Only (RO) or Read/Write (R/W) capable. [5, 6]

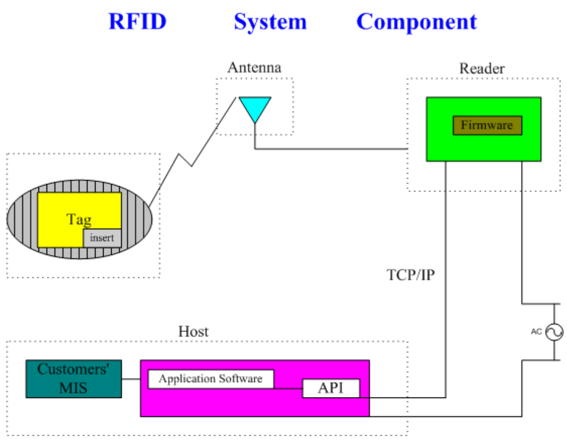


Figure 1 RFID system component

2.3 Current application of e-ticket

E-ticket is firstly brought in airline industry. In 1995, it firstly appeared in CO Airline Inc. One year later, e-ticket was widely promoted among airline companies. The concept of e-ticket was not known to global world until 1998. After two years, passengers began to reserve it on the Internet directly. In 2001, the development of e-ticket was accelerated greatly. Now, e-ticket has been widely used in UA (United Airlines) [7].

Today, there are two kinds of e-tickets in China airline industry. One is e-ticket system built by China Southern Airlines, China Eastern Airlines, etc. Another is BSP (Billing and Settlement Plan) e-ticket used by Air China Co. Ltd and Hainan Airlines Co. Ltd. In China, the first really significant e-ticket was produced in Southern Airlines on March 28th, 2000. In June, 2005, China Southern Airlines achieved 350 billion e-ticket sales. At the end of November 2005, e-ticket in Air China had covered all of 62 domestic cities, including Hong Kong, and also covered some international airlines, such as Seoul (Korean), America, Canada etc. A great improvement is that Hainan Airlines Co. Ltd has achieved “Check in without ticket” on February, 2006. IAAT has stopped providing paper ticket for domestic Airline-Ticket Agencies on October 16th, 2006 and offered 100% e-ticket in 2007. [7, 8]

At the same time, e-ticket is gradually accepted by bus industry. It is reported that RFID based IC card (a kind of e-ticket) are used in nearly above 90% domestic cities' bus industry. Also e-ticket is used in underground in many cities in China.

According to the information above motioned, e-ticket has been more and more important, and widely used in China transportation. Nevertheless, it is a pity that e-ticket is not widely used in railway industry, which is the most important vehicle in China.

3 Building a New E-Ticket System Model

In this section, we will analyze the general e-ticket system model in airline industry, on which we present our new e-ticket system based. Our new e-ticket system will inherits most of its advantages. Meanwhile, we will redesign some traditional procedures according to the specific characteristics of the railway industry.

3.1 General e-ticket system model in airline industry

In airline industry, common e-ticket system often consists of two phases:

1) Booking phase, as the Figure 2 depicted.

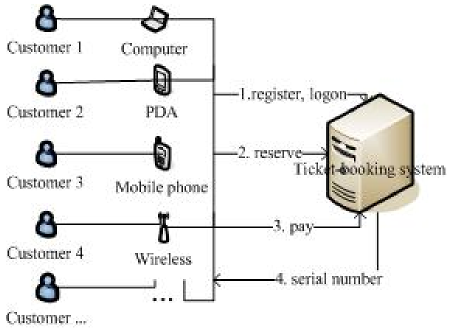


Figure 2 Booking phase in airline industry

In phase 1), passengers access to airline ticket-booking system provided by airlines via Internet, mobile phone, PDA, etc. Usually, they have to register in the ticket-booking system and provide a series of personal identification information for the system. Then, they can book the ticket. At this moment, the system asks customers to pay for it. Passengers can pay via personal Internet Banking, digital cash, etc. The system will provide customers with a unique serial number and a routine paper. This paper usually contains all of your trip information. If you need this paper, you can download and print it.

2) Check-in phase, as the Figure 3 depicted.

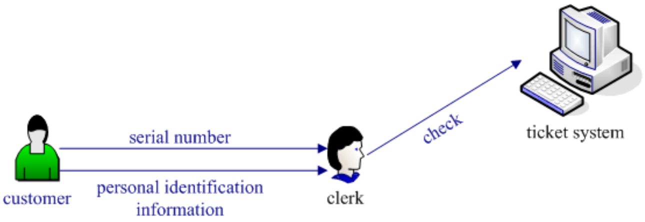


Figure 3 Check-in process

In phase 2), passengers go to the airport to check in with the serial number they have got. Meanwhile, they should provide their personal identification information such as ID card, or other valid documents. Then, the clerk can input this information to validate their booking record. If there are no mistakes in the process, you can check in and continue you trip.

3.2 Unique characteristics of china railway

Railway is the most important and complicated system of China transportation industry. Although current e-ticket systems play an important role in airline industry and bus industry, there are a sufficient number of differences between China railway and airlines industry. It has its own specific characteristics:

1) There are the biggest numbers of passengers using China railway.

2) These passengers are greatly influenced by time and area.

3) These passengers are always middle and long distance travelers.

4) The structures of passengers are often very complicated.

5) Passengers often take lots of baggage and its volume may be very big.

6) There are often transient traffic jams on check-in process in the railway station.

7) There are often transient traffic jams on check-out process in the big railway station.

3.3 Traditional ticket system in china railway

Traditional paper tickets have several special characteristics:

1) There are many kinds of tickets in railway because of its complicated ranks.

2) There is a great number of information in the ticket, such as the number of the train, price, time, beginning station, end station, etc.

3) The ticket is a kind of PI and can be used to apply for reimbursement.

4) The process of ticket sales is very complicated, for example, it allows customers to reserve or book in advance in stated days, and so on.

5) The immature technology often brings fake tickets and reselling tickets phenomenon. This greatly damages the interest of the country and customers.

Based on these characteristics, Figure 4 shows us the traditional tickets system.

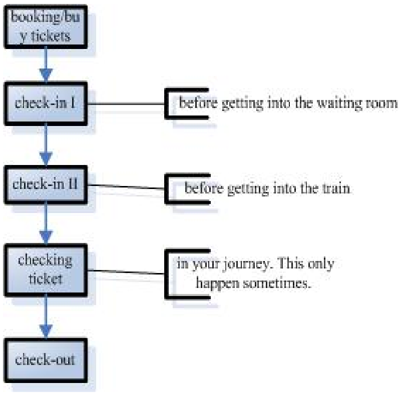

booking/buy tickets
check-in I
before getting into the waiting room
check-in II
before getting into the train
checking ticket
in your journey. This only happen sometimes.
check-out

Figure 4 Traditional tickets system in China railway

3.4 Building a new e-ticket system

The term “e-ticket system” we use has a wider meaning. The whole ticket-flow which begins with buying ticket and ends with reusing or discarding it is defined as an e-ticket system. Our new e-ticket system consists of 4 parts: booking/buying e-ticket, writing the related information into the e-ticket, check-in, and check-out.

3.4.1 New E-ticket

(1) Traditional paper ticket in railway

Traditional paper ticket in railway may include a lot of information. It generally contains the following information:

1) serial number of the ticket

2) date and time

3) beginning station and end station

4) the number of the carriage and seat

5) price

6) the number of train

7) ticket station

8) type of the train

9) period of validity

10) barcode of the ticket

11) visitors notice (often on the background)

We can find that there is so much information in a ticket. It means that when we use an e-ticket, this basic information should be included. Fortunately, this information is often stored as a record in the ticket system. It gives us an opportunity that transfer part or whole of this record into our new e-ticket.

(2) E-ticket

Our e-ticket will be an RFID based card, like a non-contact IC card in bus industry. If a customer confirms his ticket information, all of the traditional ticket information and some special e-ticket information will be written into his e-ticket. At the same time, customers’ some identification information should be scanned into the e-ticket accordingly. When customers check out, above information stored in the e-ticket can be read out and checked easily.

Our e-ticket can be reused many times. When one travel is finished, the related travel record can be reserved before your next trip. Then before your next trip, you can provide the clerk in the station the same card, she will rescan new travel information into it. The new information will replace your old travel information.

3.4.2 Buying e-ticket

Considering that it is so important to most of Chinese people and many of them don’t use computer, PDA and such kind of tools, we provide two buying ways: online and offline.

(1) Online buying

Our online buying way resembles to the airline industry. But the difference is that we abolish the booking way. Because of the lack of tickets or other reasons, people are forced to reserve the ticket online or by telephone. Passengers can only reserve a seat in the ticket system open to public, but the station can’t make sure that the seat is theirs. Now we suggest that the station open most of seats to public and allow people to buy through Internet or other way. The station will not acclaim that the seat you reserved is yours until you have finished payment. Of course, there should be a deadline. If it doesn’t receive somebody’s payment before deadline, the station will make this seat available to public. Here, the deadline is 12 hours. This means that the online buying will be closed 12 hours before the train sets out.

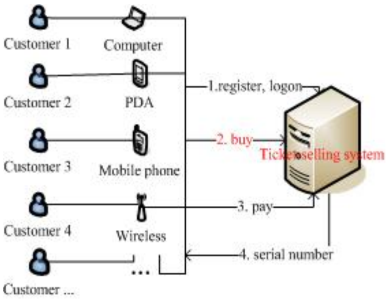


Figure 5 Online buying process

(2) Offline buying

Offline buying is similar to the traditional way. Passengers have to go to the station to buy the e-ticket. But when customers finish buying ticket, they will get an e-ticket card rather than a traditional paper ticket. Likewise, they can re-use their e-tickets in next trip as the online buyers.

(3) Writing the related information into the e-ticket

Both online and offline buyers need to get an e-ticket before check-in. When they use the e-ticket for the first time, they need to ask the clerks in the station to scan their personal identification and trip information into the e-ticket. Then these records will be written into the e-ticket automatically.

(4) Check-in

In traditional way, check-in may at least include two phase. First, before travelers coming into the waiting-room, the clerks often need to check their tickets. Then, before travelers getting into the train, the clerks check again. This process often needs a lot of clerks and time to be finished.

When people use e-ticket, they only need to be checked once before they get into the waiting-room. Because the check-in barrier machine we used only open until customers swiping their e-ticket. This measure will confirm the validity of travelers who has got into the waiting-room and can be used a final check-in in the beginning station. It can save lots of time and human cost.

(5) Check-out

In traditional way, when people arrive at the termination, they need to pass the check-out process. The terminal station also needs to send some clerks to do this business. Now, we can use the same kind of barrier machine as the check-in phase and let it do the check-out business automatically. When people check out, they only need to swipe their e-tickets. The barrier machine will read the data stored in the e-ticket out and check whether it is an effective card. If it is effective, the barrier machine will open and let the passengers pass through.

4 Discussion

E-ticket is widely used in China airline industry and bus industry now. It brings travelers with a great number of convenience and speediness. It also has saved a lot of labor and time costs for the related companies. Practice has proved that e-ticket used in China transportation industry is successful, efficient, and promising.

Railway industry is the most important transportation industry. Most of people usually use railway as their first travel choice. Therefore, bringing e-ticket into China railway is a significant thing. We do this in this paper. We design a kind of e-ticket and design a new workflow based on this kind of e-ticket.

Our new e-ticket is an RFID based card. The RFID technology has two unique advantages:

(1) It is non-contact.

(2) It can be re-read and re-written for many times.

These characteristics insure it can easily get over the traditional paper tickets' difficulties in re-use. Using this new kind of ticket require the old buying ticket, check-in, check-out process to be changed according with it consequently. So, we rebuild the buying ticket, check-in and check-out process. The new process can bring us with much convenience:

1) This ticket can be re-used. If you want to travel next time, you can use the same card. The only thing you need to do is that you go to the station to ask the clerks to update its stored information. In old way, you have to buy paper ticket when you plan to travel. When you want to travel next time, the station has to produce another paper tickets. We know, this process will waste many paper materials and other expenditures.

2) In old way, a traveler needs to check in at least twice before he gets into the train. When you arrive at the terminal station, you still have to pass the check-out process. The first check-in process occurs at his getting into waiting-room while the second occurs before his getting into the train. Maybe, you have to face another once or twice checking ticket in your trip. Our new check-in and check-out way only needs the traveler to be checked only once and the process is automatic. This can help the station reduce many check-in and check-out clerks and time.

3) Our new way greatly facilitates travelers' buying ticket. In the new process, the public can easily access to the ticket system and buy it through kinds of ways rather than that travelers can only order the ticket but can not buy it directly in traditional way. This innovation will greatly facilitate the travelers and confirm their confidence in railway.

Nevertheless, there are some problems at the same time. The ticket system facing the public directly requires the system must be safe and stable. It must provide perfect measures to prevent the system from being attacked and be in working order. Another important thing is the system must provide enough measures to protect the customers' online payment security. Also, there are many other complex problems to be resolved, such as users' checking the stored information, lost of card, etc. We will continue going to research on these problems.

5 Conclusion

E-ticket was first introduced into the airline industry. At the same time, another kind of e-ticket is widely used in bus and underground industry. All of these e-tickets have saved a lot of time and other costs and brought us with efficiency and profits.

So, we bring e-ticket into China railway industry. A new kind of adaptive e-ticket is designed in this paper. Accordingly, the old workflow process in railway has to be innovated to fit the new e-ticket and we do that.

Although the e-ticket is not used in China railway now, we believe its prospect is promising in the future. Maybe, several years later, every Chinese traveler can enjoy the efficiency and speed brought by e-ticket and make current troubled traveling by train phenomenon be a period of history.

References

[1] E-ticket-Enjoy working enjoy your business travel, http://english.yoee.com/electron/ecectron.asp?uid=zmuuuggsczabezpwgguzgpws

[2] Cuanfu Meng and Fuzhang Wang, "RFID Based E-ticket Feasibility Analysis in China Railway", Comprehensive Transportation, Apr 2006, pp.70~72

[3] Yong Lv, "Research on application of E-ticket in China Railway E-business", Railway Computer Application, Vol.14, No. 12, Dec 2005, pp.40~42

[4] Klaus Finkenzeller, Grundlagen und praktische Anwendungen induktiver Funkanlagen, Transponder und erweiterte Auflage 3, aktualisierte und erweiterte Auflage. Beijing, CA: Publishing House of Electronics Industry, 2006

[5] What is RFID? http://www.tech-faq.com/rfid.shtml

[6] Apparatus system and method of using RFID systems to help blind and visually-impaired individuals. http://www.freepatentsonline.com/70176785.html.20070802

[7] Wenyuan Qu, "E-ticket in Air Transportation", in: China Civil Aviation, Vol.25, Jan 2003, pp.54

[8] Dingquan Xu, "E-ticket Development Status Quo Analysis in China", China Collective Economy, Sep 2007, pp.49~50

Collaborative Purchasing Service in Manufacturing Grid Based on Multi-Agent *

Bing Tang   Zude Zhou   Quan Liu

School of Information Engineering, Wuhan University of Technology, Wuhan, Hubei 430070, P.R. China
Email: tangbing@whut.edu.cn

Abstract

As the development of information technology and advanced manufacturing technology, manufacturing Grid has become a reality. Collaborative resource management is an important content of manufacturing Grid. After analyzed the advantages of collaborative purchasing for enterprises, as a new application of Multi-Agent technology, the paper proposed the model of Multi-Agent-enable collaborative purchasing alliance and an automated negotiation framework in collaborative purchasing process. Finally, the paper implements a prototype system for collaborative purchasing based on Multi-Agent and then discusses the deployment of collaborative purchasing Grid services on our Manufacturing Enterprise Grid Support Platform (MEGridSP).

Keywords: Collaborative Purchasing, Manufacturing Grid, Grid Computing, Negotiation, Multi-Agent, MEGridSP

1 Introduction

It is undoubtedly that the research and development of Grid technology have provided an excellent foundation for information integration in manufacturing industry. As the next generation manufacturing model, manufacturing Grid aims at sharing all kinds of manufacturing resources [1]. Enterprise resources collaborative management in manufacturing Grid environment is a popular issue. The principle and practice of supply chain management (SCM) and collaborative commerce provides important experience for resource collaboration in manufacturing Grid [2].

The competitive market request the collaboration of small and medium-sized enterprises (SMEs), where that enterprise collaboration has very practical advantages and it is also an international tendency. As a new model, collaborative purchasing (CP) for SMEs has attracted more researchers' attention. The target of purchasing alliance is to reduce cost of material purchasing for its members, and the alliance pattern has been carried out in the world and there are satisfied results obtained [3].

Multi-Agent technology has been widely used in information network fields and distributed system fields. The communications between Agents make sure the total system becoming more knowledgeable and more functionally powerful. Multi-Agent-enabled SCM and automated negotiation in electronic commerce are currently hot spots. Collaborative purchasing based on Multi-Agent is an attempt and also an innovation [4].

The rest of the paper is organized as follows. Section 2 discusses the probability and advantages of collaborative purchasing for enterprises, especially SMEs. Section 3 briefly introduces a grid support platform named MEGridSP which is used in collaborative purchasing system. Section 4 presents a new negotiation model of collaborative purchasing based on Multi-Agent. Section 5 describes how to deploy a collaborative purchasing Grid service in manufacturing enterprise Grid. The final section offers concluding remarks.

* This paper is supported by International Cooperation Project of the Ministry of Science and Technology, China (Grant No. 2006DFA73180) and Major International Cooperation and Exchange Project of National Natural Science Foundation of China (Grant No. 50620130441).

2 Collaborative Purchasing for Smes

Theoretical background

Gartner Group gives a definition to collaborative commerce in 1999 [5]. Collaborative commerce or c-commerce involves the collaborative, electronically enabled business interactions among an enterprise's internal personnel, business partners and customers throughout a trading community. The trading community can be an industry, an industry segment, or a supply chain or supply chain segment. Supply chain management integrates planning and balances supply and demand across the entire supply chain. It ties suppliers and customers together in one concurrent business process that focuses on the ultimate customer. Collaborative commerce aims at building collaboration with customers and other enterprises, and also building collaboration inside the enterprise.

The theoretical research results of supply chain management and collaborative commerce aware us that enterprise should establish strategic fellowship to other enterprises in the same industry or in the same district to fit modern competition. There is a tight relation between supply chain optimization and collaborative commerce. Collaborative purchasing fulfils the requirement of collaborative commerce, and it also elevates the purchase efficiency and the capability of quick response through sharing resources and united purchasing activities.

Collaborative purchasing for SMEs

In order to survive in the competitive market economy, SMEs should produce commodities with high quality and low price according to the market requirements, which means SMEs need high-techniques and low material cost to make profit and seek development. Generally speaking, material cost covers 50% of total sale in manufacturing industry. With a low output quantity and limited manufacturing capability, SMEs keep in a disproportional station with large supplier in the purchasing process.

There are some problems of purchasing alone for SMEs as follows [3]:

1) With a low quantity discount. With small size and limited fund, SMEs are unable to purchase batch material, and can not enjoy a high quantity discount.

2) With a low capability of bargaining. Due to low purchasing quantity, when bargaining with large supplier, SMEs often lack of the right to negotiate a price that incline to themselves with large supplier. They could not determinate the price.

3) With a high purchasing management cost and high risk. SMEs should be responsible for supplier selection, supplier evaluation, negotiation with supplier, product tryout, etc, and they must deal with the inferior product provided by supplier. Hence, SMEs are confronted with trade risks.

Expecting to stay in an equivalent status with large supplier, SMEs must be united together to form an enterprise alliance, and congregate the purchase requirements to generate a combinational order, and bargain with supplier.

There are some advantages of collaborative purchasing for SMEs. Collaborative purchasing enables saving martial costs and transport costs for SMEs and the request of purchasing alliance can be responded quickly. Alliance enterprises may enforce the strategic fellowship through communication and collaboration.

Collaborative purchasing alliance cases

Because similar enterprises have similar demands, they may establish a purchasing website together to sharing supplier resources. This is the innovative website model for collaborative purchasing. The website often serves inside the grouped enterprise, or inside special industry, or cross some companies. It is reported that GE, FORD, Renault and NISSAN had established a collaborative purchasing website (Covisint), and business with 7500 billion US dollar were doing on the website each year. In China's steel industry, ShouGang Group, WuGang Group, AnGang Stell Company and other steel companies had founded a China united steel alliance and collaborative purchasing website (custeel), and provide bid and purchase services for mineral on the website. Shanghai Bus Group, Zhengzhou YuTong, Xiamen KingLong, etc., have set up an auto components business platform (APEP) to serves for the members in their purchasing alliance. Survey of the three alliance cases reveals the practical meaning of collaborative purchasing alliance.

Key issues of collaborative purchasing

Here are some important components and functions that a collaborative purchasing system must contain as follows:

1) Common network and information platform.

2) Material category determination and management.

3) Collaborative purchasing process model.

4) Negotiation strategy in collaborative purchasing.

5) Order management and accounting management.

3 Megridsp—Manufacturing Enterprise Grid Support Platform

Grid technology

The Grid is a name that was first coined in the mid-1990s to denote a proposed distributed computing infrastructure for advanced science and engineering projects. Grid technology is a growing information technology, where that the main purpose of Grid is to realize resource sharing and collaborative working in virtual network environment, which can eliminate the information island and resource island. As explained by Foster and Kesselman, Grid should enable 'coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organization' [6, 7].

With the first-generation Grid involving 'meta-computing' and the second-generation Grid focused on middleware and communication protocols, it is now claimed that the third-generation Grid is combining SOA (Service-Oriented Architecture) concepts and Web Service technologies to create the Open Grid Services Architecture (OGSA). OGSA is an important propositional standard by Global Grid Forum (GGF). It is an important Grid architecture after the early layered Grid architecture. The object of OGSA is to realize resource management and service sharing cross heterogeneous platform. OGSA defines Grid Service, which is a special kind of Web Service, providing service discovery, dynamically service creation, lifecycle management, notification, and so on.

The application of Grid technology has widely spread in industry fields, and the applications are moving form high performance computing fields to engineering fields, where that some new application models of Grid are emerging, such as data Grid, semantic Grid, VOD Grid, education Grid, computational Grid for bioinformatics and finance analysis, manufacturing Grid for collaborative design, production, business and management.

Manufacturing Grid

Manufacturing industry is affected by the tendency of globalization, network manufacturing and virtual manufacturing. Manufacturing enterprises are making their effort to seek solutions to survive in the global market-competition. Under continuous, variable and totally unpredictable competition, manufacturing enterprises should solve the TQCSEF problems of the new products in order to enhance the competitive capacity. But the traditional manufacturing model that only depends on the internal manufacture resources is difficult to solve the problems. Hence, in order to solve the TQCSEF problems, an enterprise should fully uses the exterior manufacturing resources and realizes the collaborative work with other enterprises under the assistance of Grid technology [1, 2].

Under these conditions, there appears the next generation manufacturing model, and that is Manufacturing Grid (MGrid). MGrid is proposed to meet the practical demands in the manufacturing industry and is a new technology enabling broad geographical distributed of all sorts of manufacturing resources using grid technology. With the MGrid, we can realize common sharing of manufacturing resources including equipment resources, material resources, applicable software systems, etc., and can also realize the collaborative design and manufacture of the same product in different places and enhance the competitive capacity of enterprises by shortening the exploring and manufacturing period of a product, and minimizing the entire cost as well. MGrid achieves the recombination of manufacturing process in form of VE (Virtual Enterprise), so as to resolve the TQCSEF problems.

About MEGridSP

Although the framework of manufacturing Grid has proposed for several years, but there is neither a standard platform, nor a commercial application specification. Supported by NSFC and Guangdong province, we have done some work to implement a manufacturing Grid follow the general manufacturing Grid architecture, named it as MEGridSP, which is a manufacturing enterprise Grid support platform, which solves the fundamental problems of Grid environment deployment, and it is expected to serve for SMEs in Shunde District of Foshan City in Guangdong province. In this paper, we just give a simple introduction, the detailed information about MEGridSP can be found in [8]. It consists of some modules, including job manager, service container, information service, data manager, domain manager, security manager, programming model, resource packaging tools, service publishing tools, etc. MEGridSP is a standard, scalable and extensible platform, which allows fast setup and deploying a Grid cross manufacturing enterprises. MEGridSP provides a network and information platform for collaborative purchasing as we have discussed in Section 2.

A GridASP framework has proposed in [9], which realizes Grid-enable application service providers (ASP) in order to realize Grid utility computing. The Platform Corporation also provides University Campus Grid solution based on ASP Portal, which has been practiced in several universities in China. Although ASP is not a newly technology, but the conception and model of ASP can be adopted in the Web portal under Grid environment.

An important feature of MEGridSP is that ASP-enabled Grid portal provides friendly shared platform which bridges the end user and resource service provider. Another feature is that plug-in-enabled mechanism makes sure Grid services can be configured dynamically.

4 Negotiation in Collaborative Purchasing Based on Multi-Agent

MEGridSP and manufacturing enterprises

MEGridSP provides a common network and information platform as we have explained in Section 2. The relationship between MEGridSP, Supplier Enterprise, Purchaser Enterprise and Grid Portal is shown as Figure 1, in which each enterprise deploys their Grid service, regarding MEGridSP platform as the infrastructure.

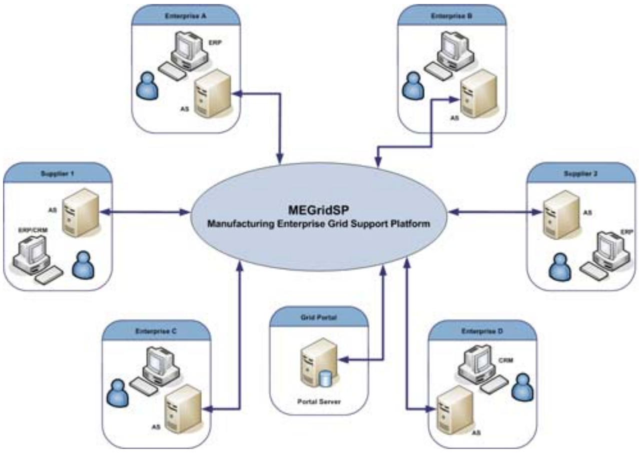


Figure 1 The relationship between MEGridSP and manufacturing enterprises

The application services (AS) of each enterprise are published on the platform and registered in Grid Information Service (GIS). The application services (AS) server is connected to the internal system of enterprise, such as ERP, SCM, and CRM system. To MEGridSP, AS is just a plug-in part as mentioned in Section 3. Each enterprise is regarded as an AS provider, and AS from different enterprises can be shared and consumed by each other complying with special policies and security protocols [9, 10, 11].

Agent communications and purchasing system model

The negotiation strategy in traditional electronic commerce process based on Multi-Agent is a static strategy. We proposed a market-driven negotiation strategy. The Agents can adjust itself according to the market activities, and select optimal strategy, and make the optimal decision instead of purchaser automatically. This method reduces the purchaser's intervention, and gives a satisfied negotiation results to purchaser. Only in the exceptional emergent situation, the Agent may seek help from purchaser [12, 13].

Altogether 9 Agents are used in the collaborative purchasing Multi-Agent system. These are Purchaser

Agent, Coordinator Agent, Control Agent, Transaction Agent, Security Agent, Contract Agent, Quote Agent, Evaluation Agent, and Supplier Agent. We only concentrate on the three Agents: Purchaser Agent (PA), Coordinator Agent (CA), and Supplier Agent (SA) in this paper.

We model the collaborative purchasing activity as the interaction between PA, CA, SA and MEGridSP. Figure 2 shows a simple communication model. The automated negotiation process lies in the interaction between CA and SA. The purchasing requirements are abstracted as the knowledge of PA. For example, purchaser may have a preference for the factor of time, or the factor of cost, and the preference is just a kind of knowledge. Coordinator is an intelligent software robot; it acts as an agency, which inclines to satisfy the benefit of purchasers, not the benefit of supplier.

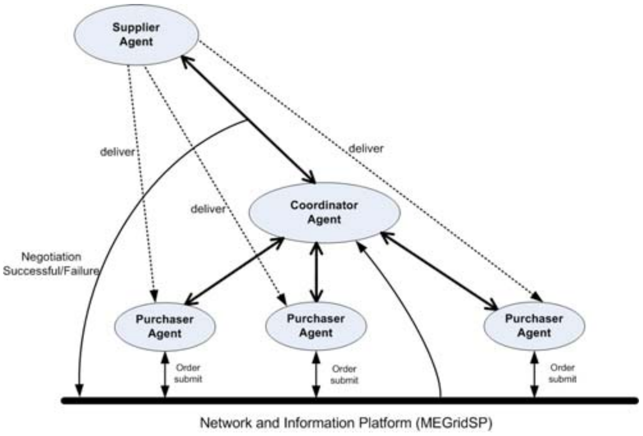


Figure 2 Communication model of collaborative purchasing system

Collaborative purchasing steps based on Multi-Agent

The sequence diagram in Figure 3 demonstrates the steps of collaborative purchasing process based on Agents. The steps are described in detail as follows:

Step 1. Purchaser submits order and defines its purchase requirements, and these requirements are extracted and learned as knowledge and strategy by PA.

Step 2. CA collects and consolidates orders from PA, and PA represents the purchaser with the willing to become a member of collaborative purchasing alliance.

Step 3. The orders are evaluated by CA, and in this stage there may be orders needed to be adjusted, we called it as dirty orders. Dirty order often means product conflict or exceptional emergent situation. The adjusting suggestion about dirty order is then sent to relate PA by CA. Generally speaking, the probability of dirty order is tiny.

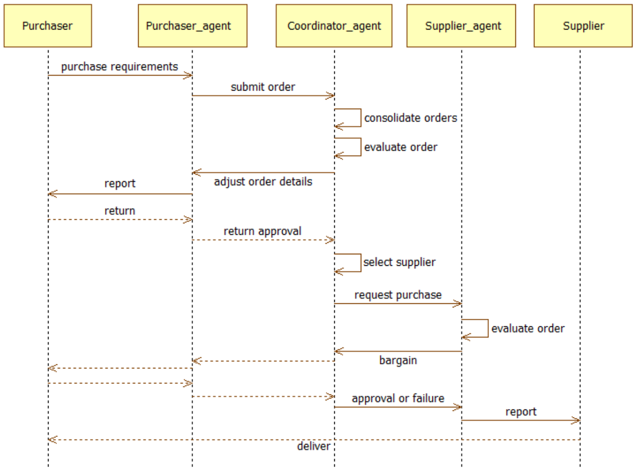


Figure 3 Sequence diagram of collaborative purchasing process

Step 4. CA selects one or several reserved supplier enterprises. When the deadline time is out, the purchase request notification is send to related SA.

Step 5. CA bargains with SA, then CA and SA will fix a price that is accepted by both of them. In this stage, CA may seek help from purchaser in some special situations. If CA and one SA can not come to an agreement, CA may negotiate with another one. The most terrible situation is that CA can not agree with any SA, but this probability is also very tiny.

Step 6. The product will be delivered to purchaser enterprise. If the members of purchase alliance are located in the same region, such as the same province, the same city, the delivering cost will reduce.

5 Collaborative Purchasing Service in Manufacturing Grid

Extension services

The collaborative purchasing service consists of the kernel service and several extension services. The kernel service is just the negotiation process in Section 4, which is accompanied by four extension components in our collaborative purchasing system. Here are four extension services and brief explanations of them, respectively:

1) *Purchase contract generation service, signature and authorization service*. In the negotiation stage, the purchasing contract will be automatically generated by CA, and then PA signs the electronic contract and generates a printable contract.

2) *CRM/ERP adapter interface service*. When CA and SA come to an agreement, the orders are recorded to the internal management system of supplier and purchase alliance member, respectively, through the CRM/ERP adapter interface service.

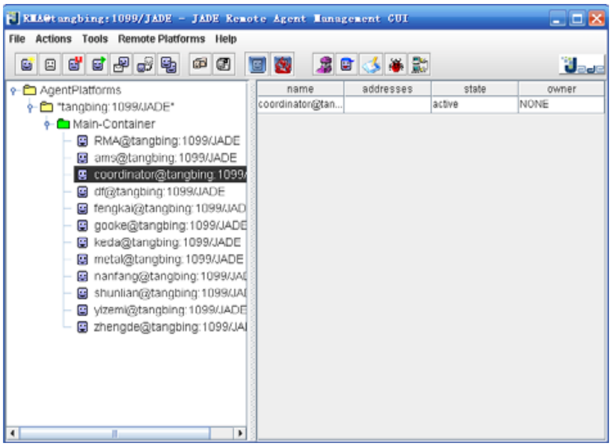
3) *Request and order tracing service*. Purchaser can easily query and tracking the status of its order from the web portal.

4) *Status altering notification service*. If the status of order is changed, such as changed from unfinished to finished, the purchase enterprise will be informed in the form of e-mail or mobile phone SMS in time.

Implementation of prototype system based on Multi-Agent

The collaborative purchasing prototype system is implemented using JADE. JADE is a software framework fully implemented in Java language, which simplifies the implementation of Multi-Agent systems through a middle-ware that complies with the FIPA specifications [14]. JADE is a fast software tools for Agent system program, you just need write some Java classes to handle logical process and define the communication rules in application layer.

At the beginning of our project, in order to research the primitive communication process flow between Agents, we use one single computer to setup the Multi-Agent system using JADE v3.5 platform. There are eight Agents registered in DF, and they are fengkai, gooke, keda, metal, nanfang, shunlian, yizemi and zhengde, which represent eight enterprises, respectively. In the collaborative system, two enterprises are defined as supplier, and they are nanfang and gooke, while other enterprises are defined as purchaser (see Figure 4).



- AgentPlatforms
  - "tangbing:1099/JADE"
    - Main-Container
      - RMA@tangbing:1099/JADE
      - ams@tangbing:1099/JADE
      - coordinator@tangbing:1099/
      - df@tangbing:1099/JADE
      - fengkai@tangbing:1099/JAD
      - gooke@tangbing:1099/JADE
      - keda@tangbing:1099/JADE
      - metal@tangbing:1099/JADE
      - nanfang@tangbing:1099/JAI
      - shunlian@tangbing:1099/JAI
      - yizemi@tangbing:1099/JADE
      - zhengde@tangbing:1099/JAI

name	addresses	state	owner
coordinator@tan...		active	NONE

Figure 4 Demo of collaborative purchasing prototype system

Deploy Collaborative purchasing services on MEGridSP

There are some add-on tools to package Agent service as Web Services (WS) and integrate the WS with Java Server Pages [15]. One solution is using the Web Service Integration Gateway (WSIG). We learn from Figure 5 that UDDI and Web Services container act as a bridge between JADE and Web Services. Upon on the architecture in Figure 5, it is convenient to write some Web Services, and make them comply with the Web Service Description Framework specification using some integrated programming tools. The collaborative purchasing services are published as application services on MEGridSP platform and consumed by Grid end users or other application services providers.

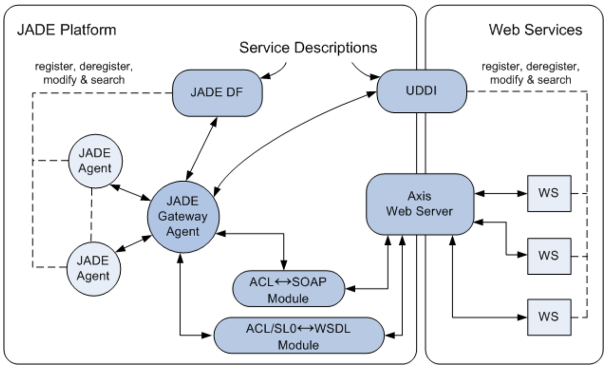
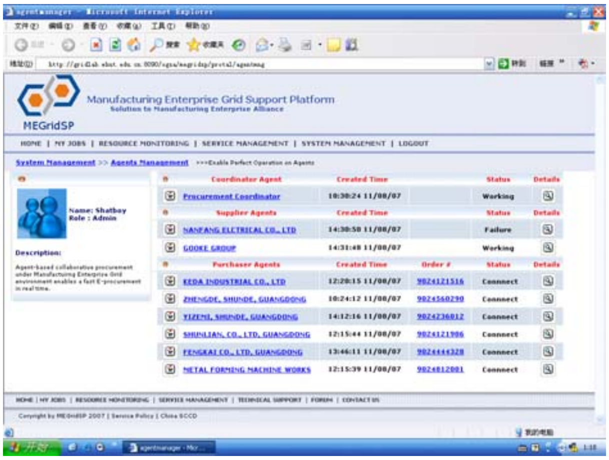


Figure 5 Architecture of the Web Services Integration Gateway

The case study is conducted based on the experiment environment of combination MEGridSP with JADE. Apache Tomcat and Apache Axis Web Server are set as the Web Services container, with MySQL as DBMS, and Globus 4.0.5 is used as Grid middleware, Eclipse MAGE plug-in is used as Grid service deployment tool, open source jUDDI is adopted to setup a Grid Service Index server. MEGridSP Grid portal and JADE v3.5 platform are interconnected according to WSIG framework.

Figure 6 shows a picture of Agent management webpage in MEGridSP Grid portal. It is a friendly visual human-computer interface, which is responsible for managing and monitoring the three important Agents, these are CA, SA and PA. It is easy to figure out the status of each Agent, including their name, created time, and order ID. Each enterprise can login to review and query the detailed negotiation log, and browse the detailed information of each enterprise and purchase order.


agentmanager - Microsoft Internet Explorer
http://gridlab.whut.edu.cn:8090/ogsa/megridsp/portal/agentmng
Manufacturing Enterprise Grid Support Platform
Solution to Manufacturing Enterprise Alliance
MEGridSP
HOME | MY JOBS | RESOURCE MONITORING | SERVICE MANAGEMENT | SYSTEM MANAGEMENT | LOGOUT

System Management >> Agents Management >>>Enable Perfect Operation on Agents

Name: Shatboy
Role : Admin

Description:
Agent-based collaborative procurement under Manufacturing Enterprise Grid environment enables a fast E-procurement in real time.

Coordinator Agent	Created Time		Status	Details
Procurement Coordinator	10:30:24 11/08/07		Working	
Supplier Agents	**Created Time**		**Status**	**Details**
NANFANG ELECTRICAL CO., LTD	14:30:50 11/08/07		Failure	
GOOKE GROUP	14:31:48 11/08/07		Working	
Purchaser Agents	**Created Time**	**Order #**	**Status**	**Details**
KEDA INDUSTRIAL CO., LTD	12:20:15 11/08/07	9024121516	Connect	
ZHENGDE, SHUNDE, GUANGDONG	10:24:12 11/08/07	9024560290	Connect	
YIZEMI, SHUNDE, GUANGDONG	14:12:16 11/08/07	9024236012	Connect	
SHUNLIAN, CO., LTD, GUANGDONG	12:15:44 11/08/07	9024121906	Connect	
FENGKAI CO., LTD, GUANGDONG	13:46:11 11/08/07	9024444328	Connect	
METAL FORMING MACHINE WORKS	12:15:39 11/08/07	9024812001	Connect	

HOME | MY JOBS | RESOURCE MONITORING | SERVICE MANAGEMENT | TECHNICAL SUPPORT | FORUM | CONTACT US
Copyright by MEGridSP 2007 | Service Policy | China SCCD
agentmanager - Mic...

Figure 6 Demo of Agent management in MEGridSP Grid portal website.

Our collaborative purchasing service is better than traditional collaborative purchasing website as introduced in Section 2.3. The collaborative purchasing service complies with the OGSA criterion, upon the manufacturing Grid infrastructure. Due to the advantages of manufacturing Grid, the collaborative purchasing service can be commonly shared by regional enterprises, and help saving material cost for enterprises. It is an emerging and potential solution to collaborative enterprise management.

6 Conclusions

With the tendency of global manufacturing, the competitive market requests the collaborative development of manufacturing enterprise. Enterprise resource collaborative management is the main goal of manufacturing Grid. The collaborative purchasing for enterprises has theoretical and practical significance, which decreases the finance cost for material purchase, accelerates the response speed of supply chain, and strengthens the cooperation and fellowship. We proposed a collaborative purchasing scheme based on Multi-Agent. This scheme has the features of intelligent, automated negotiation, standard platform, extendable architecture, and so on. The collaborative purchasing services can be deployed and integrated to MEGridSP platform easily.

References

[1] Robin G. Qiu, "Manufacturing Grid: A Next Generation Manufacturing Model", Proceedings of 2004 IEEE International Conference on System, Man and Cybernetics, 2004, pp. 4667-4672

[2] Yefa Hu, Fei Tao and Zude Zhou, "The Connotation of Manufacturing Grid & Its Key Technology", Proceedings of 2006 IEEE International Conference on Industrial Informatics, 2006, pp. 293-298

[3] Zhou Zhou, "Collaborative Purchasing in Small and Middle Sized Enterprises in Region", SMEs Sci., No.7, 2007, pp. 69-70

[4] Yang Jin, You Jianxin and Cai Yiping, "Multi-Agent System Architecture of Supply Chain Management Based on Competitive-cooperative Mechanism under Industrial Cluster", Application Research of Computers, Vol.24, No.10, 2007, pp. 263-266

[5] Rayson P. Relate, "Empower and Free (Collaborative Product Commerce)", Manufacturing Engineer, 2001, No.80, pp. 8-12

[6] I. Foster and C. Kesselman, The Grid 2: Blueprint for a New Computing Infrastructure, Morgan Kaufmann, USA, 2003

[7] I. Foster, C. Kesselman and S. Tuecke, "The Anatomy of The Grid: Enabling Scalable Virtual Organizations", International Journal of High Performance Computing Applications, Vol.15, No.3, 2001, pp. 200-222

[8] MEGridSP project, http://gridlab.whut.edu.cn/megrid

[9] Hirotaka Ogawa, Satoshi Itoh, Tetsuya Sonoda and Satoshi Sekiguchi, "GridASP: An ASP Framework for Grid Utility

Computing", Concurrency and Computation: Practice and Experience, 2007, No.19, pp. 885-891

[10] Heekwon Chae, Younghwan Choi and Kwangsoo Kim, "Component-based Modeling of Enterprise Architectures for Collaborative Manufacturing", Internatioanl Journal of Advanced Manufacturing Technology, Vol.34, No.6, 2007, pp. 605-616

[11] Dominik Vanderhaeghen and Peter Loos, "Distributed Model Management Platform for Cross-enterprise Business Process Management in Virtual Enterprise Networks", Journal of Intelligent Manufacturing, Vol.18, No.5, 2007

[12] N. R. Jennings, P. Faratin and A. R. Lomuscuo. "Auto-mated Negotiation: Prospects, Methods and Challenges", International Journal of Group Decision and Negotiation, Vol.10, No.2, 2001, pp. 199-215

[13] Chen Dejun, Li Ting and Zhou Zude, "Research on Auto Negotiation of E-commerce System on Multi-Agent", Micro-Computer Information, 2007, Vol. 23, No.7, pp. 164-165

[14] JADE, http://www. jade.tilab.com/

[15] JADE development docs, http://jade.tilab.com/doc

The Application of the Business Blueprint Engineering Design Technology in Elaborating System Functional Requirement of Upper Levels

Xiangang Chen[1] Mingge Li[2] Jingyan Sun[3]

1 School of Information Technology, Changchun Vocational Institute of Technology
Changchun 130033, Jilin, China
Email: 1 cxg2000209@sohu.com, 2 limingge66@163.com, 3 sunjingyan1111@tom.com

Abstract

Business blueprint is a flow chart carried out on computers, and a graphical model, which is not influenced by software development methodology, to elaborate system functional requirement of upper levels; Business blueprint is composed of transaction scene, institutional framework, data transaction and event. Designing a business blueprint mainly has five steps: describe transaction scene, explicit institutional framework, compile event table, ascertain data transaction and Realize by the tool of Microsoft Visio at last. The application of the business blueprint Engineering Design technology in the exploitation of large-sized software has very important significance.

Keywords: Business blueprint, transaction scene, event, data transaction, event table

1 The Significance of Business Blueprint

In the exploitation of large-sized software, establishing the system function requirement model, whichever system method of exploitation is used, starts at the business blueprint(Like Figure 1), business blueprint is a model used to elaborate system functional requirement of upper levels, and has very important significance in further system function modeling[1].

2 What is the Business Blueprint

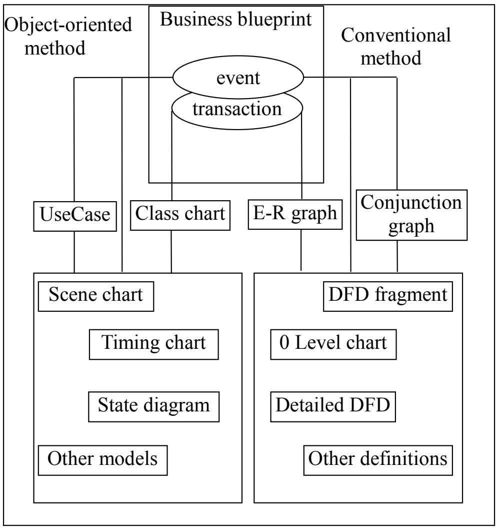


Figure 1 Demand model of object-oriented method and conventional routes

Business blueprint is a computerized operation flow model based on the system function, which established according to the business scene and the related event, transaction in the business events flow activity classification. The operation flow is a true picture of process how enterprise manage their business handling. the reason that we carry on carefully search is to develop an information system, also called target system, adapted to the enterprise business requirement. Therefore we must carry on the careful analysis to the original flow, definite the boundary of the information system, and to each activity in boundary, we must describe the department and the post it should be responsible by, the firsthand information it according to, the processing it carries on, results it obtains, activity to entere after this execution finished, this process reflects the behavior characteristic of target system, it stemmed from the operation flow, but expressed rigorousness and standard also won in the operation flow. We call it Business blueprint.

3 The Integrant Part of Business Blueprint

Business blueprint is composed of transaction scene, institutional framework, event and data transaction.

3.1 Transaction Scene

The enterprise general operation flow is an overall description of business, the requirement carries on the refinement unceasingly in the system analysis, as a result of the difference between handling object , the operation flow will have obvious change, and form many specific operation flows, which is called transaction scene[2]. take enterprise purchase and physical distribution as the example, according to the purchase material's characteristic, the transaction scene can divided into non- production material, production material, substituted storaging material, pipeline material and so on[3].

3.2 Institutional Framework

In system analysis, organization is composed of superintendent of each active department and post constitution in operation flow. Take purchases and physical distribution as the example, the superintendent has the supplier, the purchase department, the physical distribution department, the finance department and so on, the post has the supplier contact person, the purchaser, the warehouse keeper, the accountant and so on .In system analysis period, we can first determine the department then clear about the post later, and implementation the mapping between post and staff when it carried out. What should stress, since the Business blueprint, target system's organization and management, the staff management, the jurisdiction management and the access control must be able-defined and able-maintained[4].

3.3 Event

(i) The definition of event

All the processes of the system are driven or triggered by the the events, it is essential to traverse the event and then analye it during the time of defining high-level system requirements. The event refers to the meaningful matter which accurs in some specific time and the place. At the stage of system analysis, we can concentrate on the interface between the system and the user outside through analysis the influence produced by the system and regarding all the system as a black box[5]. Users, who actually perform the system, also used to describe the system requirements according to the events that affect their work; There is another division strategy, advantageous for the team-cooperation during system analysis, which divides the complex system into smaller units[6].

(ii) The type of event

Based on the character of event occured, event is classify as external event or clock event or state event.

The external event is the occurrence which triggers by the exterior entity or the system participant outside system's event. In order to distinguish the external event, at first displays the exterior entity, either the human, the organization unit, the exterior installment or other systems, which refers to gain information from the system or provides the data for the system. Second, list each external entities to provide any information, and thus trigger system or increase, or updated, or deleted, or to deal with matters such as descriptions of activities; Finally, the list to outside entities from the system access to what information, and thus an appropriate system for tracking and descriptions of activities[7]. Important external events may also come from within the company or organization needs.

Temporal event is the event that system occurs at a certain moment. A system that is scheduled at the moment, or the prior definition of a good time interval, when this moment arrived when the system of some processing and export of certain results, like the end of month, or the end of the period, at the end of the accounting treatment, the periodic statements, etc. The system, which is essential different from external event, automatically triggered by temporal event, not participants in the movement system, automatic processing and to provide the information needed by the user[8]. The best way to identify the temporal event is the way to list companies inside and outside the administration needs all the output, analyze time-related output and thus determine the clock incident[9].

State event is the event that has taken place within the system and need to address the situation triggered by the event. For example, a transaction processing to the end, the system has been found below the stock orders, it is necessary to re-orders, which generated a "landing orders," state, this has triggered system, "dealing with orders" and other processing activities. Generally speaking, the state of events generated point in time, not identified by the clock, but external events triggered by business at the end of the trigger, which is different from temporal event. Recognition from the state event starts with the result of external event[10].

(iii) Event table

Systems analyst shall identify and list as many as possible of the event, identify the target system a list of events, constantly refining each event and a detailed description. Event table is the case for firms to the various events out key information for a list of events (Like Table 1). Events critical information from the incident, and flip-flop, the sources of information, processing, value and information to return to a destination. Events are implementing a system that caused the operation of events marking; Trigger notification system is a case of things; Sources of information is defined as a system to provide data external entities or participants; Processing is an implementation time of the incident system the operation; Return value refers to the output system to deal with after the end of the entities of the ministry of information; Destination information refers to the output data receiving system outside entities or participants.

3.4 Data Thing

Things are objective of people, goods, location, composition, events, interaction; the world is composed of things. The target system is also composed of things. Objective data-processing systems of the general process is: enter a process of processing an output (IPO). Input and output data is what matters, that is the lasting memory of things, usually kept in writing or storage devices. The logistics involved in the procurement data main things for a single purchase, purchase orders, supplies, mobile documents, inventory records, financial documents and accounting entries, general ledger, inventory, accounting, and so on.

4 The Steps of Designing a Business Blueprint

Designing a business blueprint mainly has five steps: describe transaction scene, explicit institutional framework, compile event table, ascertain data transaction and Realize by the tool of Microsoft Visio at last[11].

Below we take brief supermarket purchase and physical distribution as an example to illustrate the designing of business blueprint:

4.1 Describe Transaction Scene

In the supermarket operation, the supplier is determined by the difference between supplier and product demand quantity according to the quartering regulations, Namely: When product demand quantity is small and the supplier are lack, uses the fragmentary ordering; When product demand quantity is large, the supplier are in plenty, uses the long-term order form ordering[12]; When product demand quantity is small and the supplier are in plenty , uses on-line fragmentary ordering; When product demand quantity is large, the supplier are in plenty, uses on-line tender ordering, here is the business blueprint of merchandise purchase in second kind of situation[13].

4.2 Explicit Institutional Framework

The major composed organization merchandise purchase business involves is the supplier, the purchase department, the warehousing department, the finance department, the Bank account and so on.

4.3 Compile Event Table

Recognizing all the system events, then carring on the description according to the key information such as event name, trigger, information source, processing, the returns value and the information destination and so on, and compiling into the event table at last (Like Table 1).

Table 1 Event Table of Supermarket Information Management System

Event name	Trigger	Information source	Processing	Returns value
Found purchase order form	Procurement request	Purchase department	Found new head of the order form and the detail	Order form and detail
Get goods according to the order form	Certificate of arrived shipment	Warehousing department	Get goods according to the order form and the certificate of arrived shipment	The detail of order form transaction
Checkup order form	Trades detailed	Accounting division	Checkup the received goods	the certificate of received goods
Payment	Completed goods of reception	Accounting division	Paying according to the agreement	Payment certificate
Found purchase order form	Procurement request	Purchase department	Found new head of the order form and the detail	Order form and detail

4.4 Identify Data Thing

The main things in this case: purchase orders, warehousing certificate, stock accounts, accounting documents, accounts payable, etc.

4.5 Draw Business Blueprint

Business blueprint for the customary "Lane plans," said Lane at the top of each specified responsibilities, status and their respective departments, in this case used by the graphic symbols from Microsoft Visio (Figure 2).

Business blueprint which is a flow chart for the computerization of the business, is a software development methods without affecting the definition of high-level system needs graphics model, is to solve complex software development needs of the important means of software enterprises in Europe and the United States in the application of a broad, with the China's software development companies are growing, the business blueprint for the modeling study and application of technology is of great significance.

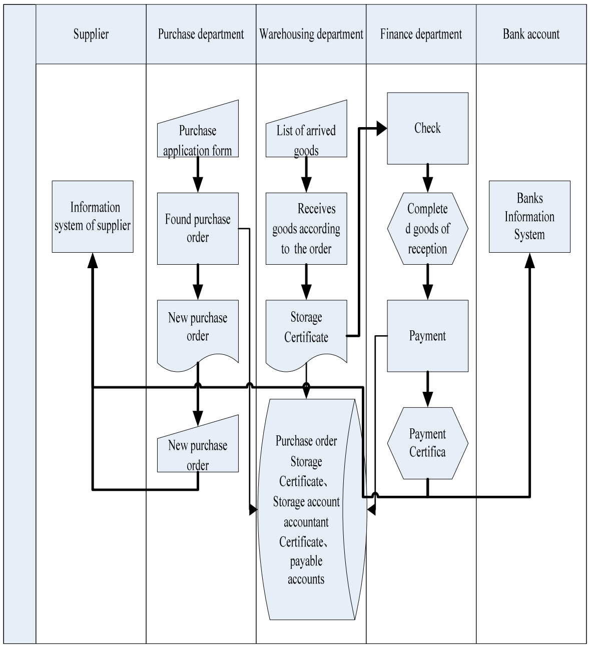


Figure 2 Business blueprint of merchandise purchase

References

[1] John W.Satzinger, Robert B.Jackson ,Stephen D.Burd, Systems Analysis and Design in a Changing World, China Machine Press,Citic Publishing House,2002

[2] Brett D.McLaughlin,Gary Pollice,David West,Head First Object-Oriented Analysis & Design, O'Reilly Media,2007

[3] Liu Jianfeng,My Summarization of Requirement Ana- lysis in Project Management,System Analyst of China 2005

[4] Mike O'Docherty,Object-Oriented Analysis & Design (Understanding System Development with UML 2.0), Tsinghua University Press,2006

[5] David Conger,Software Development in c A Practical Approach to Programming and Degin,Tsinghua University Press,2006

[6] Kenneth E.Kendall, Julie E.Kendall, Systems Analysis and Design,Tsinghua University Press,2006

[7] R.J.A Buhr, R.S.Casselman, Use Case Maps for Object-Oriented Systems,Tsinghua University Press,2006

[8] Erich Gamma, Richard Helm, Ralph Johnson, John Vlissides, Design Patterns:Elements of Reusable Object-Oriented Software, Addison-Wesley,1995

[9] John W.Satzinger,Robert B.Jackson,Stephen D.Burd. Systems Analysis and Design in a Changing World. China Machine Press,Citic Publishing House,2002

[10] Peter Coad,David North,Mark Mayfield,Object Models Strategies Patterns and Applications(2nd ed.)Prentice Hall, 1997

[11] Scott E, Donaldson, Stanley G.Siegel, Cultivating, Successful Software Development: A Practitioner's View, Prentice Hall, 1997

[12] Shao Weizhong, Yang Fuqing, Object-Oriented Systems Analysis, Tsinghua University Press, 2006

[13] Peter Coad, David North, Mark Mayfield, Object Models, Strategies, Patterns, and Applications(2nd ed.), Prentice Hall, 1997

The Application of Fault Diagnosis for Vehicle System Based on BP

Lifang Kong[1] Hong Zhang[2]

1 College of Information and Electrical Engineering, China University of Mining and Technology
Basic Education Department, Xuzhou Air Force College, Xuzhou, Jiangsu 221000，China
Email: klf030@163.com

2 College of Information and Electrical Engineering, China University of Mining and
Technology, Xuzhou, Jiangsu 221000，China
Email: klf030@163.com

Abstract

In accordance with the fault problems of vehicle systems, we introduce the pattern identification technology to diagnose them. The fault types、data gathering and signal processing are analyzed. As an example, we make an analysis of clutch fault of automobiles of applying a modified artificial neural network algorithm, while offering the fundamental theories and compute course .The algorithm is of higher convergence speed. A better result is obtained.

Keywords: vehicle systems, fault diagnosis, neural network, pattern identification

1 Introduction

As the development of the automobile industrial modern technology is improved, especially by wildly application of the computer science and controlling science on vehicles, which adds the vehicle system's complexity on one hand, and on the other hand, when the whole system appears some delicate malfunction, if it can't be detected and obviated in time, it may conduct the vehicle invalidity、paralysis, and even some tremendous tragic result. Malfunction can be explained that at lest one important variable or characteristic of the system deviates from the normal scope. Malfunction diagnosing is an integrated process which utilizes kinds of states information of the operation system being diagnosed and kinds of existed knowledge to synthesize and dispose the information, and to conclude a synthetically evaluation about the system's work state and malfunction state in the end. It is a process which changes from quantity to quality when the malfunction occurs. Going with the process, the diagnosing parameters change by all means. Malfunction diagnosing is indeed a kind of pattern classifying problem, it is just a problem how to conclude the system's state is normal or a kind of malfunction, ordinarily the malfunction pattern distribution is anomalous, so the pattern classifying method which is needed should be required to form multifarious nonlinear partitioning plane in the pattern space. NN is naturally a kind of nonlinear sorter with perfect capability for its characteristic. In this article, we apply the ameliorated algorithm of BP networks on pattern recognition of the vehicle system's malfunction. The malfunction diagnosing of the vehicle's system is a complicated systems engineering. Because of the complexity of the modern vehicle's structure, the manifold of the electrical elements and the improvement of the self controlling capability, which make it need to apply the advanced theory on disposing the system's malfunction diagnosing problems.

2 Malfunction Analysis of the Vehicle's System

2.1 Malfunction classifying

Generally speaking， the malfunction occurring in the system presents diversity in position、form and the characteristic varying with time. For a vehicle transport system, we can divide the malfunction into four parts by their positions:

1) Malfunction of engine accessories: as some accessories of the vehicle's system or even a certain part of the system appears abnormal, make the whole system can't perform its function，such as the driving system and its components go wrong、the braking system and its components go wrong、the engine goes wrong、the moving organ and the wiring system go wrong.

2) Malfunction of sensors: Nowadays all the automobiles install more kinds of sensors, especially in the engine and the braking system, if the sensor which is used to check the variable goes wrong and fails to get the information being checked well and truly, it makes people can't see actual state of the automobile, in appearance of the difference between the detected quantum and the actual quantum of the object variable, which leads the vehicle to malfunction state.

3) Malfunction of electro circuit system: the executing component in the control circuit which executes controlling orders goes wrong and can't fulfill the motion request correctly, in appearance of the difference between the system's input orders and the system's actual output, which leads part of the circuit's function invalid.

4) The lure's quality declines. Superfluous impurity such as grinded granule in the grease may lead friction suffers gravelly when the automobile keeps moving.

2.2 Malfunction data collection

Diagnosing technology is such a process that analyzes and researches from the system's symptom which is concluded from the collected signal, so the signal collecting technology is one basic of the systems diagnosing. If we collect the signal reflecting the system's actual state, the following work of diagnosis is signification. The signal collecting technology includes signal collecting and signal magnifying; research on sensors is the emphasis among the technologies. Diagnosing system's sensors are divided into the following kinds by their function: sensors for vibration, sound level meter, and sensors for sound emitter, sensors for oily liquid's grinded granule and sensors for temperature etc. Formerly, the research on sensors focuses on hardware，as that those sensors are required to have well dynamic characteristic, stethoscope, stability and strong anti-jamming characteristic. But with the inspecting and measuring system's magnifying and complication, sensors' types and amount increase quickly, multitudinous sensors form the sensors' cluster which brings the research on how to assemble the sensors layout.

2.3 Signal analyzing and disposing method

In the vehicle's system, the vehicle's signal being collected by sensors cluster is called original signal, a part of the signal can be utilized directly such as temperature、displacement; But a majority of the signal can't be utilized directly, such as liberation. Although the signal has be magnified, because it composes yawp jamming signal, generally it is difficult to reflect the system's essential problems from the single wave. In order to get the characteristic genes reflecting the vehicle state sensitively, we should utilize the signal analysis and disposal technology to transform the signal to different regions to analyze. The waves-filtering technology and the frequency composition-analyzing technology are the conventional signal- disposing methods. In recent years, with appearance of the digital waves-filtering technology 、the wavelet analysis technology 、the neural networks and illegibility technology, these enrich the content of the signal analysis.

3 BP Neural Networks and its Application

3.1 The essential principle of BP neural networks

Of all the neural networks models, the error backward propagation neural network (BP model) is popularly used. BP algorithm has been regarded as a method which extends the repeatable regression analysis to nonlinear domain. The neural networks model generally composes of input nodes、concealed nodes、output nodes and intermeddle nodes. The learning process is made up of two proportions including the forward propagation and the backward propagation. Seeking the error function is a recursive process to promulgate from the output layer to the input layer; amending the tentative value by learning the training samples repeatedly and making the tentative value change along the negative grads of the error function, in the end it stabilizes to the minimum. As the following chart is a multi-layers forward propagation neural networks model. The mathematical model is:

$$E = \sum_{P=1}^{P} E_P < \varepsilon$$

p represents all the samples in this formula; E represents the total error after training; E_P is the number p sample's output error; ε is anyone of the small positive numbers. Among those,

$$E_P = \frac{1}{2}\sum_{j=1}^{n}(d_{pj} - o_{pj}^{n}) \ o_{pj}^{l} = 1/(1+e^{-net_{pi}^{l}})$$

$$net_{pj}^{l} = \sum_{i=0}^{m1} w_{ij}^{l} o_{pi}^{l-1}$$

In above those formulas, d_{pj} , o_{pj}^{n} represent the system's expectation and networks' actual output value separately; w_{ij}^{l} is nerve cells' connection power coefficient, the original value is random; m, n, l, ml respectively is the number of nerve cells、the number of the networks' layers、the number of the connotative layers and the number of the former layer's neural cells. Adopting the BP algorithm of grad declining method exists the defect that the rapidity of convergence is slow and can not always converge to local minimum, to the function of none convexity, it is quite possible to span the whole territory's minimum spot when it searches, and it can't retain the character of descending monotonously. Being dead against above problems, we fetch two coefficients η and a in application, these two parameters η and a can affect the rapidity of the networks' learning.

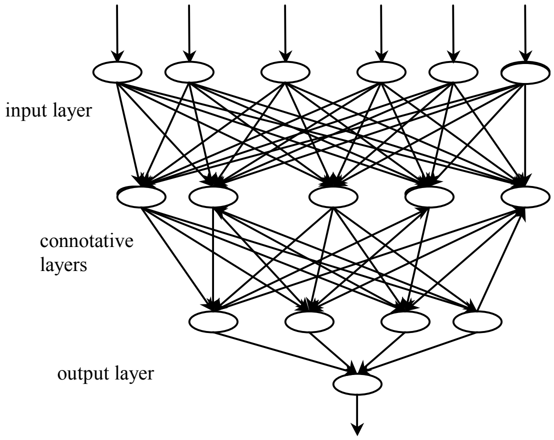


Figure 1 Multi-layers BP neural networks model

3.2 The amended BP algorithm

Applying the method of dynamically adjusting the learning rapidity and the parameters of stimulant function to amend the BP algorithm has acquired better effect. To amend the power coefficient of the networks' learning process as following:

$$\Delta w_{ij}(k) = \eta\delta_{pj}o_{pi} + \alpha\Delta_{ij}(k-1)$$

$$w_{ij}(k+1) = w_{ij}(k) + \Delta w_{ij}(k), \delta_{pj} = d_{pj} - o_{pj}^{n},$$

in above formulas, k is the iterative time (of learning); η is the relaxation coefficient; α is the inertia coefficient, both value can be put 0.5, the bigger is the value of α, the bigger is the amending value getting from error, the faster is the rapidity of learning. But if the values of η and α are too bigger, they can arose the system surging and debase the networks' capability. Generally at the beginning phase of learning, because the error is quite bigger, we should increase the parameters to enhance the rapidity of convergence. The flow of concretely computation is shown in Figure 2.

3.3 Practical application research

The features of neural networks' application in malfunction pattern recognition are following:

1) Neural networks can be applied on the malfunction pattern recognition of the unknown system models or the more complicated system models and the

nonlinear system; 2) Neural networks has the function of malfunction signal's pattern transformation and the characteristic's distillation;

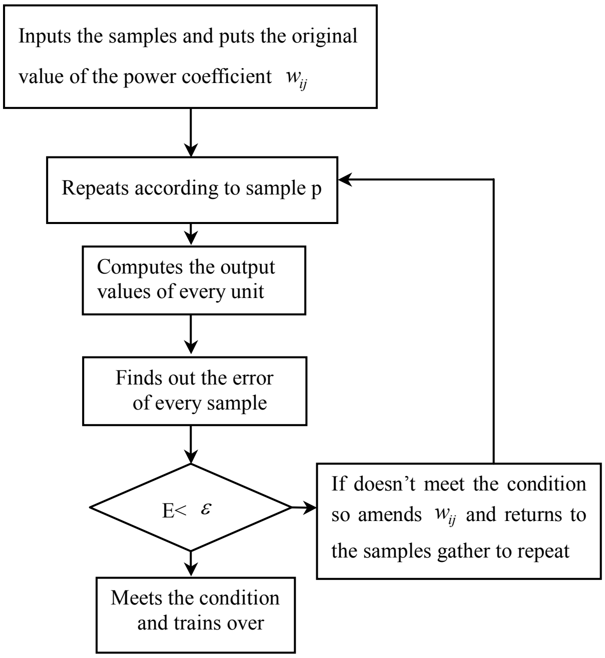

Inputs the samples and puts the original value of the power coefficient w_{ij}
Repeats according to sample p
Computes the output values of every unit
Finds out the error of every sample
E< ε
If doesn't meet the condition so amends w_{ij} and returns to the samples gather to repeat
Meets the condition and trains over

Figure 2 The flow chart of amended BP algorithm

3) Neural networks is not sensitive to the situations when the system composes the indeterminate fact、nose and the imperfect input pattern;

4) Neural networks can be applied on the malfunction diagnosing of the complicated multi-patterns;

5) Neural networks can be used to diagnose offline and can adapt the real time request. This article adopts the three layers forward feedback BP networks in malfunction diagnosing, every node of the connotative layer connects with every node of the input layer and output layer by a certain coefficient, every node of the connotative layer and output layer has a nonlinear active function, that is sigmoid function. The networks' input nodes correspond malfunction sign and the output nodes correspond malfunction causes. Doing malfunction pattern recognition, we should firstly use a passel of malfunction samples to train the model in order to ascertain the connection coefficient of the networks' structure (connotative layer and its nodes number) and nodes, after the networks has been trained perfectly, the malfunction pattern classifying is just a process that implements a nonlinear mapping from the sign gather to the malfunction gather basing on a set of sign gather, adding the malfunction sign to the neural networks' input end, we can get the proper diagnosing result. Following, we make the example of the malfunction of automobile clutch working under the hydromechanics to analyze the application of BP neural networks technology in malfunction disposal of vehicle's system. Being one of the primary part of the vehicle's transmission system, clutch will appear kinds of malfunction states as time goes by, basing on daily collection, the familiar malfunction causes are: such as the spine slides not well、the twist reed ruptures、surface of the frictional piece indurate、wears and tears、changes shape and the separate bearing gets marred、 the pulling fork and the separate bearing slide not well. Applying the BP networks which has be amended to dispose above malfunction tree model, basing on some malfunction information about clutch being obtained from some garages, distilling 22 parameters in all to act as the input samples of the networks, such as following: grease adheres the frictional piece、the spine slides not well、the rivet looses、is broken、the twist reed ruptures、the gasket is broken and shattered、the spine is wear and tear、the surface is stiffened、the frictional piece is wear and tear、the frictional piece changes shape、the pressing reed is broken、the pulling fork and the separate bearing slide not well、 the separate perches are not trim、the separate bearing gets marred、the main oil vat gyres not well、the impetus handspike adjusts improperly、the braking oil interfuses with air、the braking oil pipeline leaks、the flywheel's fixative bolts become flexible、the rubber mats are broken、the position of installation becomes flexible、the separate bearing gets marred. Making the 22 entries information to unification and marking $x_1 --x_{22}$.With malfunction, the input value is 1; the natural state is 0; to choose y_i(i =1~ 4)as the system's output, it separately denotes the situation that the clutch skids、the clutch separates incompletely、the clutch sounds exceptional and works with dithering, that the output is 1 shows there is malfunction otherwise 0

represents the system is normal, training the networks, conclude the detailed input values and the output values. Many times of iteration can result well trained applied networks. The following learning networks has completed 5112 times of iteration, its total error E has been less than any one of the given positive number, it behaves good astringency, the training results reveals the networks' computing value is basically as equal as the expected output value, which proves this algorithm is feasible, basing on the well trained networks, if only inputting the malfunction about the clutch to the networks, can we get the basically correct judgment rapidly to the malfunction of the clutch. In practical problem, we can use sensors and measuring meters to input the information of malfunction to the networks to form an automatic detecting system, the results of the networks' computing can be shown to users directly in order to obviate the malfunction in time. Adopting manifold information inosculating technologies can improve the degree of autoimmunization and veracity of the detection. There are some problems in practical application of BP networks, seeing from mathematics, it is came down to a nonlinear grad optimizing problem, so it inevitably has the problem of local minimum; the learning algorithm's rapidity of convergence is quite slow, it looks that its ability is not equal to its ambition when it faces the large dimension of problem; the networks' stability and plasticity are bad, as one of the completely learned BP networks meets a new mode, the existed connection coefficients are thrown into confusion, which lead the learned mode's information to disappear, so it need to train the networks over again before the well learned mode's information can be still applied. In addition, what is the appropriate number of the nerve cells lying up in the mesosphere of networks? If a small quantity of the learning samples' results have generic application? Such problems still need to be ulterior researched.

4 Conclusion

The malfunction diagnosing of vehicle's system is a complicated systems engineering, because of the complexity of the modern vehicle's structure, the manifold of the electrical elements and the improvement of the self controlling capability, that make it need to apply the advanced theory to dispose the system's malfunction diagnosing problems. This article applies the technology of the BP neural networks which has been amended to analyze the clutch's malfunction of the vehicle's system, this method has self-learning ability and can diagnose the system's malfunction rapidly and quite perfectly.

References

[1] Qu Liangsheng, "Intelligentized Problems of Artificial Neural Networks and Mechanical Engineering", Journal of Chinese Mechanical Engineering, Vol.2, No.1, January 1997, pp.1~4

[2] Ho T K, Hull JJ, srihari S N, "Decision combination in multiple classifier systems", IEEE Trans on PAMI, Vol.16, No.1, pp.66~75, 1994

[3] Wei Shaoyuan ect, "Application and Research of BP Neural Networks in Malfunction Diagnosing of Automobile", Journal of Modern Machine, Vol.3, 2001, pp.48~50

[4] Wang Wencheng, Neural Networks and Its Application in Automobile Engineering, Beijing: Beijing Science and Technology University Press, 1998

[5] Yan Cengren, Artificial Neural Networks and Application, Beijing: Tsinghua University Press, 1999

A Hybrid Particle Swarm Optimization and Its' Application in VRP

Yang Peng[1] Yemei Qian[2]

1 School of computer and information engineering, Zhejiang GongShang university, Hangzhou. China, 310018
Email: pengyang@mail.zjgsu.edu.cn

2 Hangzhou commercial college, Zhejiang GongShang University, Hangzhou. China, 310018
Email: Qianyemei@tom.com

Abstract

In this paper, a novel real number encoding method of Particle Swarm Optimization (PSO) for Vehicle Routing Problem is proposed. Which firstly construct a suitable mapping between problem solution and PSO particle, and in the evolution of PSO, SA algorithm is used to optimize the sequence of the customers served by each vehicle. To illustrate the effectiveness and good performance of the proposed algorithm, a number of numerical examples are carried out, and the algorithm is compared with other heuristic methods for the same problem.

Keywords: vehicle routing problem, particle swarm optimization, Simulated annealing algorithm

1 Introduction

Particle swarm optimization (PSO) algorithm is a parallel population-based computation technique originally developed by Kennedy and Eberhart[4][5], which was motivated by the organisms behavior such as schooling of fish and flocking of birds. PSO can solve a variety of difficult optimization problems. PSO's major difference from genetic algorithm (GA) is that PSO uses the physical movements of the individuals in the swarm and has a flexible and well-balanced mechanism to enhance and adapt to the global and local exploration abilities, whereas GA uses genetic operators. Another advantage of PSO is its simplicity in coding and consistency in performance. According to its advantages, the PSO algorithm is not only suitable for scientific research, but also has been widely applied in many fields. Although the PSO is developed for continuous optimization problem initially, there have been some reported works focused on discrete problems recently.

The vehicle routing problem (VRP), which was first introduced by Dantzig and Ramser (1959), is a well-known combinatorial optimization problem in the field of logistics and service operations management. The capacitated vehicle routing problem (CVRP) is an NP-hard problem for simultaneously determining the routes for several vehicles from a central depot to a set of customers, and then return to the depot without exceeding the capacity constraints of each vehicle. In practice, the problem is aimed at minimizing the total cost of the combined routes for a fleet of vehicles. Since cost is closely associated with distance, in general, the goal is to minimize the distance traveled by a fleet of vehicles with various constraints. Since the vehicle routing problem is an NP-hard problem[6], no exact algorithm can consistently solve VRP-instances with more than 50 customers; thus, the heuristic approaches are considered as reasonable choice in finding solutions for large-scale instances. Available heuristics include simulated annealing algorithms[7], tabu search algorithms [3][6], genetic algorithms [1], and ant colony algorithm[2]. In this paper, we introduce a novel hybrid algorithm based on discrete particle swarm optimization (hybrid DPSO).

2 Founddation of PSO

The general principles for the PSO algorithm are stated as follows. A particle is treated as a point in an M-dimension space, and the status of a particle is characterized by its position and velocity[4]. Initialized with a swarm of random particle, PSO is achieved through particle flying along the trajectory that will be adjusted based on the best experience or position of the one particle(called local best) and the best experience or position ever found by all particles(called global best). The M-dimension position for the *i*th iteration can be denoted as $X_i(t)=\{x_{i1}(t),x_{i2}(t)\dots,x_{im}(t)\}$,similarly, the velocity (i.e.,distance change),also an *M*-dimension vector, for the *i*th interationcan be described as $V_i(t)=\{v_{i1}(t),vi2(t),\dots,viM(t)\}$, the particle-updating mechanism for particle flying(i.e., search process) can be formulated as following.

$$V_i(t)=w(t)V_i(t-1)+c_1r_1(X_i^L-X_i(t-1))+ \\ c_2r_2(X^G-X_i(t-1) \tag{1}$$

$$X_i(t)=V_i(t)+X_i(t-1) \tag{2}$$

Where $i=1,2,\cdots,P$,and P means the total number of the particles in a swarm, which is called population size; $t=1,2,\cdots,T$, and T means the iteration limit; $X_i^L=\{x_{i1}^L,x_{i2}^L,\cdots,x_{iM}^L\}$ represents the local best of the *i*th particle encountered after *t-1* iterations, while $X^G=\{x_1^G,x_2^G,\cdots,x_m^G\}$ represents the global best among all the swarm of particles achieved so far. c_1 and c_2 are positive constants(namely, learning factors), and *r1* and *r2* are random numbers between 0 and 1; *w(t)* is the inertia weight used to control the impact of the previous velocities on the current velocity, influencing the trade-off between the global and local experiences.

3 Modified Discrete PSO for VRP

3.1 The statement and model of VRP

The capacitated vehicle routing problem is a difficult combinatorial optimization problem, and generally can be described as follows: Goods are to be delivered to a set of customers by a fleet of vehicles from a central depot. The locations of the depot and the customers are given. The objective is to determine a viable route schedule which minimizes the distance or the total cost with the following constraints:(1) Each customer is served exactly once by exactly one vehicle; (2) Each vehicle starts and ends its route at the depot; (3) The total length of each route must not exceed the constraint; (4) The total demand of any route must not exceed the capacity of the vehicle.

Assume that the depot is node 0, and *N* customers are to be served by K vehicles. The demand of customer *i* is q_i, the capacity of vehicle *k* is Q_k, and the maximum allowed travel distance by vehicle k is D_k. Then the mathematical model of the VRP based on the formulation given by Bodin et al.(1983) is described as follows:

$$\text{Minimiz}\sum_{k=1}^{K}\sum_{i=0}^{N}\sum_{j=0}^{N}C_{ij}^kX_{ij}^k \tag{3}$$

Subject to:

$$\begin{cases} X_{ij}^k=1 \ \ if\ vehicle\ k\ travels\ from\ customer\ i\ to\ j, \\ X_{ij}^k=0\ \ otherwise \end{cases} \tag{4}$$

$$\sum_{k=1}^{K}\sum_{i=0}^{N}X_{ij}^k=1,\ \ j=1,2,\cdots,N, \tag{5}$$

$$\sum_{k=1}^{K}\sum_{j=0}^{N}X_{ij}^k=1,\ \ i=1,2,\cdots,N, \tag{6}$$

$$\sum_{i=0}^{N}X_{it}^k-\sum_{j=0}^{N}X_{tj}^k=0,\ \ k=1,2,\cdots,K;t=1,2,\cdots,N, \tag{7}$$

$$\sum_{i=0}^{N}\sum_{j=0}^{N}d_{ij}^kX_{ij}^k\le D_k,\ \ k=1,2,\cdots K, \tag{8}$$

$$\sum_{j=0}^{N}q_j(\sum_{i=0}^{N}X_{ij}^k)\le Q_k,\ \ k=1,2,\cdots,K, \tag{9}$$

$$\sum_{j=1}^{N}X_{0j}^k\le 1,\ \ k=1,2,\cdots,K, \tag{10}$$

$$\sum_{i=1}^{N}X_{i0}^k\le 1,\ \ k=1,2,\cdots,K, \tag{11}$$

$$X_{ij}^k\in\{0,1\},\ \ i,j=0,1,2,\cdots,N;k=1,2,\cdots,K, \tag{12}$$

where *N* represents the number of customers, and *K* is the number of vehicles, and C_{ij}^k is the cost of traveling from customer *i* to customer *j* by vehicle *k* and d_{ij}^k is the travel distance from customer *i* to customer *j* by vehicle *k* The objective function Eq.(3) is to minimize the total cost by all vehicles. Constraints

Eqs.(4) and (5) ensure that each customer is served exactly once. Constraint Eq.(7) ensures the route continuity. Constraint Eq.(8) shows that the total length of each route has a limit. Constraint. Eq.(9) shows that the total demand of any route must not exceed the capacity of the vehicle. Constraints Eqs.(10) and (11) ensure that each vehicle is used no more than once. Constraint Eq.(12) ensures that the variable only takes the integer 0 or 1.

3.2 Particle representation

One of the key issues in designing a successful PSO algorithm is the representation step, i.e. finding a suitable mapping between problem solution and PSO particle. In this paper, we setup a *N*-dimension search space, *N* is the total number of customer to be served, $X_i=\{x_{i1},x_{i2},...,x_{iN}\}$, $x_{i1},x_{i2},...,x_{iN}$ is an arrange of the customer's number, denotes the *i*th particle's position in the population, according to the restriction of vehicle's capacity, the encoding can be decompose to several sections, every section denote the customers and the order served by a vehicle. For example, there are 10 customer and 3 vehicle, if a particle's position is:{5,3,7,2,1,4,10,6,9,8}, and decompose to {5,3,7|2,1,4,10|6,9,8}, then it maps to the solution as follow:

Vehicle 1: 5→3→7; vehicle 2: 2→1→4→10; vehicle 3: 6→9→8.

In the particles evolution, the encode of particle will be a real number, then through a process we can transfer it into a integer number, suppose: $x_i(t)'=f(x_i(t-1),v_i(t-1),p_i,g_i)=\{x_{i1}',x_{i2}',\cdots,x_{iN}'\}$ and $\{x_{ij}\mid j=1,2,\cdots N\}$ is a real number, then sort it into $x_i''=\{x_{i1}',x_{i2}',\cdots,x_{iN}'\}$, find the index of x_{ij}' in x_i'' and substitute the original real number. For example, $x_i(t)'=\{3.5,2.3,3.7,7.3,2.0,\quad 8.1,6.2,4.5,9.0,1.8\}$, sort it to $x_i''=\{1.8,2.0,2.3,\quad 3.5,3.7,4.5,6.2,7.3,8.1,9.0\}$, find the index and substitute , then $x_i(t)'$ can be converted into $x_i(t)=\{4,3,5,8,2,2,7,6,10,1\}$ and further decompose it into a solution.

This representation makes each customer can be served by only one depot, and no more than the capacity of the vehicles, moreover, it can increase the rate of vehicle full loaded.

3.3 Hybrid strategy of evolution

Simulated annealing (SA) algorithm is introduced by Kirkpatrick in 1983, which is a metastrategy local search method that attempts to avoid producing the poor local maximum inherent in the steepest ascent method. It employs additional random acceptance and selection strategies. The random acceptance strategy allows occasional downhill moves to be accepted with certain probabilities. SA algorithm has produced good results for many scheduling problems. It starts from an initial solution *s* and a high temperature T_0, and then the temperature is gradually decreased, at each temperature, the search will be performed a certain iteration, every iteration begins from a new produced solution and computers the difference of objective function between current solution and the last, then decides whether to accept the current solution. When the termination condition is satisfied, the algorithm will stop. For the VRP, SA algorithm can be used to optimize the sequence of the customers served by each vehicle.

In the process of sequencing the customers in each route using SA algorithm, the neighborhood selection rule greatly influences the performance of the solution for VRP. We adopt pair-exchange to obtain neighbors, namely, swap the positions of adjacent elements. After exchanging the customers in the same route of a particle using pair exchange rule every time, the fitness of the new solution is calculated. If the fitness is improved, the new solution is accepted. Otherwise, the new solution is accepted with the probability of Blotzmann.

An SA algorithm generally must be carefully designed as the choice of its parameters might affect the quality of the solution and computation. Control parameters were set according to problem characteristics. Through many experiments, we found that the solutions and running time are both better when initial temperature is set according to the maximal difference in fitness value between any two neighboring solutions. The length of temperature denotes the number of moves made at the same temperature, and generally, it is set according to the size of neighborhood solutions for a given solution. In SA optimization process, the temperature is gradually lowered.

3.4 The procedure of the proposed algorithm

Te procedure of the hybrid PSO can be stated as following:

```
Begin
Initialize parameters: swarm size, maximum of generation, $\alpha$, $\beta$, $w$, $c1$, $c2$; Set $t0$, $tf$, $\lambda$, $R$ ; t:=1;
Initialize the particles' position $X$ and the particles' velocity;
Evaluate each particle's fitness according to Eq.(3);
Obtain $X_g$ and $X_p$;
Repeat
Compute $V(t+1)$ according to Eq.(1);
Obtain $X(t+1)$ according to $V(t+1)$;
Compute each particle's fitness according to Eq.(2);
Find new $X$g and $X$p and updated$d$;
Carry out SA subprogram on each route of each particle;
Compute the particle's fitness;
Update $X$g and $X_p$;
t=t+1;
Until (one of termination conditions is satisfied)
Output the optimization results;
End
```

For each route of each particle, executes the SA subprogram as following:

```
SA algorithm subprogram
{$t_n$=$t_0$;
Repeat
$r$=1;
Repeat
Generate a neighboring solution $s'$ from $s$ by the pair exchange rule;
Compute fitness of $s'$, then $\Delta=Fit(s')-Fit(s)$;
If ($\Delta$<0) $s'$ is accepted;
Else If ($rand$<exp($-\Delta/t_n$)) $s'$ is accepted;
Update the best solution found so far; $r$=$r$+1;
Until ($r$>$R$)
$t_n$= $\lambda\times t_n$;
Until ($t_n$<$t_f$)}
```

4 Experimental Results and Discussions

To illustrate the effectiveness and good performance of the proposed algorithm, various kinds of benchmark instances with different sizes have been selected for the computation. We programmed the algorithms in Matlab 6.5 and ran them on Intel Pentium IV CPU 3.0 GHz with 512 M RAM.

For problems of a little larger scale, computational experiment was conducted on the instances from Vehicle Routing Data Sets (http://www.Branchandcut.org/VRP/data/) and the results were compared with those from GA with 2-opt proposed by Wang[10] and those from branch and cut method.

The parameters for Hybrid PSO: maximum number of generations: 100; swarm size: 30; $\alpha=0.3$; $\beta=0.7$; $w=0.2$; $c1=0.3$; $c2=0.5$; $t_0=30$; $t_f=0.1$; $\lambda=0.9$; and the parameters for GA with 2-opt: Crossover rate: 0.75; Mutation rates: (Swap: 0.05, Insertion: 0.15, Inversion: 0.01); Population size and termination condition are set according to the problem scale respectively

Table 1 The results of experiments for the instances and comparison

Problem	N	K	Branch and Cut	GA		hybrid PSO	
			distance	distance	time	distance	time
A-n33-k5	32	5	661	661	39.6	661	31.9
A-n46-k7	42	7	914	928	136.4	914	126.3
B-n35-k5	34	5	955	955	46.9	955	36.6
B-n45-k5	44	4	751	762	129.3	751	118.7
E-n30-k3	29	3	534	534	30.5	534	28.7
E-n51-k5	50	5	521	531	289.6	525	273.8
E-n76-k7	75	7	682	697	498.7	688	485.2
F-n72-k4	71	4	237	246	468.5	244	388.5
P-n76-k4	75	4	593	605	528.4	602	488.9
P-n101-k4	100	4	681	706	1213.2	694	952.3
M-n101-k10	100	10	820	836	992.1	824	862.9
M-n121-k7	120	7	1034	1068	1643.1	1038	1547.7

Branch, cut, and price (BCP) is an LP-based branch and bound technique for solving mixed integer linear programs (MILPs). In BCP, both cuts and variables are generated dynamically throughout the search tree, allowing the solution of large-scale instances, but they need more time than generic algorithm(GA) as the website stated.

In our experiments' results as table 1, N is the number of customers, K is the number of vehicles. It can be observed that the results of the proposed hybrid PSO algorithm are very close to the best known solution as Branch and Cut methods. Even if for the large-scale problem, the solutions of hybrid PSO are also very good. The computational speed of an application is also an important means of measuring the ability of an algorithm. Therefore, running time of each algorithm is reported. Through comparison, we can find that the speed of proposed hybrid PSO is better than that of the other two methods for most of the test instances. Therefore, our proposed algorithm is a feasible and effective approach for solving the VRP.

5 Conclusion

In this paper, we propose a new hybrid optimization approach combining discrete particle swarm optimization and simulated annealing to solve the vehicle routing problem. The performance of the approach is evaluated for comparison with the results obtained by other methods for a number of benchmark instances. Although the global optimality cannot be guaranteed, the performance of the results is relatively good. Moreover, the proposed algorithm is efficient in running time, which illuminated the validity and advantage of the proposed method.

References

[1] Baker, B.M., Ayechew, M.A.. A genetic algorithm for the vehicle routing problem. Computers & OperationsResearch, 2003,30(5):787-800

[2] Bodin, L., Golden, B.L., Assad, A., Ball, M.O.. The state of the art in the routing and scheduling of vehicles and crews. Computers & Operations Research, 1983,10:69-221

[3] CHEN Ai-ling et al. Hybrid discrete particle swarm optimization algorithm for capacitated vehicle routing problem. Journal of Zhejiang Univ SCIENCE A. 2006 7(4):607-614

[4] Eberhart, R., Kennedy, J., A New Optimizer Using Particle Swarm Theory. Proceeding of the Sixth International Symposium on Micro Machine and Human Science, 1995,pp.39-43

[5] Kennedy, J., Eberhart, R.. Particle Swarm Optimization. Proceeding of the 1995 IEEE International Conference on Neural Network, 1995, pp.1942-1948

[6] Laporte, G.. The vehicle routing problem: an overview of exact and approximate algorithms. European Journal of Operational Research, 1992, 59(3):345-358

[7] Osman, I.H.. Metastrategy simulated annealing and tabu search algorithms for the vehicle routing problem. Annals of Operations Research, 1993, pp.1945-1950

[8] Salman, A., Ahmad, I., Madani, S.A. Particle swarm optimization for task assignment problem. Microprocessorsand Microsystems, 2002, 26 (8): 363- 371

[9] Shi, Y., Eberhart, R.. Empirical Study of Particle Swarm Optimization. Proceedings of Congress on Evolutionary Computation, 1999,41(4):421-451

[10] Wang, Z.Z., Cheng, J.X., Fang, H.G., Qian, F.L. A hybrid optimization algorithm solving vehicle routing problems. Operations Research and Management Science, 2004,13:48-52 (in Chinese)

LS-SVM with Forgetting Factor and its Application*

Zhaoqing Song

Department of Control Engineering, Naval Aeronautical and Astronautical University, Yantai,
Shandong, 264001, China
Email: szqln@163.com

Abstract

The Least Square Support Vector Machine (LS-SVM) is an important method of Support Vector Machine (SVM), but the method cannot be used for online identification, and will lead to the problem of computation inflation. An online LS-SVM with forgetting factor is presented by combining the method of LS-SVM with forgetting factor. The presented method considers the effect of the historical data and emphasizes the effect of the current data. The computation speed of LS-SVM is hastened and the identification precision of LS-SVM is improved. In the end, the presented method is applied to the modeling of chaotic time series. The simulation example verifies the validity of the presented method.

Keywords：Support Vector Machine; Least Square; Forgetting Factor; rectangle windows; regression

1 Introduction

Support Vector Machine (SVM) was first presented by Vapnik in 1995[1] and received double recognition from the aspects of theory studies and engineering applications in recent years. SVM based on the method of Kernel learning was regarded as a very popular method and successful example in the field of machine learning, and is a very compelling development direction [2]. The standard training algorithm of SVM can be expressed as to solve the quadratic programming (QP) problem with linear inequality restriction [3]. Suppose we wish to perform regression from a training set $\boldsymbol{S}=\left\{\boldsymbol{s}_i \middle| \boldsymbol{s}_i=(\boldsymbol{x}_i,y_i),\boldsymbol{x}_i\in\boldsymbol{R}^n,y_i\in R\right\}_{i=1}^{l}$, the form of the regression function can be expressed as following

$$y(\boldsymbol{x})=\boldsymbol{w}\cdot\boldsymbol{\varphi}(\boldsymbol{x})+b \tag{1}$$

where $\boldsymbol{\varphi}(\boldsymbol{x})$ is the feature mapping, $\boldsymbol{w}$ and b are the regression parameters to be solved. The regression problem of standard SVM is to solve the minimum problem described in the following

$$\begin{cases}\min\boldsymbol{Q}(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*)=\dfrac{1}{2}\|\boldsymbol{w}\|^2+\gamma\sum\limits_{i=1}^{l}(\boldsymbol{\xi}+\boldsymbol{\xi}^*)\\ \text{s.t.}\quad y_i-\boldsymbol{w}\cdot\boldsymbol{\varphi}(\boldsymbol{x}_i)-b\le\varepsilon+\xi_i\\ \boldsymbol{w}\cdot\boldsymbol{\varphi}(\boldsymbol{x}_i)-b-y_i\le\varepsilon+\xi_i^*\\ \xi_i,\xi_i^*\ge 0,i=1,2,\cdots,l\end{cases} \tag{2}$$

where $\boldsymbol{\xi}=(\xi_1,\xi_2,\cdots,\xi_l)^T$ and $\boldsymbol{\xi}^*=(\xi_1^*,\xi_2^*,\cdots,\xi_l^*)^T$, γ is the predefined punished coefficient. $\varphi(\boldsymbol{x}_i)$ satisfies the following conditions

$$\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_j)=K(\boldsymbol{x}_i,\boldsymbol{x}_j),i,j=1,2,\cdots,l \tag{3}$$

where $K(\cdot,\cdot)$ is the kernel function.

Along with the increment of the problem scale, the QP problem spends more time and space on computation. When the scale of the training set is very big, the scale of the QP problem is too big to be solved. Inasmuch as this, researchers presented many improved methods about the learning algorithm of SVM. As a result, many new algorithms were presented. Among of these, the LS-SVM is the most famous one.

2 Least Square Support Vector Machine

The LS-SVM is a transmutation algorithm in essential. The main thinking of the transmutation

* Foundation item: Postdoctoral Science Foundation of China (20070410359)

algorithm is to transmute the equation by adding function items or variables or coefficients, and to form diversiform algorithms with some advantage.

Suykens et al transformed the linear inequality restriction of the minimum problem of the standard SVM to a group of linear equality restriction creatively so as to the training of SVM be equivalent to the solving of a group of linear equations. This method is called the LS-SVM [4]. The LS-SVM advances the solving efficiency and reduces the learning difficulty and accelerates the applications of SVM [5-7].

The LS-SVM method presented by Suykens can be described as the solving of the following minimum problem

$$\begin{cases} \min Q(\boldsymbol{w},\boldsymbol{e}) = \dfrac{1}{2}\|\boldsymbol{w}\|^2 + \dfrac{\gamma}{2}\sum_{i=1}^{l}\boldsymbol{e}_i^2 \\ \text{s.t.} \quad y_i = \boldsymbol{w}\cdot\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i, i = 1,2,\cdots,l \end{cases} \tag{4}$$

where $\boldsymbol{e} = (e_1, e_2, \cdots, e_l)^T$. The Lagrangian function of the minimum problem（4）can be described as following

$$\begin{aligned} \boldsymbol{L}(\boldsymbol{w},b,\boldsymbol{e},\boldsymbol{a}) &= \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{\gamma}{2}\sum_{i=1}^{l}e_i^2 \\ &-\sum_{i=1}^{l}a_i(\boldsymbol{w}\cdot\varphi(\boldsymbol{x}_i) + b + e_i - y_i) \end{aligned} \tag{5}$$

where $\boldsymbol{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_l \end{bmatrix}^T$. According to the equilibrium condition of Eq.（4）, we can know that

$$\begin{cases} \dfrac{\partial \boldsymbol{L}}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{l} a_i\varphi(\boldsymbol{x}_i) = 0 \\ \dfrac{\partial \boldsymbol{L}}{\partial b} = -\sum_{i=1}^{l} a_i = 0 \\ \dfrac{\partial \boldsymbol{L}}{\partial e_i} = \gamma e_i - a_i = 0 \\ \dfrac{\partial \boldsymbol{L}}{\partial a_i} = \boldsymbol{w}\cdot\varphi(\boldsymbol{x}_i) + b + e_i - y_i = 0 \end{cases} \tag{6}$$

From Eq.（6）, we can know that

$$\boldsymbol{w} = \sum_{i=1}^{l} a_i\varphi(\boldsymbol{x}_i) \, , e_i = \frac{1}{\gamma}a_i$$

If we cancel $\boldsymbol{w}$ and e_i in Eq.（6）, we can get a group of linear equations as following

$$\begin{bmatrix} 0 & \bar{\mathbf{1}}^T \\ \bar{\mathbf{1}} & \boldsymbol{Z}\boldsymbol{Z}^T + \gamma^{-1}\boldsymbol{I} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{a} \end{bmatrix} = \begin{bmatrix} 0 \\ \boldsymbol{y} \end{bmatrix} \tag{7}$$

where $\boldsymbol{Z} = \begin{bmatrix} \varphi(\boldsymbol{x}_1) & \varphi(\boldsymbol{x}_2) & \cdots & \varphi(\boldsymbol{x}_l) \end{bmatrix}^T$,

$\boldsymbol{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_l \end{bmatrix}^T$, $\bar{\mathbf{1}} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}^T \in R^l$,

$\boldsymbol{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_l \end{bmatrix}^T$.

Introduction the notation $\boldsymbol{\Omega} = \boldsymbol{Z}\boldsymbol{Z}^T$, $\Omega_{ij} = K(\cdot,\cdot)$, then the item $\boldsymbol{\Omega} + \gamma^{-1}\boldsymbol{I}$ in Eq.（7）is called a kernel interrelated matrix. If we let $\boldsymbol{A} \equiv \boldsymbol{\Omega} + \gamma^{-1}\boldsymbol{I}$, then the Eq.（7）is equal to the following

$$\begin{bmatrix} \mathbf{0} & \bar{\mathbf{1}}^T \\ \bar{\mathbf{1}} & \boldsymbol{A} \end{bmatrix} \begin{bmatrix} b \\ \boldsymbol{a} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{y} \end{bmatrix} \tag{8}$$

From Eq.（8）we can get that

$$b = \frac{\bar{\mathbf{1}}^T \boldsymbol{A}^{-1}\boldsymbol{y}}{\bar{\mathbf{1}}^T \boldsymbol{A}^{-1}\bar{\mathbf{1}}} \tag{9}$$

$$\boldsymbol{a} = \boldsymbol{A}^{-1}(\boldsymbol{y} - b\bar{\mathbf{1}}) \tag{10}$$

According to Eq.（9）and（10）and $\boldsymbol{w} = \sum_{i=1}^{l} a_i\varphi(\boldsymbol{x}_i)$, we can get the regressive function as following

$$\begin{aligned} y(\boldsymbol{x}) &= \boldsymbol{w}\cdot\boldsymbol{\varphi}(\boldsymbol{x}) + b \\ &= \sum_{i=1}^{l} a_i\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}) + b \\ &= \sum_{i=1}^{l} a_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \end{aligned} \tag{11}$$

We call $\boldsymbol{a}$ and b the parameters of the regressive function. From Eq.（11）, we can see that the key problem depends on the calculation of the inverse of the kernel interrelated matrix, *namely*, $\boldsymbol{A}^{-1}$.

When we set eyes on the advantages of the LS-SVM, we should also pay attention on the disadvantages of the LS-SVM. For example, the number of the support vector of LS-SVM algorithm is equal to the length of the training set. Then, for the LS-SVM algorithm, it doesn't possess the properties of sparseness and robustness any longer. The dimension of Ω in the LS-SVM algorithm is equal to the length of the training set, *namely*, the number of the elements in Ω is equal to the square of the length of the training set. On condition like this, a big training set means a huge calculation burden and maybe leads to the problem of computation inflation.

According to the problem of LS-SVM lacking of the properties of sparseness and robustness, some improved algorithms were presented, for example [8-13].

In this paper, an online LS-SVM with forgetting factor is presented. The online LS-SVM has a quick calculating speed and can solve the problem of computation inflation of the LS-SVM. The presented online LS-SVM with forgetting factor is a rectangle windows algorithm in essential. It considers the effect of the historical data and emphasizes the effect of current data and has more advantages than the standard LS-SVM.

3 LS-SVM with Forgetting Factor

In fact, the rectangle windows algorithms consider some finite past data only, and those before these data are eliminated entirely. Suppose the training samples set is $\boldsymbol{S}=\left\{\boldsymbol{s}_i \middle| \boldsymbol{s}_i=(\boldsymbol{x}_i,y_i),\boldsymbol{x}_i\in\boldsymbol{R}^n,y_i\in R\right\}_{i=1}^{l}$, the width of the rectangle window is m, then the learning samples set at k time can be described as $\{\boldsymbol{x}(k),\boldsymbol{y}(k)\}$, where $\boldsymbol{x}(k)=\left[\boldsymbol{x}_{k-m+1}\ \boldsymbol{x}_{k-m+2}\ \cdots\ \boldsymbol{x}_k\right]$, $\boldsymbol{y}(k)=\left[y_{k-m+1}\ y_{k-m+2}\ \cdots\ y_k\right]^T$, $\boldsymbol{x}_k\in\boldsymbol{R}^n$, $y_k\in R$. The kernel matrix Ω, the parameter $\boldsymbol{\alpha}$ and b to be solved can be described as followings

$$\boldsymbol{\Omega}_k(i,j)=K(\boldsymbol{x}_{k\text{-}m+i},\boldsymbol{x}_{k\text{-}m+j}),i,j=1,2,\cdots,m \qquad (12)$$

$$\boldsymbol{\alpha}(k)=\left[\alpha_{k-m+1}\quad\alpha_{k-m+2}\quad\cdots\quad\alpha_k\right]^T \qquad (13)$$

$$b(k)=b_k \qquad (14)$$

then the output of the regressive LS-SVM at time k is

$$y_k=\sum_{i=k-m+1}^{k}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i)+b(k) \qquad (15)$$

Let $\boldsymbol{Q}_k=\boldsymbol{\Omega}_k+\boldsymbol{I}/\gamma$, we have

$$\begin{bmatrix}\boldsymbol{0} & \bar{\boldsymbol{1}}^T\\ \bar{\boldsymbol{1}} & \boldsymbol{Q}_k\end{bmatrix}\begin{bmatrix}b(k)\\ \boldsymbol{a}(k)\end{bmatrix}=\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{y}_k\end{bmatrix} \qquad (16)$$

Let $\boldsymbol{P}_k=\boldsymbol{Q}_k^{-1}$, we can get $\boldsymbol{\alpha}(k)$ and $b(k)$ as followings

$$b(k)=\frac{\bar{\boldsymbol{1}}^T\boldsymbol{P}_k\boldsymbol{y}_k}{\bar{\boldsymbol{1}}^T\boldsymbol{P}_k\bar{\boldsymbol{1}}} \qquad (17)$$

$$\boldsymbol{a}(k)=\boldsymbol{P}_k(\boldsymbol{y}_k-\frac{\bar{\boldsymbol{1}}^T\boldsymbol{P}_k\boldsymbol{y}_k\bar{\boldsymbol{1}}}{\bar{\boldsymbol{1}}^T\boldsymbol{P}_k\bar{\boldsymbol{1}}})=\boldsymbol{P}_k(\boldsymbol{y}_k-b_k\bar{\boldsymbol{1}}) \qquad (18)$$

The Eq.（17）and（18）are called the rectangle windows algorithm. This method abandons the past data simply and doesn't consider the effect of the past data. The basic thinking of the forgetting factor method is to add forgetting factor on the past data. The method considers the effect of the past data and emphasizes the effect of current data [14].

Define $\boldsymbol{\theta}_k$, Φ_k and $\boldsymbol{z}_k$ as followings

$$\boldsymbol{\theta}_k=\begin{bmatrix}b_k\\ \boldsymbol{\alpha}(k)\end{bmatrix} \qquad (19)$$

$$\Phi_k=\begin{bmatrix}\boldsymbol{0} & \bar{\boldsymbol{1}}^T\\ \bar{\boldsymbol{1}} & \boldsymbol{Q}_k\end{bmatrix} \qquad (20)$$

$$\boldsymbol{z}_k=\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{y}_k\end{bmatrix} \qquad (21)$$

Suppose the function $J_k(\boldsymbol{\theta})$ can be expressed as a quadratic function

$$J_k(\boldsymbol{\theta})=(\boldsymbol{\theta}-\boldsymbol{\theta}_k)^T\boldsymbol{M}_k^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_k)+\beta_k \qquad (22)$$

Let's choose the index function for parameter estimation as following

$$\begin{aligned}J_{k+1}(\boldsymbol{\theta})&=\alpha J_k(\boldsymbol{\theta})+(\boldsymbol{z}_{k+1}-\Phi_{k+1}\boldsymbol{\theta})^T(\boldsymbol{z}_{k+1}-\Phi_{k+1}\boldsymbol{\theta})\\&-\alpha(\boldsymbol{z}_{k+1}-\Phi_{k+1}\boldsymbol{\theta}_k)^T(\alpha\boldsymbol{I}+\Phi_{k+1}\boldsymbol{M}_k\Phi_{k+1})^{-1}(\boldsymbol{z}_{k+1}-\Phi_{k+1}\boldsymbol{\theta}_k)\end{aligned} \qquad (23)$$

where $0<\alpha<1$ is called forgetting factor. $\boldsymbol{M}_k^{-1}$ and m are both positive definite symmetry matrixes. The rectangle windows recursive algorithm of LS-SVM with forgetting factor can be described as followings

$$\boldsymbol{\theta}_{k+1}=\boldsymbol{\theta}_k+\boldsymbol{N}_{k+1}(\boldsymbol{z}_{k+1}-\Phi_{k+1}\boldsymbol{\theta}_k) \qquad (24)$$

$$\boldsymbol{N}_{k+1}=\boldsymbol{M}_k\Phi_{k+1}(\alpha\boldsymbol{I}+\Phi_{k+1}\boldsymbol{M}_k\Phi_{k+1})^{-1} \qquad (25)$$

$$\boldsymbol{M}_{k+1}=\frac{\boldsymbol{M}_k}{\alpha}-\frac{\boldsymbol{M}_k}{\alpha}\Phi_{k+1}(\alpha\boldsymbol{I}+\Phi_{k+1}\boldsymbol{M}_k\Phi_{k+1})^{-1}\Phi_{k+1}\boldsymbol{M}_k \qquad (26)$$

Proof: According to Eq.（22）and（23）, we have

$$\begin{aligned}J_{k+1}(\boldsymbol{\theta})&=\alpha(\boldsymbol{\theta}-\boldsymbol{\theta}_k)^T\boldsymbol{M}_k^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_k)\\&+(\boldsymbol{z}_{k+1}-\Phi_{k+1}\boldsymbol{\theta})^T(\boldsymbol{z}_{k+1}-\Phi_{k+1}\boldsymbol{\theta})\\&-\alpha(\boldsymbol{z}_{k+1}-\Phi_{k+1}\boldsymbol{\theta}_k)^T(\alpha\boldsymbol{I}+\Phi_{k+1}\boldsymbol{M}_k\Phi_{k+1})^{-1}\\&\times(\boldsymbol{z}_{k+1}-\Phi_{k+1}\boldsymbol{\theta}_k)+\alpha\beta_k\end{aligned} \qquad (27)$$

Let's introduce the following notation

$$\boldsymbol{M}_{k+1}^{-1}=\alpha\boldsymbol{M}_k^{-1}+\Phi_{k+1}\Phi_{k+1} \qquad (28)$$

$$\boldsymbol{\theta}_{k+1}=\boldsymbol{M}_{k+1}(\alpha\boldsymbol{M}_k^{-1}\boldsymbol{\theta}_k+\Phi_{k+1}\boldsymbol{z}_{k+1}) \qquad (29)$$

$$\beta_{k+1}=\alpha\beta_k \qquad (30)$$

Applying the lemma of matrix generalized inverse on Eq.（28）leads to

$$Y=[y_i]^T \ i\in[1,m] \qquad (31)$$

Substituting the Eq.（31）into Eq.（29）yields to

$$\boldsymbol{\theta}_{k+1}=\boldsymbol{\theta}_k+\boldsymbol{M}_k\Phi_{k+1}(\alpha\boldsymbol{I}+\Phi_{k+1}\boldsymbol{M}_k\Phi_{k+1})^{-1}(\boldsymbol{z}_{k+1}-\Phi_{k+1}\boldsymbol{\theta}_k) \qquad (32)$$

Let

$$\boldsymbol{N}_{k+1}=\boldsymbol{M}_k\Phi_{k+1}(\alpha\boldsymbol{I}+\Phi_{k+1}\boldsymbol{M}_k\Phi_{k+1})^{-1} \qquad (33)$$

then we have

$$\boldsymbol{\theta}_{k+1}=\boldsymbol{\theta}_k+\boldsymbol{N}_{k+1}(\boldsymbol{z}_{k+1}-\Phi_{k+1}\boldsymbol{\theta}_k) \qquad (34)$$

According to the Eq.（28）to（30）and（32）to（34）, we have

$$J_{k+1}(\boldsymbol{\theta})=(\boldsymbol{\theta}-\boldsymbol{\theta}_{k+1})^T\boldsymbol{M}_{k+1}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_{k+1})+\boldsymbol{\beta}_{k+1} \qquad (35)$$

Obviously, $\boldsymbol{M}_{k+1}^{-1}$ is a positive definite symmetry matrix. The initial parameter $\boldsymbol{M}_0$ is defined as $\boldsymbol{M}_0=c^2\boldsymbol{I}$ and c is a large sufficiently real number. Thus the proof is completed. We can see that the regressive LS-SVM rectangle windows recursive algorithm with forgetting factor is an online recursive algorithm.

4 Simulation Example

Consider the Mackey-Glass chaotic time series, the Mackey-Glass chaotic time series is described as following

$$\frac{d(x(t))}{dt}=\frac{0.2x(t-\tau)}{1+x^{10}(t-\tau)}-0.1x(t) \qquad (36)$$

where τ is the delay time of the series and $\tau\geq 17$. When $\tau=17$, Mackey-Glass chaotic time series is shown in Figure 1.

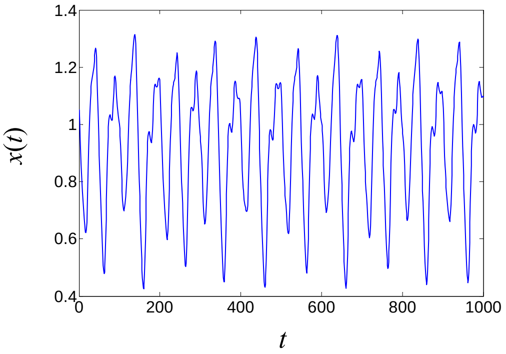


Figure 1 Mackey-Glass chaotic time series

The LS-SVM with forgetting factor (24-26) is adopted to set up the model of the chaotic time series, $\gamma=10000$, and the kernel function is taken as $K(x,y)=\exp(-\frac{\|x-y\|}{2\sigma^2})$, $\sigma=0.25$, the forgetting factor is taken as $\alpha=0.08$. The identified error is shown in Figure 2.

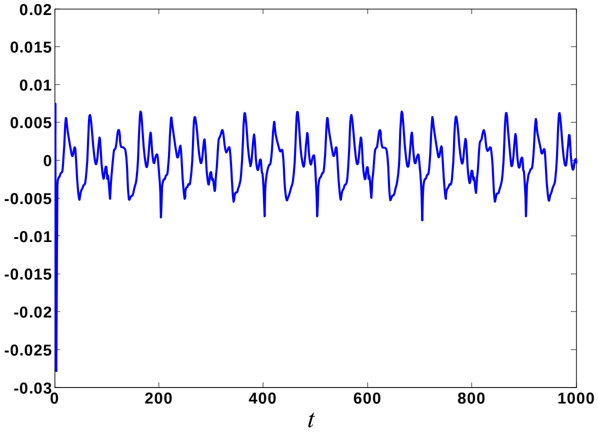


Figure 2 The identified error

The identified mean square error is 1.0492×10^{-5}. From the simulation result, we can see that a better simulation precision can be gained by using the LS-SVM with forgetting factor.

5 Conclusions

The LS-SVM translates the solving of SVM from QP problem to a group of linear equations and advances the solving efficiency and reduces the learning difficulty and accelerates the applications of SVM. But at the same time, the LS-SVM maybe results in computation inflation and can’t be used for online identification. This paper presented an online LS-SVM algorithm with forgetting factor by combining the forgetting factor method of LS with the SVM. The presented method considers the effect of the past data and predigests the computation of LS-SVM. In the end, the presented method is applied to the modeling of chaotic time series and a better result is gained. The simulation example verifies the validity of the presented method.

References

[1] V.N.Vapnik. The Nature of Statiscal Learning Theory. New York: Springer-Verlag, 1995

[2] Nello Cristinaini, John Shawe-Yaylor. An Introduction to Support Vector Machine.Beijing: China Machine Press, 2005

[3] V.N. Vapnik. Statistical Learning Theory. Wiley,1998

[4] J.A.K. Suykens, J. Vandewalle. Least squares support vector machine classifiers. Neural Processing Letters (1999) 9: 293~300

[5] T.V. Gestel, J.A.K. Suykens, et al. Benchmarking Least Squares Support Vector Machine Classifiers. Machine Learning,2004,54: 5~32

[6] D. Anguita1 and A. Boni. Digital Least Squares Support Vector Machines. Neural Processing Letters, (2003) 18: 65~72

[7] D. Tsujinishi, S. Abe. Fuzzy least squares support vector machines for multiclass problems. Neural Networks, (2003) 16: 785~792

[8] J.A.K. Suykens, L. Lukas, et al. Sparse approximation using least squares support vector machines. In Proc. Of the IEEE International Symposium on Circuits and Systems (ISCAS 2000), Geneva, Switzerland, (2000): 757~760

[9] Y.G.Fan,P.Li,Z.H.Song. Dynamic Weighted Least Squares Support Vector Machine.Control and Decision,2006,21（10）:1129-1134

[10] C.G.Wu. Study on Generalized Chromosome Genetic Algorithm and Iterative Least Squares Support Vector Machine Regression.. University of Jilin, 2006

[11] M.Y.Ye, X.D.Wang, H.R.Zhang. Chaotic time series forecasting using online least squares support vector machine regression. ACTA Physical Sinica, 2005,54（6）: 2568-2573

[12] H.R.Zhang and X.D.Wang. Incremental and Online Learning Algorithm for Regression Least Squares Support Vector Machine.Chinese Journal of Computers,2006, 29(3):400-406

[13] D.C.Wang and B.Jiang. Online sparse least square support vector machines regression. Control and Decision,2007,22（2）:132- 137

[14] H.F.Wang and D.J.Hu. Sparse Least Squares Support Vector Machines.Computer Engineering and Applications, 2005,33:68-70

Research of Decision Tree Arithmetic Based on Privacy Preservation*

Qiuyu Zhang   Luning Zheng

College of Computer and Communication, Lanzhou University of Technology, Lanzhou, Gansu, 730050, China

Email:zhangqy@lut.cn, zln1208@mail2.lut.cn

Abstract

As privacy becoming a more and more important topic in the data mining field, the technique of data mining regarding privacy preservation has developed quickly. It is different from the exiting technique of data interference and safety computing. Avoiding betraying the original data, after analyzed the exiting technique of privacy preservation and traditional decision tree computing, we bring forward a method that uses probability instead of original data in the construction of decision tree. It was validated to be a method that supports high veracity in classification, so that it meets the requirement in privacy preservation.

Keywords：Privacy preservation; Data mining; Decision tree; Probability

1 Introduction

Data mining[1] has been widely used in finance, medication and retail businesses. However, data owners are facing problems such as: they don't know how to manage the data, they look forward to professional data analyzing and mining institution to manage their data and from which provide them wanted information. On the other hand, they don't want their original data to be betrayed to the data analyzing institution[2]. Because the data is tied up with the privacy of the company or personal information (for instance, the trading information of a company or personal ID information), once this data is used unlawfully, it could lead to trouble or economy loss for the company or persons.

With the demand for special privacy-preserving method, the data mining is spring out, and becoming the hot topic for research.

In the data mining, there are two main category of information for privacy preservation:

1) preserving the original data information such as identify user's ID information (name, phone number and address etc.).

2) Preserving the rules after mined information, such as the rule used for determine customers' credit standing for bank.

2 The Privacy Preservation Method

At present, there are two main methods aim at data mining arithmetic based on privacy preservation[3]: random interference method and safety multimode computing. In simple words, the random interference method is a method that adds random measure to the data that needs protection. For instance, x is the data needs protection and r is the random noise which meets certain distribution. Provide $x+r$ to a data-mining instrument for data mining, if it is known the distribution of original data could be re-distributed to some extent. Safety multimode computing is used where there is distributed database. The original data was in possession of few people, but each of them only

* The Natural Science Foundation of Gansu (3ZS062-B25-037)

possesses mutually exclusive part. It is necessary to combine all those parts for data mining.

However, each person doesn't want others to know what data he has, so when conveying information to each other, only common interested data is wish to be computed and private data will not be betrayed[1,2].

For example, the participants input x_i, only need calculate $\sum_{i=1}^{n} x_i$, but do not want to be leaked their input x_i. In this article, it mainly discusses how to protect the concentrate distributed data.

3 The Traditional Decision Tree

Decision tree[4] is a method that widely used in data miming and equipment study. The key structure of decision tree is based on the choice of division property. The division property is determined by information entropy. In other words, it will choose the property, which will make the information to be most profitable, to be the division property. Suppose data aggregation D is divided to k different category, named $c_1, c_2, \ldots, c_k$, and $T(c_i)$ is all the subject in category c_i, then the entropy of this data aggregation as following:

$$E(T) = \sum_{i=1}^{k} (-\frac{|T(c_i)|}{|T|} * \log \frac{|T(c_i)|}{|T|}) \qquad （1）$$

There are n properties in the data aggregation named $A_1, A_2, \ldots, A_n$. Of which A_i takes finite values $a_1, a_2, \ldots, a_n$, and these values divide the aggregation to p subsets. $T(a_i, c_j)$ means the property value is a_i and belongs to aggregation $c_{i.}$, A_i attributes of the decision tree classification for the conditional entropy:

$$E(T \mid A) = \sum_{i=1}^{p} \frac{|T(a_i)|}{|T|} * E(T(a_i)) \qquad （2）$$

The change of entropy called attribute A_i gain information for classification, namely:

$Gain(A) = E(T) - E(T \mid A)$

We choice A as the largest classification attributes for a_i, at the same time A is to let $E\ (T|A)$ minimum.

However, in order to calculate $|T(c_i)|$ and $|T(a_i)|$, it need to get information from original data, so that it is hard to prevent let out of data[5]. Here, there is a method for calculating $E\ (T|A)$ even without original data.

4 Decision Tree for Privacy Preservation

4.1 Improvement for Decision Tree Based on Privacy Preservation

Combine Eq.（1）and Eq.（2）, the result as following:

$$\begin{aligned}
E(T \mid A) &= \sum_{i=1}^{p} \frac{|T(a_i)|}{|T|} * E(T(a_i)) \\
&= \frac{1}{|T|} \sum_{i=1}^{p} [|T(a_i)| * \sum_{j=1}^{k} -\frac{|T(a_i, c_j)|}{|T(a_i)|} * \log(\frac{|T(a_i, c_j)|}{|T(a_i)|})] \\
&= \sum_{i=1}^{p} [\frac{|T(a_i)|}{|T|} * \sum_{j=1}^{k} -\frac{|T(a_i, c_j)|/|T|}{|T(a_i)|/|T|} * \log(\frac{|T(a_i, c_j)|/|T|}{|T(a_i)|/|T|})] \\
&= \sum_{i=1}^{p} [p(a_i) * \sum_{j=1}^{k} -\frac{p(a_i, c_j)}{p(a_i)} * \log(\frac{p(a_i, c_j)}{p(a_i)})] \\
&= \sum_{i=1}^{p} [p(a_i) * \sum_{j=1}^{k} -p(c_j \mid a_i) * \log(p(c_j \mid a_i))] \\
&= -\sum_{i=1}^{p} [p(a_i) * \sum_{j=1}^{k} p(c_j \mid a_i) * \log(p(c_j \mid a_i))]
\end{aligned}$$

As we can see, the calculation of $E\ (T|A)$ is relevant to $p(a_i)$ and $p(c_i|a_j)$. However, as the decision tree performing the calculation to certain information entropy, the data aggregation is getting smaller.

For instance, if 'humidity' is determined to be the division property, then the data aggregation is not all the data in 'outlook', instead it should be under condition 'outlook =sunny'. So the probability in the formula should change to:

$$\begin{aligned}
E(T|A) = -\sum_{i=1}^{p} [p(a_i | a_s = A_{sl}, \ldots, a_1 = A_{1m}) \\
* \sum_{j=1}^{k} p(c_j | a_i, a_s = A_{sl}, \ldots, a_! = A_{!m}) \\
* \log(p(c_j | a_i, a_s = A_{sl}, \ldots, a_1 = A_{1m}))]
\end{aligned}$$

Where, $a_1, a_2, \ldots, a_s$ are the parent node. Grandparent node is the root node.

Suppose the probability of certain property is only relevant to property of parent node, the probability of category property is relevant to both current property and parent node property, then; the above formula can be translated to:

$$\sum_{i=1}^{p}[p(a_i|a_s = A_{sl}) * \sum_{j=1}^{k} p(c_j|a_i, a_s = A_{sl}) * \log(c_j|a_i, a_s = A_{sl}))]$$

So, for the calculation, it is only need to know the property, probability of properties, and probability of property and category, while no need to know the original data. And the information entropy could be simplified as:

$$E(T|A) = -\sum_{i=1}^{p} p(a_i|a_s = A_{sl}) * \sum_{j=1}^{k} p(c_j|a_i, a_s = A_{sl}) * \log(p(c_j|a_i, a_s = A_{sl})))] \quad （3）$$

Thus, it can protect privacy and hide original data. According to above formula, it uses probability instead of original data. It becomes few probability matrixes. It only needs to send those probability matrixes to data analyzing institution.

It also can use decision tree method to reconstruct decision tree[6,7,8,9]. Suppose there are *k* properties, and the domain of each has n_i scatter values. So, in total there are $n=n_1 + n_2 + ... + n_k$ scattered property values.

$Matrix0 = \{a_1, a_2, \cdots, a_n\}$, let this matrix every value that attribute a value, $a_1 = A_{11}$, $a_2 = A_{12}$,..., $a_n = A_{knk}$.

$Matrix\ 1 = \{p(a_i)\}$, this matrix representation of the value of each attribute probability. $p(a_i) = |T(a_i)|/|T|$ ·

$Matrix2 = \{p(c_i)\}$, the matrix representation of each of the categories of probability.

$Matrix3 = \{p(c_i, a_j)\}$ ·

$Matrix4 = \{p(a_i, a_j)\}$, the matrix representation attribute a_i , the probability at the same time.

$Matrix\,5 = \{p(c_i|a_j, a_s)\}$.

We can obtain the calculation of conditional probability by the five matrixes.

$$Matrix3' = p(c_i|a_j = A_{jl}) = p(c_i, a_j = A_{jl}) \big/ p(a_j = A_{jl})$$

$$Matrix\ 4' = p(a_i|a_s = A_{sl}) = p(a_i, a_s = A_{sl}) \big/ p(a_s = A_{sl})$$

$$Matrix5' = p(c_j|a_i, a_s = A_{sl}) = \frac{p(c_j, a_i, a_s = A_{sl})}{p(a_i, a_s = A_{sl})}$$

Because cannot access to original data, it can not test if the samples are in the same category. Thus it is not realistic and time consuming to use all property to construct decision tree. It needs secateur. Thus it needs special treatment when determining if it is leaf node.

After determined a certain distribution property, first scan the probability:

$$Matrix3' = \{p(c_i, a_j)\}$$

$$Matrix\,5' = \{p(c_i|a_j, a_s)\}$$

Examine whether category probability of those property values once to be 0. If it is, then all the samples are in the same category and it is a leaf node.

Also, when the probability of a certain category appears to be heavy proportion, if it is more than 98%, then the node is a leaf node as well, otherwise, continues its distribution property.

4.2 Algorithm Illustration

Algorithm: Privacy preserving construction of a decision tree using ID3 algorithm

```
Require: R: Set of attributes to be considered,
    M: Set of Matrixes to be considered,
    C= {c1, c2,..., ck}: Set of possible categories,
    Vmax: a threshold given by person.
1: if R is empty then
2: Return a leaf node whose category is set to the dominant category among the Matrix3’ or Matrix5’
3: end if
4: if all $p(c_i|a_j,a_s)=0$ and $p(c_l|a_j,a_s)\neq 0$ ($i\neq l$)
5: Return a leaf node whose category is set to $c_l$
6: end if
7: if all $p(c_l|a_j,a_s)=\max\{ p(c_i|a_j,a_s)\} \geqslant V_{max}$
8: Return a leaf node whose category is set to $c_l$
9: end if
10: Determine the attribute A that best classifies the objects and assign it as the test attribute for the current tree node /* the conditional entropy value E(T|A) for each attribute A is obtained using Eq.（3）, and the attribute with the minimum E(T|A) value (i.e.,
```

the highest gain) is selected as the best attribute. */

11: Create a new node for every possible value a_i of A, and recursively call this method on it with $R'=(R-\{A\})$

4.3 Sample Analysis

Against the data in the following table, brief description of algorithm.

Table1 Attributes

DAY	Outlook	Temp.	Humidity	Windy	PLAY?
1	sunny	hot	high	false	No
2	sunny	hot	high	true	No
3	overcast	hot	high	false	Yes
4	rain	mild	high	false	Yes
5	rain	cool	normal	false	Yes
6	rain	cool	normal	true	No
7	overcast	cool	normal	true	Yes
8	sunny	mild	high	false	No
9	sunny	cool	normal	false	Yes
10	rain	mild	normal	false	Yes
11	sunny	mild	normal	true	Yes
12	overcast	mild	high	true	Yes
13	overcast	hot	normal	false	Yes
14	rain	mild	high	true	No

In the data pretreatment stage, carries on discretization processing to all continual attributes, causes each attribute value space separate is several values.

In this example, the attribute outlook value is sunny, overcast, rain. The humidity value is high, normal. The windy value is true, false. The temp value is hot, mild, and cool. Use of the data in the table according to the method proposed in this paper, the original data will be hidden for the probability matrix structure decision tree, the results are the following:

$$Matrix0 = [sunny, overcast, rain, hot, mild, cool, high, normal, ture, false]$$

$$Matrix1 = [0.357, 0.286, 0.357, 0.286, 0.429, 0.286, 0.5, 0.5, 0.429, 0.571]$$

$$Matrix2 = [0.643, 0.357]$$

$$Matrix3 = \begin{bmatrix} 0.143, 0.286, 0.214, 0.143, 0.286, & 0.214, 0.214, 0.429, 0.214, 0.419 \\ 0.214, 0.000, 0.143, 0.143, 0.143, & 0.071, 0.286, 0.071, 0.214, 0.143 \end{bmatrix}$$

$$Matrix4 = (slightly)$$

$$Matrix\ 5 = (slightly\ )$$

Then by calculating, could be obtain Matrix3', Matrix4', Matrix5'.

After calculation, four attributes of information gain are 0.2337, 0.2372, 0.2690, 0.2819, while the root node is attribute '*outlook*'.

When 'outlook = overcast', $p(no|overcast) = 0$, $p(yes|overcast) = 1$, so generate leaf node, and marking its category for 'Yes'.

When 'outlook = rain', the remaining three attributes of information gain are 0.2624, 0.2615, 0.1651.

When 'outlook = sunny', the remaining three attributes of information gain are 0.2073, 0.1651, 0.2615.

So on the conditions of the division property is 'humidity'. In accordance with this method, and ultimately the decision tree to be:

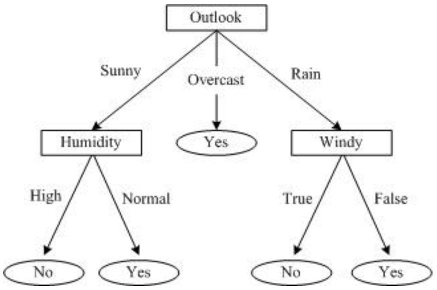


Figure 1 The original data hidden for the probability matrix structure decision tree

5 Model Evaluation

The veracity of categorizing is crucial in the evaluation of decision tree. The traditional decision tree is using original data to construct the tree, and using testing result to evaluate the categorizing veracity of the tree[10,11].

In this article, it didn't use the original data in constructing the tree. So, when evaluating the model, in addition to the testing data aggregation, the original data aggregation is used as well. Because here is using the probability matrix to construct the tree, there will be some error in evaluation of the tree, it can’t be 100% same as using the original data. Even using testing data, it will still have some error. Using both error values as parameter in evaluating the model. The smaller the error is, the higher veracity the tree has.

According to the method mentioned above, we constructed the decision tree by the processed data matrix.

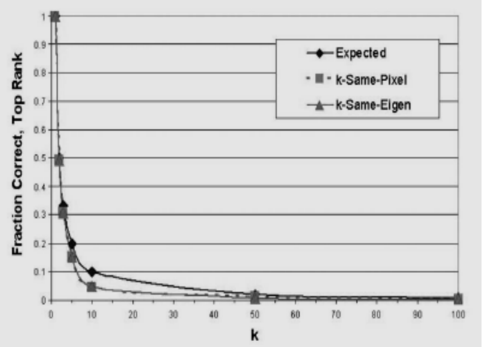


Figure 2 Evaluating the model

Then to evaluate tree by original data and testing data, it can be find that as the amount of the original data increasing, the veracity of the tree is higher.

6 Conclusions

The article mainly discussed using the probability in construction of the decision tree, which can hide the original data. Although it may lower the veracity of the tree, it avoids betraying the original data. Specially, when the amount of original data is huge and error is in the tolerable range, the difference can be ignored.

In this article mainly talked about the preservation of privacy for concentrate-distributed data. In reality, many cases data is saved in many computers. There is horizontal and vertical distribution of data. How to preserve those data will be the main topic for later research.

References

[1] Verykios VS, Bertino E, Fovino IN, et al. State of the art in Privacy Preserving Data Mining[J]. ACM SIGMOD Record, 33,2004, pp.50-57

[2] Vaidya J, Clifton C. Privacy pre serving association rule mining in vertically partitioned data[R]. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002, pp. 639–644

[3] Pin Lv, Nian-sheng Chen, and Wu-shi Dong. Study of Data Mining Technique in Presence of Privacy Preserving[J]. Computer Technology and Develop- ment, 16（7）, 2006, pp. 147-149

[4] Xiu-hong Ma, Jian-she Song. Research on Decision Tree in Data Mining[J]. Computer Project and Application, 1,2004, pp.185-214

[5] Hui-ping Lu, Xue-feng Tong. Elementary research on the building of privacy preserving decision tree[J]. Computer Application, 25（6）,2005, pp.1382-138

[6] Kantarcioglu M, Clifton C. Privacy-preserving distributed mining of association rules on horizontally partitioned data[C]. The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02), 2002

[7] Lindell Y, Pinkas B. Privacy preserving data mining[C]. Advances in Cryptology CRYPTO 2000, Springer-Verlag, 2001

[8] Du Wenliang, Zhan Zhijun. Building Decision Tree Classifier on Private Data[C]. Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining, 2002

[9] W.L.Du, Z.J.Zhan. Using randomized response techniques for privacy preserving data mining[A]. In Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and data Mining[C]. Washington D C, USA, 2003,505-510

[10] Z.Huang, W.Du, B.Chen. Deriving private information from randomized data[A]. SIGMOD 2005[C]. Baltimore, Maryland, USA: ACM SIGMOD, 2005, 27-48

[11] Yuan-yuan Lv, Xue-ming Shen. Decision Tree Building Based on Privacy preserving[J]. Computer Project, 32（3）,2006, pp.96-98

Parameter Estimation of Complex Functions Based on Quantum-behaved Particle Swarm Optimization Algorithm

Min Xu[1] Wenbo Xu[2]

1 School of Science, Jiangnan University, Wuxi, Jiangsu, 214122, China

Email: xm786@yahoo.com.cn

2 School of Information Technology, Jiangnan University, Wuxi, Jiangsu, 214122, China

Email: xwb@jiangnan.edu.cn

Abstract

This paper, quantum-behaved particle swarm optimization (QPSO) algorithm is developed for some serious disadvantages of traditional parameter estimation methods of complex functions in statistics. QPSO algorithm is an improved algorithm of particle swarm optimization algorithm (PSO). QPSO algorithm and PSO algorithm are detailed introduced and studied. A new method that using least-squares estimation of complex functions based on QPSO algorithm is developed. Several tests are made by computer. It indicates that QPSO algorithm can estimate parameters of complex functions correctly. It can calculate simply and constringe fast. Through comparing with traditional PSO algorithm, QPSO algorithm’s advantages are testified.

Keywords：Particle Swarm Optimization; Quantum-behaved Particle Swarm Optimization; Parameter Estimation

1 Introduction

In mathematical statistics, regression analysis is a commonly used method, in accordance with the variable data to analyze the relationship between variables. In return for these issues often need to estimate the parameters of the regression equation. However, there are many limitations when we use the traditional method of least-squares estimation. This method may be very difficult for some complicated functions and would not be able to form a standard equation. So the method of least-squares estimation will not be applied successfully. This paper introduced the quantum-behaved particle swarm optimization (QPSO) algorithm application to complex functions parameter estimation problems. Traditional particle swarm optimization (PSO) algorithm in the application of basic easily fall into local minima. QPSO algorithm enhances the overall search capability; it can converge faster and make more accurate estimations of the parameters.

2 Traditional method of parameter estimation

Under the assumption that the linear regression equation:

$$f(x_n, x_{n-1}, ..., x_1) = a_n x_n + a_{n-1} x_{n-1} + ... + a_1 x_1 + a_0 \quad （1）$$

If a known sample values: $X = [x_{ij}]$ $i \in [1, m]$; $j \in [1, n]$ (m is the number of samples, n is dimension of samples); observation function values of the samples:

$$Y = [y_i]^T \ i \in [1, m].$$

Using least-squares estimation to estimate the parameters of Eq.（1）, we need to calculate summation of sampling errors:

$$e = \sum (y_i - f(x_{in}, x_{i(n-1)}, ..., x_1))^2 \qquad （2）$$

Then seek partial derivative of the error equation in the following equations form:

$$\begin{cases} \dfrac{\partial \varepsilon}{\partial a_n} = 0 \\ \dfrac{\partial \varepsilon}{\partial a_{n-1}} = 0 \\ \dots \\ \dfrac{\partial \varepsilon}{\partial a_1} = 0 \end{cases} \qquad （3）$$

Solving this equation group, we can get a_i. It's the least- squares estimation of Eq.（1）.

At this point, summation of sampling errors e must have partial derivatives. Linear function normally under the partial derivative, we will be able to form standard matrix equation. Then we can go for the standard equation. But to some of the complex functions, it could be very difficult to calculate the partial derivatives. So we can not create standards equation. At this time, least-squares estimation can not succeed.

3 PSO algorithm

3.1 Classics PSO Algorithm

PSO algorithm is proposed by Kennedy J and Eberhart RC in 1995[1]. This approach stemmed from the study of biological groups. It is one of the heuristic optimization algorithms. The basic algorithm of PSO is: there is a given objective function. Under an acceptable amount of time and cost, through searching in a scheduled multidimensional space, every particle evaluates its own location information in every iteration. In this process, information about the optimal location is shared by all particles. Then they adjust their speed and position by constant comparison and following through optimal location. Eventually, the optimal solution or near optimal solution is found. [2][3]

The velocity and position of particles are constantly updated by the following formula:

$$\begin{aligned} V_i(t+1) &= V_i(t) + c_1 * rand_1 * (P_i - X_i(t)) \\ &\quad + c_2 * rand_2 * (P_g - X_i(t)) \end{aligned} \qquad （4）$$

$$X_i(t+1) = X_i(t) + V_i(t+1) \qquad （5）$$

Factor c_1 is called own learning factor. c_2 is called society learning factor. The two factors usually take value between $[0,2]$. They adjust the most step when particles flying to individual optimal particle and overall optimal particle. $rand_1$ and $rand_2$ each is a random number between $[0,1]$. Particle's speed is limited between $[-V_{\max}, V_{\max}]$ in order to prevent particles flying out of the solution space. p_i is the historical optimal location. g_i is global optimum position of all particles.

PSO algorithm was immediately attracted wide attention, and many researchers improve it. A. Chatterjee and Siarry invented adding dynamic and inertia factor [4] [5] into the algorithm. This method has become more classical PSO algorithm. The convergence of the algorithm is more accurate and more comprehensive.

In inertia weight method, the speed update equation is:

$$\begin{aligned} V_i(t+1) &= w * V_i(t) + c_1 * rand_1\ (P_i - X_i(t)) \\ &\quad + c_2 * rand_2\ (P_g - X_i(t)) \end{aligned} \qquad （6）$$

Here w is inertia weight to control the impact of the previous rate to current speed. Its formula:

$$w = \left[\frac{(iter_{\max} - iter)^n}{iter_{\max}{}^n} \right] * (w_{initial} - w_{final}) + w_{final} \qquad （7）$$

Here $iter_{\max}$ is the largest number of iterations. The number of iterations is determined by actual need. [6]

a) Quantum-behaved Particle Swarm Optimization Algorithm

As p_i is a constant as same as random numbers. Therefore, the PSO algorithm is a linear method. In fact, the thinking of individuals in the biological groups is very complex and uncertain. The phenomenon is just like particles have quantum behaviors. And the application of quantum theory in PSO is called QPSO algorithm [7] [8].

In QPSO, particles update their position information in accordance with the following formulas [9]:

$$mbest = \frac{\sum_{i=1}^{M} p_i}{M} = \left(\frac{\sum_{i=1}^{M} p_{i1}}{M}, \frac{\sum_{i=1}^{M} p_{i2}}{M}, \ldots\ldots, \frac{\sum_{i=1}^{M} p_{id}}{M}\right) \qquad （8）$$

$$p = (\varphi_1 * p_i + \varphi_2 * p_g)/(\varphi_1 + \varphi_2) \qquad （9）$$

$$X(t+1) = p \pm \beta * \left| mbest - X(t) \right| * \ln\left(\frac{1}{u}\right) \qquad （10）$$

Here $mbest$ is called middle value optimal location. M is the number of particles. p_i is the historical optimal location $pbest$. g_i is global optimum position of all particles $gbest$. φ_1 、φ_2 and u are all random numbers between $[0,1]$. β is a factor called creativity of coefficient. β is the only parameter in the algorithm.

The basic steps of QPSO algorithm are as follows:

1) Initialize the locations of particles $X(i)$ randomly and initialize the number of particles M .

2) Calculate value of fitness function according to optimal function. Compare the value with individual historical optimal value. If current value is better than prior historical optimal value, then replace it with historical optimal value. Otherwise, don't replace.

3) Evaluate the fitness of all particles, get $gbest$. Then calculate $mbest$. Update the information of particles according to Eq.（8）（9）（10）.

4) Judge whether reach accuracy or get the best fitness value, if it does not, then return to step（2）to continue. Otherwise, end iterative.

This shows that the advantage lies in QPSO algorithm, such as only one parameter, programming simply and converging fast.

4 Least-squares estimation of complex functions based on QPSO algorithm

In this paper, a new method is developed that using least-squares method to estimate parameters based on QPSO algorithm. Converge parameters to stable value through iterative. This method need not to go for Eq.（3）, and the form of estimated function is not limited. So it can be used to estimate complex function's parameters in the least -squares sense. Experiments show that this method is fast and effective. Through comparing with classics PSO algorithm, the experimental results show that calculation of QPSO is much simpler than PSO, program of QPSO is simpler, QPSO converges faster, and it is estimated more accurately by QPSO.

In experiment, assuming known samples $X_{m\times k}$ and corresponding value of the observation function $Y_{m\times l}$. Known from the expertise that the sample is a duty to observe in a multi-function mapping. The style of this function is known, now the parameter values are needed to be estimated.

Supposing the function which is to be estimated is $f(X_{1\times k})$ and its parameter vector is $A_{1\times n}$ .Then $A_{1\times n}$ is which to be estimated. Now setting there are n particles. The position of every particle is a q -dimensional vector $P_{n\times q}$, and then we use P_i as the estimated location of the particles $A_{1\times q}$. The goal is to let $f(X_{1\times k}|P_i)$ converge to optimal values, then the convergence value is the least-squares parameter estimates of $f(X_{1\times k})$ .Therefore the fitness function in particles optimization process is:

$$Fitness_i = \sum_{i=1}^{M}(Y_i - f(X_i|P_i))^2 \qquad （11）$$

So the least-squares estimation is completed when fitness function meets its minimum value. Here the form of function which to be estimated is not limited. A complex function form and multi-function are allowed as long as the function value can be calculated.

5 The test

Using above method under Windows XP and Matlab 7.1, we do several tests to testify. Particles' size is set to 200. In order to avoid random error, the algorithm runs 50 times. We take the average results of 50 times. β is reduced from 1.0 to 0.5. In Eq.（10）, $"\pm"$ is decided by the random number between $(0,1)$. When the random number is lager than 0.5, we take $"-$", others we take $"+"$.

Test function 1: $L_i = x_1 e^{ix_2}$. This is a function with two parameters x_1 and x_2 [10]. In order to analyze the results, we presume $x_1 = 5.42$ $x_2 = -0.25$. Now generate 50 independent variables between [50,100] randomly. Then calculate corresponding L value to simulate the observation of this function. Then by observing independent variable i and function value L , we use the least-squares estimation method based on the above QPSO algorithm to estimate two parameters x_1 and x_2 . After testing, the result is shown in Table 1.

Table 1 shows that when the numbers of iterations are same, QPSO algorithm is better. It can get more accurately result. The error is smaller. The parameters which are estimated are closer to the true value. And the value of fitness function is better.

Figure 1 shows the comparison of the randomly selected 50 samples dots and the function’s image using the estimation results of QPSO algorithm. Comparison chart shows than QPSO algorithm is feasible to estimate the parameters of the complex function and curve fitting accuracy is very high.

Figure 2 and Figure 3 show the parameters of the two algorithms convergence chart. They show that convergence speed of parameters is faster in QPSO than in PSO.

Table 1 algorithm results (test function 1)

Iterate 100	PSO		QPSO	
	Estimate parameter	Error rate (%)	Estimate parameter	Error rate (%)
x_1	5.4588	0.7159	5.4180	0.3333
x_2	－0.2501	0.0400	－0.2500	0.0000
Fitness value	9.1303e－016		4.0211e－017	

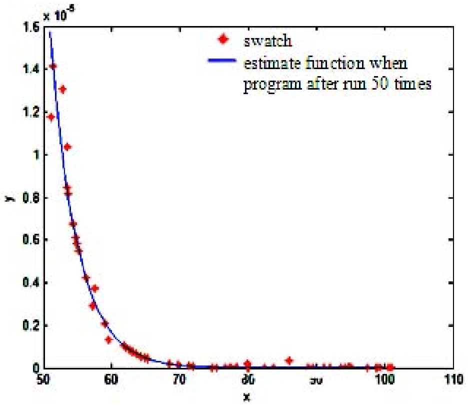


Figure 1 comparison of estimated function image using QPSO and the sample dots

convergent image of parameter x_1 and x_2

Test function 2: $y = P_1 + \frac{P_2}{1 + \ln(e^{P_3(x-P_4)})}$. This is a complex function with four parameters P_1 P_2 P_3 and P_4. Presume $P_1 = 100$ $P_2 = 200$ $P_3 = 0.3$ $P_4 = 75$. Then use the same method to complete the test. Table 2 and Table 3 give the results of the experiment. In the experiment, we respectively run PSO and QPSO algorithms on this function, and we respectively iterate 100 and 1000 times.

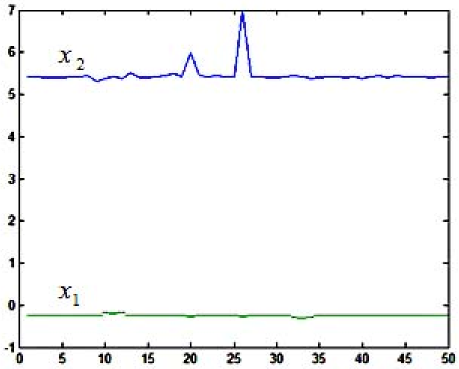


Figure 2 use PSO algorithm, convergent image of parameter $x_1$andx_2

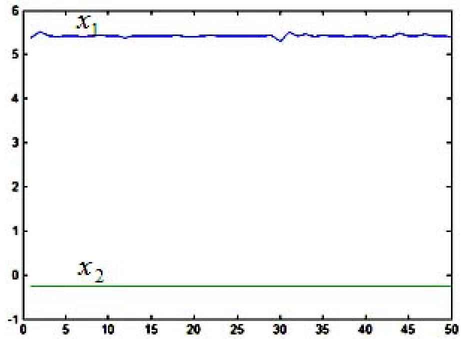


Figure 3 use QPSO algorithm

The results show that:（1）Under the same number of iterations, parameters estimated by QPSO algorithm are closer to the presumed value. Error rate is smaller. The fitness function is optimized.（2）When the number of iterations changes from 100 to 1000, parameters estimate have higher accuracy. Error is smaller and the fitness function is better.

Figure 4 shows the result of QPSO when the number of iteration is 1000. We can see that the function image is very closer to the image using sample dots. The result show that QPSO algorithm used to estimate the parameters is feasible and Curve Fitting more accurate.

Figure 5 and Figure 6 show the convergence trends of fitness function when the number of iteration is 1000. Comparison shows that QPSO algorithm is more accurate and converging faster.

Table 2 algorithm results when iterating 100 times (test function 2)

Algorithm results	PSO		QPSO	
	Estimate parameter	Error rate (%)	Estimate parameter	Error rate (%)
P_1	260.4634	160.4634	104.8713	4.8713
P_2	220.6671	10.3336	204.9841	2.4921
P_3	0.1977	34.1000	0.3065	2.1667
P_4	82.3656	9.8208	74.1601	1.1199
Fitness Average Value	2.019*e*+007		4.1114*e*+003	

Table 3 algorithm results when iterating 1000 times (test function 2)

Algorithm results	PSO		QPSO	
	Estimate parameter	Error rate (%)	Estimate parameter	Error rate (%)
P_1	100.4668	0.4668	100.0526	0.0526
P_2	189.4155	5.2923	199.8467	0.0767
P_3	0.2615	12.8333	0.2998	0.0667
P_4	73.0313	2.6249	74.2164	1.0448
Fitness Average Value	1.8342*e*+005		32.1571	

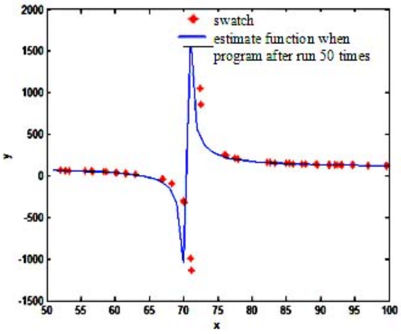


Figure 4 comparison of estimated function image using QPSO and the sample dots

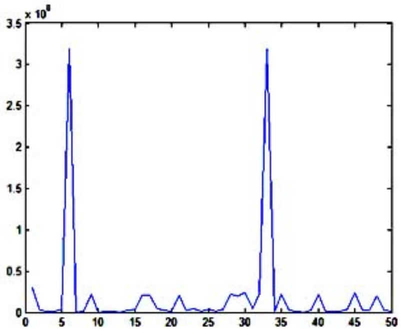


Figure 5 use PSO algorithm, convergent image of fitness function

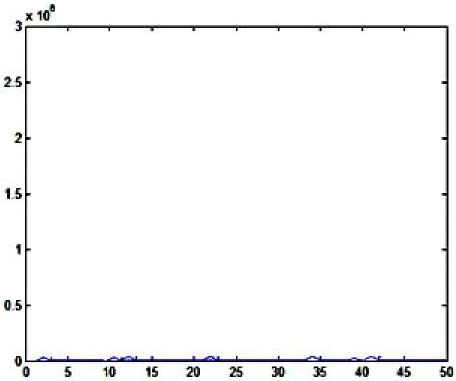


Figure 6 use QPSO algorithm, convergent image of fitness function

6 TAG

This paper presents a new method to estimate parameters of complex functions, including *e* transcendental functions. The method is based on QPSO algorithm. Comparing with traditional PSO algorithm, there are several advantages. Such as it calculates much simpler, it is easier to program, it converge faster and it has stronger global search capability. The parameters estimated are more accurate. Tests show the effectiveness of the algorithm.

References

[1] Kennedy J, Eberhart RC, “Particle swarm optimization”, Proceedings of the IEEE International Conference on Neural Networks, Vol. IV, Piscataway, NJ: IEEE Service Center, 1995, pp. 1942-1948

[2] AG LI, Z JIA, FM BAO, “ Particle Swarm Optimization Algorithm”, Computer Engineering and Applications, Vol. 38, No.21, 2002, pp.1-3

[3] XF XIE, WJ ZHANG,ZK YANG, ”A Summary of Particle Swarm Optimization Algorithm”, Control and Decision, Vol.18, No.2, 2003, pp.129-134

[4] A. Chatterjee, P. Siarry. “Nonlinear inertia weight variation for dynamic adaptation in particle swarm optimization”, Computers and Operations Research, Vol.33, No.3, 2006, pp. 859-871

[5] Shi Y, Eberhart R C. “Empirical study of particle swarm optimization”, Proceedings of the IEEE Congress on

Evolutionary Computation. Piscataway, NJ: IEEE Press, 1999, pp.1945-1950

[6] JCh ZENG,Q JIE, ZhH CUI, Particle Swarm Optimization Algorithm, Beijing: Science Press,2004

[7] J SUN, WB XU, "A Global Search Strategy of Quantum-behaved Particle Swarm Optimization", Proceedings of IEEE Conference on Cybemetics and Intelligent Systems, 2004, pp.111-116

[8] J SUN, B FENG, WB XU, "Particle Swarm Optimization with Particles Having Quantum Behavior", Proceedings of 2004 Congress on Evolutionary Computation, 2004, pp. 325-331

[9] WB XU, JB JIANG, J SUN, "Multi-Stage portfolio optimization using QPSO", Computer Applications,Vol.26, No.7,2006,pp.1682-1685

[10] XZh Wang, Theory and Application of Nonlinear Model Parameter Estimation, Wuhan: Wuhan University Press, 2002

Solving Traveling Salesman Problem by Genetic Ant Colony Optimization Algorithm [*]

Shang Gao[1,2]

1 School of electronics and information, Jiangsu University of Science and Technology,
Zhenjiang, Jiangsu 212003, China

2 State Key Laboratory of CAD＆CG, Zhejiang University ,Hangzhou, Zhejiang ,310027,China

Email: gao_shang@hotmail.com

Abstract

By use of the properties of ant colony algorithm and genetic algorithm, a hybrid algorithm is proposed to solve the traveling salesman problems. First, it adopts genetic algorithm to give information pheromone to distribute. Second, it makes use of the ant colony algorithm to get several solutions through information pheromone accumulation and renewal. Finally, by using across and mutation operation of genetic algorithm, the effective solutions are obtained. Compare with the simulated annealing algorithm, the standard genetic algorithm, the standard ant colony algorithm, and statistics initial ant colony algorithm, all the 16 hybrid algorithms are proved effective. Especially the hybrid algorithm with across strategy B and mutation strategy B is a simple and effective better algorithm than others.

Keywords：Ant Colony Algorithm; Genetic Algorithm; Traveling Salesman Problem; Simulated Annealing Algorithm

1 Introduction

Inspired by the behavior of real ants, Marco Dorigo first introduced the colony optimization approach in his Ph.D. thesis in 1992 and expanded it in his further work. The characteristics of artificial ant colony include a method to construct solutions that balances pheromone trails and a problem-specific heuristic, a method to both reinforce and evaporate pheromone, and local search to improve the constructed solutions. The ACO[1] methods have been successfully applied to diverse combinatorial optimization problems including traveling salesman, quadratic assignment, vehicle routing[2], telecommunication networks[3], graph coloring, constraint satisfaction, Hamitonian graphs and scheduling. Genetic algorithms (GAs) or more generally, evolutionary algorithms [4] have been touted as a class of general-purpose search strategies for optimization problems. GAs use a population of solutions, from which, using crossover, mutation and selection strategies, better and better solutions can be produced. GAs can handle any kind of objective functions and any kind of constraints without much mathematical requirements about the optimization problems, and have been widely used as search algorithms in various applications. Various GAs have been proposed in the literature [5,6] and shown superior performances over other methods. As a consequence, GAs seemed to be nice approaches for solving TSP. However, GAs may cause certain degeneracy in search performance if their operators are not carefully designed [6]. A genetic algorithm (GA) is a metaheuristic inspired by the efficiency of natural selection in biological evolution. Genetic algorithms have been applied successfully to a wide variety of combinatorial optimization problems and are the subject of numerous recent books [7-8] and conference proceedings. Unlike traditional heuristics (and some metaheuristics like tabu

[*] Project supported by State Key Laboratory of CAD＆CG, Zhejiang University

search) that generate a single solution and work hard to improve it, GAs maintain a large number of solutions and performcomparatively little work on each one. Several researchers (see [9] and the references contained within) have implemented GAs for the standard TSP, with mixed results. The GA in [9] found new best solutions for some well studied benchmark problems. Recently, there are many search activities over artificial ants, which are agents with the capability of mimicking the behavior of real ants [10,11]. The agents are sufficiently intelligent to exploit pheromone information that has been left on the traversed ground. Agents can then choose a route according to the amount of pheromone. The larger amount of pheromone is on a route, the larger is the probability of selecting the route by agents. With such concept, a population-based algorithm, Ant Colony Optimization (ACO), has been widely used as a new cooperative search algorithm [10]. In this paper, a novel algorithm of genetic ant colony optimization (GACO) for traveling salesman problem is proposed.

2 The basic ACO algorithm

In this section we introduce the basic ACO algorithm. We decided to use the well-known traveling salesman problem as benchmark, in order to make the comparison with other heuristic approaches easier. Given a set of *n* towns, the TSP can be stated as the problem of finding a minimal length closed tour that visits each town once. We call d_{ij} the length of the path between towns *i* and *j*. In the case of Euclidean TSP, d_{ij} is the Euclidean distance between *i* and *j* (i.e., $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. An instance of the TSP is given by a graph (*N*, *E*), where *N* is the set of towns and *E* is the set of edges between towns (a fully connected graph in the Euclidean TSP).

Let $b_i(t)\,(i = 1, 2, \cdots, n)$ be the number of ants in town i at time t and let $m = \sum_{i=1}^{n} b_i(t)$ be the total number of ants. Each ant is a simple agent with the following characteristics:

- it chooses the town to go to with a probability that is a function of the town distance and of the amount of trail present on the connecting edge.
- to force the ant to make legal tours, transitions to already visited towns are disallowed until a tour is completed (this is controlled by a *tabu* list).
- when it completes a tour, it lays a substance called trail on each edge (i, j) visited.

Let $\tau_{ij}(t)$ be the intensity of trail on edge (i, j) at time *t*. Each ant at time *t* chooses the next town, where it will be at time $t+1$. Therefore, if we call an iteration of the ACO algorithm the m moves carried out by the m ants in the interval $(t, t+1)$, then every *n* iterations of the algorithm (which we call a cycle) each ant has completed a tour. At this point the trail intensity is updated according to the following formula

$$\tau_{ij}(t+n) = \rho\tau_{ij}(t) + \Delta\tau_{ij} \qquad (1)$$

where ρ . is a coefficient such that (1- ρ) represents the evaporation of trail between time *t* and $t+n$,

$$\Delta\tau_{ij} = \sum_{k=1}^{m} \Delta\tau_{ij}^{k} \qquad (2)$$

where $\Delta\tau_{ij}^{k}$ is the quantity per unit of length of trail substance (pheromone in real ants) laid on edge (i, j) by the k-th ant between time *t* and $t+n$. It is given by

$$\Delta\tau_{ij}^{k} = \begin{cases} \dfrac{Q}{L_k} & \text{if k- th ant uses edge (i, j)} \\ & \text{in its tour} \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

where Q is a constant and L_k is the tour length of the k-th ant.The coefficient ρ must be set to a value ρ <1 to avoid unlimited accumulation of trail. In our experiments, we set the intensity of trail at time 0, $\tau_{ij}(0)$, to a small positive constant *c*.

In order to satisfy the constraint that an ant visits all the n different towns, we associate with each ant a data structure called the *tabu list*, that saves the towns already visited up to time t and forbids the ant to visit them again before n iterations (a tour) have been completed. When a tour is completed, the *tabu* list is used to compute the ant’s current solution (i.e., the distance of the path followed by the ant). The *tabu* list is then emptied and the ant is free again to choose. We define $tabu_k$ the dynamically growing vector, which contains the *tabu* list of the kth ant, $tabu_k$ the set obtained from the elements of $tabu_k$, and $tabu_k(s)$ the s-th element of the list (i.e., the s-th town visited by the k-th ant in the current tour).

We call *visibility* η_{ij} the quantity $1/d_{ij}$. This quantity is not modified during the run of the ACO algorithm, as opposed to the trail, which instead changes according to the previous formula（1）.

We define the transition probability from town i to town j for the k-th ant as

$$p_{ij}^{k}(t)=\begin{cases}\dfrac{\tau_{ij}^{\alpha}(t)\cdot\eta_{ij}^{\beta}}{\sum\limits_{s\in allowed_k}\tau_{is}^{\alpha}(t)\cdot\eta_{is}^{\beta}} & if\ j\in allowed_k\\ 0 & otherwise\end{cases}\qquad(4)$$

where $allowed_k$ = $\{N\text{-}tabu_k\}$ and where α and β are parameters that control the relative importance of trail versus visibility. Therefore the transition probability is a trade-off between visibility (which says that close towns should be chosen with high probability, thus implementing a greedy constructive heuristic) and trail intensity at time *t* (that says that if on edge (i,j) there has been a lot of traffic then it is highly desirable, thus implementing the autocatalytic process).

3 The basic genetic algorithm

Genetic algorithms start with a set of randomly selected chromosomes as the initial population that encodes a set of possible solutions. In GAs, variables of a problem are represented as genes in a chromosome, and the chromosomes are evaluated according to their costs using some measures of profit or utility that we want to optimize. The recombination typically involves two genetic operators: crossover and mutation. The genetic operators alter the composition of genes to create new chromosomes called offspring. The selection operator is an artificial version of natural selection, a Darwinian survival of the fittest among populations, to create populations from generation to generation, and chromosomes with better cost have higher probabilities of being selected in the next generation. After several generations, GA can converge to the best solution. Let P(t) and C(t) are parents and offspring in generation t. A usual form of general GA is shown in the following:

```
Procedure: General GA
Begin
$t \leftarrow 0$;
Initialize P(t);
Evaluate P(t);
While (not match the termination conditions) do
Recombine P(t) to yield C(t);
Evaluate C(t);
Select P(t+1) form P(t) and C(t);
$t \leftarrow t+1$;
End;
End;
```

Recently, genetic algorithms with local search have also been considered as good alternatives for solving optimization problems. The local search for TSP, 2-opt approach, can be implemented after crossover and mutation operators

4 Hybrid algorithm

4.1 Modified ACO Algorithm

In traditional ACO Algorithm, the initialization of the pheromone matrix is equal. Ants need iterate many numbers to find the best tour. We can generate a large amount of tours (e. g. 100 tours), and then we choose some better tours (e. g. 30 tours). At last ants lay trail only on these better tours. These trails affect following ants.

When ant completes a tour, it always lays called trail on each edge visited. If the tour is worse, ant also lay the trail on each edge. These trails disturb the following ants, so the ACO algorithm’s convergent speed is very slow. We can calculate the length of tour firstly. and then we compare with the given value. If the length of tour is less than the given value, we update trail values. Otherwise don’t update the trail values.

4.2 Mutation operators

There are following methods to generate a new tour C_1 from the tour C_0.

Mutation operator A Choose two cities j_1 and j_2 from the tour C_0 by randomly, and then swap j_1 with j_2 in the tour C_0, so the new tour is C_1. For example, suppose C_0=2 3 4 1 5 7 9 8 6, $j_1=3$ and $j_2=9$, so C_1=2 9 4 1 5 7 3 8 6.

Mutation operator B Choose a city j_1 from the tour C_0 by randomly, and then swap j_1 with the next visited city. For example, suppose C_0=2 3 4 1 5 7 9 8 6, $j_1=3$, so C_1=2 4 3 1 5 7 9 8 6.

Mutation operator C A modified solution C_1 is generated from C_0 by randomly choose two cities j_1 and j_2 and reversing the sequence in which the cities in between cities j_1 and j_2 are traversed, i.e. the 2-change generation mechanism. For example, suppose C_0=2 3 4 1 5 7 9 8 6, $j_1=3$ and $j_2=9$, so C_1=2 9 7 5 1 4 3 8 6.

Mutation operator D Choose two cities j_1 and j_2 from the tour C_0 by randomly, and then insert city j_1 into the latter of j_2 city. For example, suppose C_0=2 3 4 1 5 7 9 8 6, $j_1=3$ and $j_2=9$, so C_1=2 4 1 5 7 9 3 8 6.

4.3 Crossover Operators

There are many different types of crossover operators, but we discuss several usual crossover operators as following. Let's suppose we have two parent tours given by

old1=1 2 3 4 5 6 7 8 9

old2=9 8 7 6 5 4 3 2 1.

Crossover operators A We will swap a substring from old2 that will be called the donor, and this substring is selected randomly. Then we insert the substring into the beginning (or the end) of old1. After the insertion of the substring we must delete the cities of the receptor that were included in the new substring added. For example the substring 6 5 4 3 of donor old2 is the one that will be inserted into the beginning of old1, and we delete the cities of the receptor that were included in the new substring added. After that, we get the configuration

new1=6 5 4 3 1 2 7 8 9.

Crossover operators B We will swap a substring from old2 randomly. Then we insert the substring into the matching section of old1. After the insertion of the substring we must delete the cities of the receptor that were included in the new substring added. For example the substring 6 5 4 3 is from donor old2, so after rearrangement we have new1=1 2 6 5 4 3 7 8 9.

Crossover operators C We will swap a substring from old2 randomly. Then we insert the substring into the random section of old1. After the insertion of the substring we must delete the cities of the receptor that were included in the new substring added. For example the substring 6 5 4 3 is from donor old2, and if city 7 is the random city of old1, so after rearrangement we have new1=1 2 7 6 5 4 3 8 9.

Crossover operators D The substring 6 5 4 3 is from donor old2 randomly, then we insert the substring into the position of city 6 of old1, and we delete the cities of the receptor that were included in the new substring added, so

new1=1 2 7 6 5 4 3 8 9.

4.4 Hybrid Algorithm

The hybrid algorithm solving TSP can be expressed as follows:

Step1. $nc \leftarrow 0$ （nc is iteration number）. Generate 100 tours, and choose the better 30 tours from these 100 tours, and pheromone laid on edge of these 30 better tours.

Step 2. Choose the next city j according to formula（4）.

Step 3. The j th tour $C_0(j)$ is required to crossover to generate an offspring $C_1(j)$ according to the mutation probability. Calculate the new evaluation values.

Step 4. Compute L_k $(k=1,2,\cdots,m)$ (L_k is the length of tour done by ant k). Save the current best tour.

Step 5. If the L_k is less the given value, update trail values according to formula（1）,formula（2）and formula（3）.

Step 6. $nc \leftarrow nc+1$.

Step7. If the iteration number nc reaches the maximum iteration number, then go to Step 9. Otherwise, go to Step 2.

Step 8. Print the current best tour.

5 Experimental results

This section compares the results of simulated annealing algorithm, genetic algorithm, ACO algorithm and hybrid algorithms on traveling salesman problem of 30 cities. The parameters of simulated annealing algorithm are set as follows: the initial temperature $T = 100000$, the final temperature $T_0 = 1$,and annealing velocity $\alpha = 0.99$. The parameters of the genetic algorithm optimization toolbox (GAOT) used to solving TSP are set as follows: the population $N = 30$, the cross probability $P_c = 0.2$, and the mutation probability $P_m = 0.5$. The parameters of the hybrid algorithms are set as follows: $\alpha = 1.5$, $m = 30$, $\beta = 2$,and $\rho = 0.9$. 20 rounds of computer simulation are conducted for each algorithm, and the results are shown in Table 1. The optimal tour of 30 cities by hybrid algorithm is shown in Figure 1. All the 16 hybrid algorithms are proved effective. Especially the hybrid algorithm with across strategy B and mutation strategy B is a simple and effective better algorithm than others. The crossover operators of particle swarm optimization are different from those of genetic algorithm. In hybrid optimization, the crossover operator happened between the individual with local optimum and global optimum, so the capability of offspring got to be improved.

Table 1 Testing Rresult of Algorithms

Algorithms	Average solutions	Best solutions	Worst solutions
Aimulated annealing algorithm	438.5223	424.6918	479.8312
Genetic algorithm	483.4572	467.6844	502.5742
Basic ACO algorithm	550.0346	491.9581	599.9331
Crossover operators A +Mutation operator A	439.4948	425.6490	456.7721
Crossover operators A +Mutation operator B	441.9257	428.7296	455.2382
Crossover operators A +Mutation operator C	437.0028	426.6002	446.2394
Crossover operators A +Mutation operator D	438.7750	425.4752	455.2929
Crossover operators B +Mutation operator A	438.9350	424.6354	457.9062
Crossover operators B +Mutation operator B	431.4987	423.7406	447.6865
Crossover operators B +Mutation operator C	435.4220	424.9003	447.3223
Crossover operators B +Mutation operator D	439.4777	426.1972	465.9935
Crossover operators C +Mutation operator A	444.1723	429.3803	459.4925
Crossover operators C +Mutation operator B	438.5871	426.3076	455.5854
Crossover operators C +Mutation operator C	440.4201	427.6016	454.8674
Crossover operators C +Mutation operator D	439.5524	424.4643	461.7948
Crossover operators D +Mutation operator A	439.0477	424.6727	451.8001
Crossover operators D +Mutation operator B	436.0081	423.7406	460.6230
Crossover operators D +Mutation operator C	438.8091	425.8201	455.4830
Crossover operators D +Mutation operator D	436.4577	423.9490	457.3155

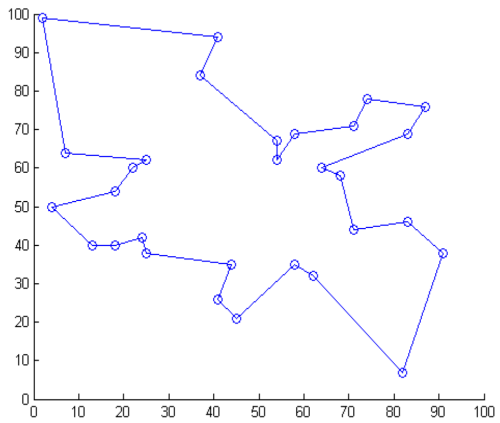


Figure 1 The optimal tour of 30 cities by hybrid algorithm

6 Conclusions

In this paper, we presented a novel algorithm of genetic ant colony optimization (GACO) for traveling salesman problem. It keeps the advantages of ant colony optimization and GAs. From our simulation for those test problems, the proposed algorithm indeed can find the best solutions or optimal solutions. In other words, the proposed algorithm seems to have admirable performance. Experiments for benchmark problems show the hybrid algorithm better than other algorithms.

The following problems need to be considered. 1. The parameters and their affect on the performance of the optimization should be studied in more detail. 2. How to explore hybrid algorithm application to continuous space problem should be investigated. 3. The hybrid algorithm's convergent speed, or the efficiency, should be worth further investigating. 4. How to evaluate the quality of hybrid algorithm and other algorithms is still a problem.

References

[1] A. Colorni, M. Dorigo, and V. Maniezzo, "An investigation of some properties of an ant algorithm", Proc. Of the Parallel Problem Solving from Nature Conference (PPSN'92). Brussels, Belgium: Elsevier Publishing,1992, pp.509-520

[2] Bernd Bullnheimer, F. Richard. Hartl, and Christine Strauss, "Applying the ant System to the vehicle routing problem". 2nd Metaheuristics International Conference (MIC-97). Sophia-Antipolis, France, 1997, pp.21-24

[3] Gianni Di Caro, and M. Dorigo, "Mobile agents for adaptive routing", Proceedings of the 31th Hawaii International Conference on system Sciences. Big Island of Hawaii,1998, pp.74-83

[4] T. Bäck, U. Hammel, and H. P. Schwefel, "Evolutionary computation: Comments on the history and current state," IEEE Trans. On Evolutionary Computation, Vol. 1, No. 1, 1997,pp.3-17

[5] M. Gen, and R. Cheng, Genetic Algorithms and Engineering Design, John Wiley & Sons Inc., 1997

[6] L. Jiao, and L. Wang, "Novel genetic algorithm based on immunity," IEEE Transactions on Systems, Man and Cybernetics, Part A, Vol. 30, No. 5, 2000, pp.552 –561

[7] K. F. Man, K. S. Tang, and S. Kwong, Genetic Algorithms: Concepts and Designs. Springer, New York, 1999

[8] G. Winter, J. Periaux, M. Galan, and P. Cuesta, editors. Genetic Algorithms in Engineering and Computer Science. Wiley, New York, 1995

[9] G. Laporte, A. Asef-Vaziri, and C. Sriskandarajah, "Some applications of the generalized travelling salesman problem," Journal of the Operational Research Society,Vol 47, No12,1996,pp.1461–1467

[10] M. Dorigo, and L. M. Gambardella, "Ant colony system: A cooperative learning approach to the traveling salesman problem," IEEE Trans. On Evolutionary Computation, Vol. 1,1997, pp.53-66

[11] R. Beckers, J. L. Deneubourg, and S. Goss, "Trails and u-turns in the selection of the shortest path by the ant Lasius Niger," Journal of Theoretical Biology,Vol. 159, 1992, pp.397-415

Training Support Vector Machines with Quantum-behaved Particle Swarms Optimization

Hui Li   Wenbo Xu   Jun Sun

College of Information, Jiangnan University, Wuxi, 214122

Email: lihui19992003@yahoo.com.cn

Abstract

Large number of example vectors brings difficulties for quadratic programmin g problem with support vector machines, traditional methods may be impossible.Quantum-behaved Particle Swarm Optimiz ation presented by the author is a new method of optimization. It is better than classical Particle Swarm Optimiza- tion(PSO for short) in convergence and stability of the overall. Testify QPSO has determinate applied value in the field of support vector machines, and it is a new way for quadratic programming problem with a large number of example vectors.

Keywords: SVM; QPSO; Classification; MNIST

1 Introduction

Support Vector Machines(SVMs)are a young and important addition to the machine learning toolbox.Having been formally introduced by Boser er al.[1],SVMs have proved their worth-in the last decade there has been a remarkable growth in both the theory and practice of these learning machines.

Training a SVM requires solving a linearly constrained quadratic optimization problem.This problem often involves a matrix with an extremely large number of entries ,which make off-the-shelf optimization packages unsuitable.In the last decade,many researchers paid much attention to optimization algorithm,using Evolution Strategies to solve a linearly constrained quadratic optimization problem.Based on the primitive PSO ,we propose a global-convergence-guaranteedPSO,Quantum-behaved Particle Swarm Optimization algorithm (QPSO).QPSO enhances the global search ability of PSO algorithm,has just only one parameter,easy to realize and to select the parameter,and is more stable than original PSO.So in this paper a QPSO is adapted and shown to be ideal in optimizing the SVM problem.

This paper gives an overview of the SVM algorithm, and explains the main methodologies for training SVMs.QPSO is discussed as an alternative method for solving a SVM’s quadratic programming problern. Experimental results on character recognition illustrate the convergence properties of the algorithms.

2 Support Vector Machines

Traditionally, a SVM is a learning machine for two-class classification problems, and learns from a set of l N-dimensional-example vectors x_i , and their associated classes y_i ,i.e.

$$\{x_1, y_1\}, \cdots \{x_l, y_l\} \in R^N \times \{\pm 1\} \qquad (1)$$

The algorithm aims to learn a separation between the two classes by creating a linear decision surface between them.This surface is,however,not created in input space,but rather in a very high-dimensional feature space.The resulting model is nonlinear,and is accomplished by the use of kernel functions.The kernel function k gives a measure of similarity between a pattern x,and a pattern x_i from the training set.The decision boundary that needs to be constructed is of the form

$$f(x)=\sum_{i=1}^{l} y_i\alpha_i k(x,x_i)+b \qquad (2)$$

where the class of x is determined from the sign of f(x).The α_i are Lagrange multipliers from a primal quadratic programming(QP) problem,and there is an α_i for each vector in the training set.The value b is a threshold. “Support vectors” define the decision surface, and correspond to the subset of nonzero α_i .These vectors can be seen as the “most informative” training vectors.

Training the SVM consists of finding the values of α_i .By defining a Hessian matrix H such that $(H)_{ij}=y_iy_jk(x_i,x_j)$, training can be expressed as a dual QP problem of solving

$$\min W(\alpha)=\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j k(x_i,x_j)-\sum_{i=1}^{l}\alpha_i \qquad (3)$$

subject to one equality constraint

$$\sum_{i=1}^{l} y_i\alpha_i=0 \qquad (4)$$

and a set of box constrains

$$0\le\alpha_i\le C\ i=1,2,3,\ldots,l \qquad (5)$$

Training a SVM thus involves solving a linearly constrained quadratic optimization problem.

3 SVM training methods

The QP problem is equivalent to finding the maximum of a constrained bowl-shaped objective function.Due to the definition of the kernel function,the matrix H always gives a convex QP problem,where every local solution is also a global solution[2].Certain optimality conditions-the Karush-Kuhn-Tucker(KKT) conditions[2]-give conditions determining whether the constrained maximum has been found.

Solving the QP problem for real-world problems can prove to be very difficult.The matrix H has a dimension equal to the number of training examples. A training set of 60,000 vectors gives rise to a matrix H with 3.6 billion elements, which does not fit into the memory of a standard computer.

For large learning tasks, off-the-shelf optimization packages and techniques for general quadratic programming quickly become intractable in their memory and time requirements.A number of other approaches, which allow for fast convergence and small memory requirements, even on large problems, have been invented:

Chunking

The chunking algorithm is based on the fact that the non-support vectors play no role in the SVM decision boundary.If they are removed from the training set of examples, the SVM solution will be exactly the same

Decomposition

Decomposition methods are similar to chunking, and were introduced by E.Osuna[9],.The large QP problem is broken down into a series of smaller subproblems,and a numeric QP optimizer solves each of these problem,It was suggested that one vector be added and one removed from the subproblem at each iteration,and that the size of the subproblems should be kept fixed.The motivation behind this method is based on the observation that as long as at least one α_i violating the KKT conditions is added to the previous subproblem,each step reduces the objective function and maintains all of the constrains.In this fashion the sequence pratical convergence,researchers add and delete multiple examples.

Sequential Minimal Optimization

The most extreme case of decomposition is Sequential Minimal Optimization (SMO)-where the smallest possible optimization problem is solved at each step.Becauce the α_i must obey the linear equality constrain,two α_i is chosen to jointly optimized.No numerical QP optimization is necessary ,and after an analytic solution,the SVM is updated to reflect the new optimal values.

With the exception of SMO, a numeric QP library is needed for training a SVM.An intuitive and alternative approach is to use PSO to optimize each decomposed subproblem.The QPSO algorithm is easy to implement, and certain properties of the QPSO make it ideal for the type of problem posed by SVM training.

4 Quantum-Behaved particle swarm optimization

Quantum-behaved Particle Swarm Optimization (QPSO)was proposed[4][5] by Sun et al,it is a new algorithm model based on PSO. Quantum-behaved Particle Swarm Optimization algorithm that outperforms traditional PSOs in search ability as well as having less parameter to control.In QPSO algorithm.

Evolution formula of the particle is:

$$mbest = 1/M\sum_{i=1}^{M} P_i = (1/M\sum_{i=1}^{M} P_{i1,} \cdots, 1/M\sum_{i=1}^{M} P_{id}) \quad (6)$$

$$\mathrm{PP_{id}} = \phi \times \mathrm{P_{id}} + (1-\phi)\times \mathrm{P_{gd}} \ \phi = \mathrm{rand} \quad (7)$$

$$\mathrm{x_{id}} = \mathrm{PP_{id}} \pm \alpha \times \left|\mathrm{mbest_d} - \mathrm{x_{id}}\right| \times \ln(1/\mathrm{u}) \ u = rand \quad (8)$$

Where, mbest is the middle position of the particle swarm(pbest); $\mathrm{PP_{id}}$ is the random point between P_{id} and P_{gd} , α is the only parameter of the QPSO algorithm. Commonly let $\alpha = (1.0-0.5)\times(MAXITER - T)/MAXITER + 0.5$ Where T is the current number of iterations, $MAXITER$ is the maximum number of iterations.

The QPSO algorithm flow:

Set the iteration number T to zero .Initialize the swarm.

Evaluate the performance $\mathrm{f(p_i^{(t)})}$ of each particle.

Evaluate new $\mathrm{p_{id}}$ of each particle.

Evaluate new $\mathrm{p_{gd}}$.

Evaluate mbest by formula（6）.

Evaluate the random point $\mathrm{PP_{id}}$ of each particle by equation（7）.

Move each particle to its new position ,according to equation（8）.

1) Make T=T+1,go to step 2,and repeat until terminal condition satisfied.

5 Training the SVM

Using QPSO to solve the SVM QP problem requires criteria for optimality,a way to decompose the QP,and a way to extend QPSO to optimize the SVM subproblem.

Since H is a positive semi-define matrix (the kernel function used is positive semi-define),and the constriains are linear,the Karush-Kuhn-Tucker (KKT)conditions are necessary and sufficient for optimality[2].A solution α of the QP problem,as stated in equalities（3）-（5）,is an optimal solution if the following relations hold for each α_{i} :

$$\alpha_{\mathrm{i}} = 0 \Leftrightarrow \mathrm{y_i f(x_i)} \ge 1$$

$$0 < \alpha_{\mathrm{i}} < \mathrm{C} \Leftrightarrow \mathrm{y_i f(x_i)} = 1$$

$$\alpha_{\mathrm{i}} = \mathrm{C} \Leftrightarrow \mathrm{y_i f(x_i)} \le 1 \quad (9)$$

where i is the index of an example vector from the training set.

Decomposing the QP problem involves choosing a subset, or "working set", of variables for optimization. The working set,called set B, is created by picking q sub-optimal variables from all l α_{i} .The working set of variables is optimized while keeping the remaining variables(called set N)constant.The general decomposition algorithm works as follows:

While the KKT conditions for optimality are violated:

Select q variables for the working set B.The remaining l-q variables(set N)are fixed at their current value.

Use QPSO to optimize $\mathrm{W}(\alpha)$ on B.

Return the optimized α_{i} from B to the original set of variables.

Terminate and return α .

A concern in the above algorithm is to select the optimal working set.The decomposition method presented here is due to [3],and works on the method of feasible directions.The idea is to find the steepest feasible direction d of ascent on the objective function W as defined in equation（3）,under the requirement that only q components be nonzero.The α_{i} corresponding to these q components will be included in the working set.Finding an approximation to d is equivalent to solving

Maximise $\nabla \mathrm{W}(\alpha)^{\mathrm{T}} \mathrm{d}$

Subject to $\mathrm{y^T d} = 0$

$\mathrm{d_i} \ge 0$

if

$\alpha_{\mathrm{i}} = 0$

$d_i \le 0$

if

$\alpha_i = C$

$d_i \in \{-1, 0, 1\}$

$\left|\{d_i : d_i \ne 0\}\right| = q$

For $y^T d$ to be equal to zero,the number of elements with sign matches between d_i and y_i must be equal to the number of elements with sign mismatches between d_i and y_i .Also, d should be chosen to maximize the direction of ascent $\nabla W(\alpha)^T d$ .This is achieved by first sorting the training vectors in increasing order according to $y_i \nabla W(\alpha)_i$ .Assuming q to be even a "forward pass" selects $q/2$ examples from the front of the sorted list,and a "backward pass" selects $q/2$ examples from the back.A full explanation of this method is given by P.Laskov.

It is necessary to rewrite the objective function（3） as a function that is only dependent on the working set.Let α be split into two sets $\alpha_B$andα_N .If α , y and H are appporpriately rearranged,we have

$$\alpha = \begin{pmatrix} \alpha_B \\ \alpha_N \end{pmatrix}, \quad Y = \begin{pmatrix} Y_B \\ Y_N \end{pmatrix}, \quad H = \begin{pmatrix} H_{BB}\, H_{BN} \\ H_{NB}\, H_{NN} \end{pmatrix}$$

Since only α_B is going to be optimized.W is rewritten in terms of α_B.If terms that do not contain α_B are dropped,the optimization problem remains essentially the same.Also,since H is symmetric,with " ,the problem is to find:

$$\min_{\alpha_B} W(\alpha_B) = 1/2\alpha_B{}^T H_{BB}\alpha_B - \alpha_B{}^T(e - H_{BN}\alpha_N) \quad (10)$$

Subject to

$$\alpha_B{}^T y_B + \alpha_N{}^T = 0 \;\; \alpha_B \ge 0 \;\; C1 - \alpha_B \ge 0 \quad (11)$$

Implementing Quantum-behaved Particle Swarm Optimization

When the decomposition algorithm starts, a feasible solution that satisfies the linear constraint $\alpha^T y = 0$,with constraints $0 \le \alpha_i \le C$ also met, is needed.The initial solution is constructed in the following way:

Let c be some real number between 0 and C,and γ some positive integer less than both the number of positive examples ($y_i = +1$)and egative examples ($y_i = -1$)in the training set.Randomly pick a total of γ positive examples,and γ negative examples,and initialize their corresponding α_i to c. By setting all other α_i to zero, the initial solution will be feasible.

The value 2γ gives the total number of initial support vectors,and since these initial support vectors are a randomly chosen guess,it is suggested that the value of γ be kept small.

In optimizing the q-dimensionalsub problem,

QPSO requires that all particles be initialized such that $\alpha_B{}^T y_B + \alpha_N{}^T y_N = 0$ is met .This is done as follows:

Set each particle in the swarm to the q-demensional vector α_B.

Add a random q-demensional vector δ satisfying $y_B{}^T \delta = 0$ to each particle, under the condition that the particle will still lie in the hypercube $[0, C]^q$.

Initializing the swarm in this way ensures that the initial swarm lies in the set of feasible solutions $P = \{P \mid AP = -\alpha^T{}_N y_N\}$, allowing the flight of the swarm to be defined by feasible directions.

For faster convergence, the vector $v^{(t)}$ used to adjust the global best particle, can be chosen as an approximation to the partial derivative $\nabla W(\alpha_B)$, subject to $y_B{}^T v^{(t)} = 0$.

6 Experimental results

The SVM training algorithm presented in this paper was tested on the MNIST dataset[7].The MNIST dataset is a database of optical characters,and consists of a training set of 60,000 handwritten digits.Each digit is a 28 by 28 pixel gray-level image,equivalent to a 784-dimensional input vector .Each pixel corresponds to an integer value in the range of 0(white) to 255(back)

For training a SVM on the MNIST dataset, the character'8'was used to represent the positive examples, while the remaining digits defined the negative examples.Training was done with the kernel function: $k(x, x_i) = \exp(-\frac{\| x - x_i \|^2}{1.0^2})$

For an optimal solution to be found in the following PSO experiments, the KKT conditions needed to be satisfied within an error threshold of 0.02.Optimization of the working set terminated when the KKT conditions on the working set were met within an error of 0.001, or when the swarm has optimized for five hundred iterations.

The following parameters defined the experimental QPSO: By letting γ =10,a total of 20 initial support vectors were chosen to start the algorithm. The value of Contraction-Expansion Coefficient α is set to 0.7, along with iteration increases; the value of α linearily reduces to 0.3, so

$$\alpha = (0.7 - 0.3) * (\text{ MAXITER} - \text{T}) / \text{MAXITER} + 0.3$$

For each experiment the upper bound C was kept at 100.0.

Experimental results show successful and accurate training on the MNIST database.The influence of different working set sizes on the QPSO training algorithm, its scalability, as well as its relation to other SVM training algorithm were examined.

Influence of working set sizes

Experiments on different working set sizes were done on the first 20.000 elements of the MNIST database.Results are shown in Table I,and indicate that a working set of size q=4 gives the fastest convergence time and fewest support vectors.

Tabale 1 Inpluence of different working set sizes on the irst 20.000 elements of the mnist dataset

Working set size	Working set selections	Time	SVs
4	8,782	01：47：23	1,631
6	8,213	02：31：40	1,598
8	7,502	02：46：17	1,651
10	10,023	04：40：11	1,701
12	9,667	06：15：23	1,653

Scalability of the QPSO approach

Table 2 Scalability:training on the mnist dataset

MNIST elements	QPSOWorking Set selections	QPSO time	QPSO SVs	PSO time	PSO SVs	SMO Time	SMO SVs
10.000	3.898	00:01:26	1.012	00：29：49	1,022	00：01：29	1,032
20.000	8.782	00:05:46	1.564	02：17：43	1,631	00：06：14	1,647
30.000	12.428	01:50:21	1.798	04：50：11	1,988	00：13：22	2,012
40.000	15.725	03:14:26	2.234	08：14：26	2,353	00：22：46	2,355
50.000	22.727	07:50:23	2.134	15：05：09	2,728	01：46：38	2,740
60.000	25.914	10:11:21	3.001	20：54：15	3,025	04：38：11	3,043

Scalability of the QPSO algorithm was tested by training on the first 10.000,20.000,etc.examples from the MNIST dataset,as shown in Table II.In each case a working set of size 4 was used .The experimental results indicate that the QPSO training algorithm shows quadratic scalability,and scales as □ $l^{2.1}$.

In Table II, the QPSO approach is compared to PSO and SMO. when solves the QP problem, the PSO algorithm is inferior to the QPSO algorithm obviously in the training time.

The experiment demonstrated that the QPSO algorithm trains SVM has a better effect. In the experiment also observes, the kernel function's selection and the penalty parameter C choice has the important influence to the experimental result.

7 Conclusion

The experimental result indicated that QPSO has determinate applied value in the field of support vector machines, and it is a new way for quadratic programming problem with a large number of example vectors.

References

[1] B.E. Boser.I.M.Guyon,and V.N.Vapnik,” A training algorithm for optimal margin classifiers,” in D.Haussler, editor, Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, pages144-152, Pitsburgh, PA,1992,ACM Press

[2] R.Fletcher,Practical Methods of Optimization.John Wiley and Sons,Inc.2nd edition.1987

[3] T.Joachims." Making large-scale SVM learning practical," in Advances in Kernel Methods-Support Vector Learning, B. Scolkopf, C.J,C Rurges, and A.J.Smola.editors. pages 169-184. MITPress, Cambridge, MA, 1999

[4] Sun, J. and Xu W.B.. A Global Search Strategy of Quantum-behaved Particle Swarm Optimization[C]. Proceedings of IEEE conference on Cybernetics and Intelligent Systems, 2004, 111 – 116

[5] Sun, J. Feng B. and Xu W.B.Particle Swarm Optimization with Particles Having Quantum Behavior[C]. Proceedings of 2004 Congress on Evolutionary Computation, 2004, 325-331

[6] P.Laskov." Feasible direction decomposition algorithms for training support vector machines," in MachineLearning Volume46, N.Cristianini, C.Campbell,and Chris Burges, editors, pages315-349,2002

[7] MNIST Optical Character Database at AT & T Research, http://yann.lecun.com/exdb/mnist

[8] Joachims, Making large-scale SVM learning practical[C], in Advances in Kernel Methods –Support Vector Learning, Bernhard Scholkopf, Christopher J. C. Burges, and Alexander J. Smola (eds.).pages 169-184.MIT Press, Cambridge, MA, 1999

[9] E.Osuna,R. Freund.and F.Girosi, "Support vector machinese: Training and applications, "A.I.Memo AIM-1602, MIT A.I.Lab,1996

Evaluation Model of Construction of Container Terminal with ANN

Chunling Liu[1] Zhijun Wu[2] Zhiping Zuo[2]

1 School of Electronic & Information, Wuhan University of Science and Engineering, Wuhan, 430073, China

Email: Chunringliu5@yahoo.com.cn AND

2 School of Economics & Management, Wuhan University of Science and Engineering, Wuhan, 430073, China

Abstract

Subjective factors such as weight values, to a minimum degree, effect final results in the course of evaluation of Port construction. It is more likely to lead to decision-making prone to unreliable and inefficient. A new decision-making supporting method was presented to evaluate construction of container terminal according to the characters of container terminal transportation system. Meanwhile a corresponding dual evaluation (horizontal and perpendicular evaluation) model and its application were also provided in this paper in basis of artificial neural network with view to obtaining reliable and accurate final result.

Keywords： container terminal; decision support; artificial neural network; container transportation

1 Introduction

Construction of port takes into accounts not only its natural circumstances and regional economic conditions, but also layout of national transportation system and freight transport economy. Container terminal acted as high technological one is quite different from normal coal and bulk terminals in some aspects. Technologically, a container terminal system request both handling equipments and management techniques to meet demands in handling cargo efficiently and effectively. As a result, it makes sense that its equipment features automation and profession. From economical prospective, it is expensive to construct container terminal and to build container ship and relevant handling equipments. Therefore, to strictly evaluate proposed construction of container terminal is imperative, thus contributing to improving operation efficiency, reducing risk of ships and port congestion and guaranteeing both enterprises gains and society welfare.

There are various approaches employed to evaluate construction of port in many literatures. However, most of them were modeled only based on port enterprises economic indicators, i.e., just considered gains of port enterprises in term of NPV, IRR, PBP (Pay-back period), sensitivity analysis [1][2][3] and ignored a key factor -- the whole society in the process of evaluation. Meanwhile these methods didn’t shake off subjective factors such as weight values, to a minimum degree, which effected final result in the course of evaluation. Consequently, it is more likely to lead to decision-making prone to unreliable and inefficient. In particular, container transportation system involves four subsystems-maritime navigation operation subsystem, handling operation subsystem, warehouse operation subsystem and collection & distribution operation subsystem among which exist inter-relationship. So this paper tries to explore a new approach – a dual evaluation method and establishes the corresponding model as well for evaluating construction of container terminal, in which we apply Artificial Neural Network (ANN) to eliminate subjective influences from decision-makers.

2 Frame structure of the evaluation method

With view to obtaining reliable and accurate final result, evaluating construction of container terminal needs to inspect and to investigate container terminal system from two-dimension facets—production transportation sector and technological economy sector. The former is linked with logistics system, the latter mainly with capital flow system. Whether the two facets of container terminal are compatible and consistent is crucial for stakeholders such as investors, port enterprise, banks and owners. In order to comprehensively appraising investment feasibility of container terminal, analyzing these two facets is converted into horizontal and perpendicular evaluation respectively. In detail, perpendicular evaluation is to appraise technological, operational, economic and environmental factors from container terminal technological economy; while horizontal evaluation is to optimize collection & distribution subsystem, warehouse subsystem, handling subsystem of container terminal in term of production logistics system which ensures the whole system remains production equilibrium and reaches maximum production capacity. Combined with the two evaluations, each factor links and restricts one another, thus forming the frame of dual evaluation method. See Fig. 1

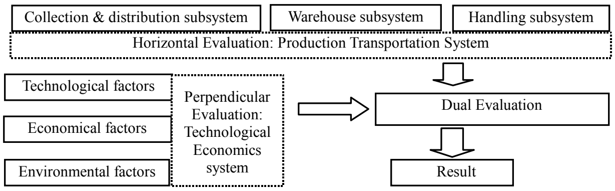

Collection & distribution subsystem
Warehouse subsystem
Handling subsystem
Horizontal Evaluation: Production Transportation System
Technological factors
Economical factors
Environmental factors
Perpendicular Evaluation: Technological Economics system
Dual Evaluation
Result

Figure 1 Framework of the dual evaluation method

3 Perpendicular evaluation

3.1 Perpendicular Evaluation Index System

According to the characteristics of container terminal system, there are three aspects chosen to establish multi-level index system for estimating feasibility of construction of container terminal, they are technological, economic and environmental index, as is shown in Table 1.

3.2 Model of Perpendicular Evaluation

A Back-Propagation (BP) ANN is employed to carry out the perpendicular evaluation, in which factor values used to estimate container terminal construction refer to as input vectors, while the final result of the perpendicular evaluation as output ones, then sufficient samples will be used to train the BP ANN model. Usually, different values of input vectors create different output values, in this way, we obtain relevant coefficients to correct inner expression of the BP ANN model through self-learning or self-training of ANN until the perpendicular model with ANN is well-trained and can be used as an effective quantitative and qualitative tool to evaluate construction of proposed container terminal.

The ANN is a non-linear dynamic system model that contains many simple non-linear computing nodes or joints, and there are three layers, each of which consists of some joints. Within the perpendicular model, there are 1(final result), 8 and 24 joints (as seen in Tab.1) in output, middle and input layers of the perpendicular ANN respectively. Between input layer and output layer is middle layer, which connects input layer forward and output layers backward, the degree of their correlations are donated as relevant coefficients. Through repeatedly adjusting these coefficients, we continuously train the perpendicular ANN model, thus reaching a goal of well-trained model. Each of joints can be expressed as Sigmoid Function, i. e.:

$$f(\mu_j) = [1+exp(-\mu_i)]^{-1}$$
$$= \{1+exp[-(\sum W_{ij}X_j - Q_j)]\}^{-1}$$

where W_{ij} is coefficients of input, middle and output layers, Q_j is threshold.

3.3 Specify and Standardize the Estimated Factor Values

In the input layer, there are following factors which influence construction of container terminal, some of these factor values can be expressed as arithmetic number itself: Area of warehouse u_{111}, Rail & highway transportation capacity u_{112}, Container-opening & patching capacity u_{113}, Productive capacity of transportation u_{114}, Productive capacity of handling equipment u_{134}, Freight amount u_{231}, NPV u_{311}, Dispose of garbage u_{223}, IRR u_{312}, PBP u_{313}, Sensitivity analysis u_{314}, NPV u_{321}, Sensitivity analysis u_{323}, IRR u_{322}. The others are quantified in the following ways.

1) Through fuzzy mathematics the factors u_{232}

(million RMB), u_{221} (decibel)and u_{222} traces out subject function, their subjection functions are

Table1 Major factors of perpendicular evaluation

Standard Level	Factor Level	Sub-factor Level
Technological Set (u_1)	Conditions of collection & distribution (u_{11})	Area of warehouse (u_{111}) Rail & highway transportation capacity (u_{112}) Container-opening & patching capacity (u_{113}) Productive capacity of transportation (u_{114})
	Harbor natural conditions (u_{12})	
	Handling equipment and technology (u_{13})	Selection of handling equipment (u_{131}) Handling technology (u_{132}) Feasibility of handling equipment (u_{133}) Productive capacity of handling equipment (u_{134})
Environmental Set (u_2)	Investment attraction (u_{21})	
	Environmental pollution (u_{22})	Water pollution (u_{221}) Noise pollution (u_{222}) Dispose of garbage (u_{223})
	Conditions of source of cargo (u_{23})	Freight amount (u$_{231}$) Regional economical development (u_{232}) Freight flow stability (u_{233})
Economical Set (u_3)	Financial estimation (u_{31})	NPV (u_{311}) IRR (u_{312}) PBP (u_{313}) Sensitivity analysis (u_{314})
	State economical estimation (u_{32})	NPV (u_{321}) IRR (u_{322}) Sensitivity analysis (u_{323})
	State economical layout (u_{33})	

$$U_1(x_s)=\begin{cases}1 & x_s\in(1600,+\infty)\\ 0.001(x_s-600) & x_s\in(600,1600)\\ 0 & x_s\in(-\infty,600)\end{cases}$$

$$U_2(x_s)=\begin{cases}1 & x_s\in(-\infty,80)\\ 0.025(120-x_s) & x_s\in(80,120)\\ 0 & x_s\in(120,+\infty)\end{cases}$$

$$U_3(x_s)=\begin{cases}1 & x_s\in(-\infty,1)\\ 0.5(3-x_s) & x_s\in(1,3)\\ 0 & x_s\in(3,+\infty)\end{cases}$$

2) Through qualitative analysis specify the factors values including u_{33}, u_{12}, u_{233}, u_{152}, u_{21}, u_{151}, u_{114}, as seen in Table 2.

Table 2 Specified criterion of input factor values

Factors	Good（Ⅰ）	Normal（Ⅱ）	Bad(Ⅲ)
State economical layout	1~0.8	0.7~0.5	0.3~0
Investment attraction	1~0.7	0.5~0.4	0.3~0.1
Freight flow stability	1~0.8	0.5~0.4	0.3~0.1
Harbor natural conditions	1~0.8	0.5~0.4	0.1~0
Selection of handling	1~0.6	0.5~0.4	0.3~0.1
equipment	1~0.7	0.6~0.4	0.3~0.1
Handling technology	1~0.6	0.5~0.4	0.3~0
Feasibility of handling equipment			

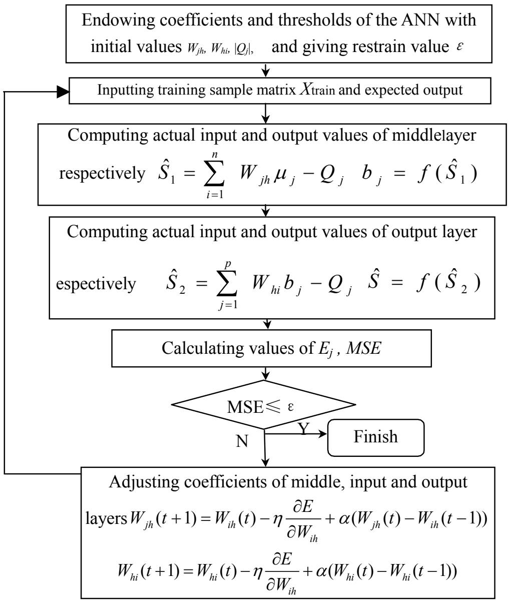

Endowing coefficients and thresholds of the ANN with initial values W_{jh}, W_{hi}, $|Q_j|$, and giving restrain value ε
Inputting training sample matrix X_{train} and expected output
Computing actual input and output values of middlelayer respectively $\hat{S}_1=\sum_{i=1}^{n}W_{jh}\mu_j-Q_j$ $b_j=f(\hat{S}_1)$
Computing actual input and output values of output layer espectively $\hat{S}_2=\sum_{j=1}^{p}W_{hi}b_j-Q_j$ $\hat{S}=f(\hat{S}_2)$
Calculating values of E_j, MSE
MSE≤ ε
Y
N
Finish
Adjusting coefficients of middle, input and output layers $W_{jh}(t+1)=W_{ih}(t)-\eta\frac{\partial E}{\partial W_{ih}}+\alpha(W_{jh}(t)-W_{ih}(t-1))$
$W_{hi}(t+1)=W_{hi}(t)-\eta\frac{\partial E}{\partial W_{ih}}+\alpha(W_{hi}(t)-W_{hi}(t-1))$

Figure 2 The learning procedure of the three-layer BP ANN

of the category of 1) and 2) although the numbers subject to [0,1], so we must standardize factor values next, why should be so is because the estimated factors have different measurement units and grades, which make it difficult to compare with one another in both different factors and schemes. The goal of standardizing factor values is to make all factor values confined to [0,1] and comparable. The approaches vary in two circumstances:

Case 1 Factor values which the more are, the better can be dealt with:

$$F_j = (u_j - u_{j\min})/(u_{j\max} - u_{j\min})$$

Case 2 Factor values which the more are, the worse can be dealt with:

$$F_j = 1-(u_j - u_{j\min})/(u_{j\max} - u_{j\min})$$

where F_j is standardized values of factor X_j, u$_{jmin}$, u_{jmax} is the minimum and maximum values of the j^{th} factor .

(2) Learning Algorithm of the Network[4]

In order to obtain coefficients of each joint of the ANN, MSE of training sample is used to be a target function that trains the ANN. As for given sample matrix X=（u_{ij} ）$_{n\times p}$, it is partitioned into two child matrixes: one is training matrix X_{train}=（u_{ij} ）$_{n\times j}$, the other is testing matrix X_{test}=(u_{ij}) $_{n\times(p-j)}$, and if each column in training matrix X_{train} input the ANN, then the output element $\widehat{S}_j$ and expected output S_j would be created error which defined as: $E_j = (S_j - \widehat{S}_j)^2/2$

The overall sample deviation is

$$MSE = \sum_{k=1}^{m}(S_j - \widehat{S}_j)^2/2M$$

The littler MSE is, the better coefficients of the ANN is. Studies show, in most cases, X_{train} and X belong to the same distribution and have the same MSE, so X can testify the ANN.

The learning procedure of the three-layer BP ANN is shown in Fig.2

If the ANN model was trained well, it could evaluate proposed schemes B = (B_1, B_2, ...B_n), and obtained the best one: Y= max f (B_1, B_2, ...B_n) where Y is the best among the proposed schemes.

where t is number of change, η is learning variance, α is state variance, W_{jh} is coefficients of input and middle layers, W_{hi} is coefficients of middle and output layers, j={1, 2, 3, …, 24}, h={1, 2, 3, …, 8}, i= {1}.

4 Horizontal Evaluation

The purpose of horizontal evaluation for container terminal construction is to optimize the whole production transportation system, in other words, in order to fully take advantage of proposed container terminal and the relevant equipments, it is required to keep every subsystem of container terminal link up to their maximum capacity.

Container transportation system consist maritime navigation subsystem, handling operation subsystem, warehouse subsystem and collection &distribution subsystem. Coordinative and consistent operation of the four subsystems plays a key role in smooth production in the container terminal transportation system. Basically, the number of port throughput depends on the weakest point at the whole system rather than the largest one. Therefore, modern efficient container transportation requires improving overall efficiency of the whole system, which should meet such demand of:

$$P_{Collection\&Distribution} \geqslant P_{Warehouse} \geqslant P_{Handling} \geqslant P_{Navigation}$$

Where $P_{Navigation}$ is overpass capacity of container terminal. In addition, if $P_{Collection\&Distribution}$, $P_{Warehouse}$, $P_{Handling}$ can not satisfy the above inequation, extra capital that invested in these subsystems should be added into the initial overall capital, thus we need to re-analyze financial and state economic conditions(u_{31} and u_{32}).

All in all, horizontal evaluation puts emphasis on operation in production process; on the other hand, perpendicular evaluation focuses on technological and economic feasibility. Only combining both can be created creditable and feasible estimated result.

So, to combine the two evaluation, and consider some special needs (for instance, some investors request that PBP ≤ 10 years) in the process of decision-making, the dual evaluation model can be expressed as:

Target Function Y= max f (B_1, B_2, ...B_n)（Perpendicular evaluation）

S.T:

$$\begin{cases} u_{1jk} \,?\, a_1 \textit{ and、or } u_{1jk} \,?\, a_2 \\ u_{2jk} \,?\, a_3 \textit{ and、or } u_{2jk} \,?\, a_4 \\ \quad \vdots \qquad\qquad\quad \vdots \\ u_{nmp} \,?\, a_{2n-1} \textit{ and、or } u_{nmp} \,?\, a_{2n} \end{cases}$$

$$P_{Collection\&Distribution} \,?\, P_{Warehouse} \,?\, P_{Handling} \,?\, P_{Navigation} \qquad \textit{(Horizontal evaluation)}$$

Where u_{nmp} is the p^{th} factor value of the m^{th} factor layer of the nth standard layer；

a_{2n-1}，a_{2n} is standard value.

5 Application

We used various economic indictors as tested samples after selected some typical cases of container terminal construction nationwide such as Dalian, Qingdao, Huangyi, Tianjing, Shanghai and Guangdong from the year 1995 to the year 1997.$\eta$and$\alpha$are constant, η∈(0.15～0.015)，α=0.075, note that if η intend to increase，it means weight values change fiercely，thus leading to learning process fluctuate. But η can not select too small value otherwise this would slows down speed of learning，in order to make η increase gradually in the process of learning, we used conditional and recycling causes to fulfill this objective.

Through trained network with different samples until value ofεmeets requirement, we tested and evaluated nine themes[5], comparison between the result from dual model and the one from fuzzy model is presented in Table 3.

Table 3 A schematic representation of results from two models

Number of theme	Sequence of Dual Model	Sequence of Fuzzy Model
1	7（O）	8（O）
2	7	3
3	5	5
4	1	1
5	8（O）	7（一）
6	2	2
7	4	4
8	6	6
9	9	9

From the results in Table 3, among nine themes, the sequence of seven ones remain original order, in particular, the order of the first six themes in term of dual model is consistent with those in term of Fuzzy model. Although the results of theme 1 and 5 have differences in two models, the sequence are close nearly, the reason for it is that the two themes both need reconstruct based on the original ports, theme 1 need not invest in distribution system, while theme 5 need it to balance every capacity of all segments in system, due to the economic result of theme 1 and theme 5 are not good, so additional investment worse economic result of theme 5, in this way, theme 1 should better than theme 5.

From the analysis above, it is easily concluded that the dual model is more reliable and accurate. Currently the model has developed relevant software used in Communication Ministry, China for evaluation.

References

[1] D.M. Tolmson. Port and Harbor Evaluation. Cambridge University Press, 1998, pp.455-480

[2] Report of Economic Evaluation for Gaogang Ten-thousand-ton Port. Waterborne Transportation Research. Institute of Communication Ministry, Beijing, China, 1990, pp. 56-60

[3] P. H. Wang. Study of DSS for Port Evaluation. The Thesis of Wuhan Waterborne Transportation College, Wuhan, 1995, pp.34-38

[4] A. C. Goh. Back-Propagation neural network for modeling complex systems. Artificial Intelligence in Engineering, 1995, 9, pp.143-151

[5] J. Z. Li. Study of IDSS for Evaluation of Container Terminal Construction. The Thesis of Wuhan Transportation University of Science, Wuhan, 1997

[6] C. Q. Zhang. Storage space allocation in container terminals. Transportation Research: Part B, 2003, 37(10), pp.883-903

[7] K.H. Kim, P. K. Tae. A note on a dynamic space allocation method for outbound containers. European Journal of Operational Research, 2003, 148(1), pp.922-101

[8] P. Preston, E. Kozan. An approach to determine storage locations of containers at sea port terminals. Computers & Operations Research, 2001, 28(10), pp.983-995

[9] E. Nishimura.Yard trailer routing at a maritime container terminal. Transportation Research: Part E, 2005, 41(1), pp.53-76

[10] T. Dilek, I. B. Laura. A two-phase tabu search approach to the location routing problem. European Journal of Operational Research, 1999, 116(1), pp.87-99

An Improved Grouping Model In DNA Computing

Wei Liu Shouxia Sun Chunling Liu Jing Ju

Department of mathematics and information, Ludong University, Yantai, Shandong, China
Email: liuyiwei1030@yahoo.cn

Abstract

DNA computing is a means of solving a class of intractable computational problems, in which the computing time can grow exponentially with problem size. Based on the Adleman-Lipton model, some grouping strategy and the operations are designed to study the commonly algorithm model of 3-SAT in DNA computing. The parallel algorithm is proposed to control the running time of algorithm to constant.

Keywords：Grouping Model; Probe Modules; DNA Computing

1 Introduction

DNA computing is to solve computational problems by employing molecular biology laboratory techniques to manipulate DNA strings. It has been developed rapidly these days since the first algorithm was proposed by Adleman[1]. Since then, many groups have worked on NP-complete problems with the Adleman-Lipton model computing [3-13].For example, Ouyang's work on solving the maximal clique problem[3] and Faullhammer's RNA solution to chess problems[5] are further demonstrations of molecular-based computations. Next DNA computing on surfaces [7] and DNA computing by DNA hairpin formation [4] were proposed to solve 3-SAT problem.

In this paper, a DNA-based grouping algorithm model[7] is proposed to solve a general CNF Boolean formula F with n-variables, m-clauses instance of the 3-SAT problem. An unique grouping Strategy [13] is implemented to divide m-clauses of the given Boolean formula into [m/4] or [m/4] +1 parts. By making using of biological operations, the proposed algorithm can effectively control running-time to constant.

The paper is organized as follows. Section 2 describes the background of SAT problem and probe modules. Section 3 offers the novel grouping strategy to implement self-assembly. Finally, a conclusion is drawn in section 4.

2 Background

2.1 SAT Problem

The SAT (Propositional Satisfiability) problem, which is the first one ever shown to be NP-complete, is to find assignments of a set of Boolean variables that lead the output of a Boolean formula to be true. It is a simple search problem known as one of the hardest NP problems. Every NP problem can be seen as the search for a solution that simultaneously satisfies a number of logical clauses. In a subclass of the SAT problem, called conjunctive normal form CNF-SAT [1], Boolean formulas are restricted to the following form:

$$F = C_1 \wedge C_2 \wedge \ldots \wedge C_m \qquad (1)$$

Where each C_t is a "clause", and '∧', '∨', '¬' are logical AND, OR, NOT operations, respectively. If each clause C_t contains not more than three variables, such as $C_t = x_i \vee x_j \vee x_k$, that is 3-SAT, which is decided whether it is a given 3-CNF formula is satisfiable. Boolean variables x_i is allowed to range over "true" and "false" values. A literal is either a variable x_i or its negation $\neg x_i$.

2.2 Probe Modules

Probe module is an important tool in the DNA computing, we can use it to capture target sequence by hybridize. Generally, we need to create all the probe modules which will be used during the process before starting an algorithm. The design of probe module must follow some protocol [1][2], i.e.

For each clause C_t of the formula F, a probe module is created by adding the corresponding 5′-end Acrydite-modified oligonucleotides probe to polyacrylamide gel. If the t^{th} clause C_t contains literal x_i, the corresponding Acrydite-modified probe $\overline{x}_i^T$ is added, where $\overline{x}_i^T$ denotes the Watson-Crick complement of x_i. If the t^{th} clause C_t contains literal $\neg x_i$, the corresponding Acrydite-modified probe $\overline{x}_i^F$ is added, where $\overline{x}_i^F$ denotes the Watson-Crick complement of $\neg x_i$.

3 Grouping algorithm model

3.1 Synthesize Library Strand Sequence

Here proposes a novel design of grouping strategy, let each library strand consists of four different forms, and the different form of library strands will be put into four different tubes to implement the self-assembly algorithm. All of the library strand sequences must be synthesized before the algorithm starts.

For every library strand sequence L_j in S, we design four kinds of function segments show as follows:

$$S^{(1)}:5'-\left(L_j,h_j\right)-3'$$

$$S^{(2)}:3'-\left(\overline{h}_j,L_j,h_j\right)-5'$$

$$S^{(3)}:5'-\left(\overline{h}_j,L_j,h_j\right)-3'$$

$$S^{(4)}:3'-\left(\overline{h}_j,L_j\right)-5'$$

Where

(1) L_j encode all possible truth assignments of library strand in S;

(2) 5′-(h_j)-3′denotes 5′-3′ sticker end of every library sequence L_j, $3'-\left(\overline{h}_j\right)-5'$ denotes its Watson-Crick complement;

(3) While 3′-(h_j)- 5′ denotes 3′-5′ sticker end of every library sequence L_j, $5'-\left(\overline{h}_j\right)-3'$ denotes its Watson-Crick complement;

(4) $S=\{S^{(1)},S^{(2)},S^{(3)},S^{(4)}\}$;

(5) h_j encode a sticker end for every strand L_j, *so that* L_j only hybridize with one of it's three forms.

3.2 Algorithm

The grouping algorithm describe as follows:

(1) Use quaternion to divide the m clauses into [m/4] or [m/4] +1 groups;

(2) Search the answer of each group which can satisfy four clauses;

(3) Continue to use quaternion to divide the left tubes into [m/16] or [m/16] +1groups;

(4) Search the answer of each new group which can satisfy sixteen clauses;

(5) Such-and-such circulate, till get the answer of formula F.

3.3 Operations

For formula F of the formula (1), the operation of the algorithm shows as follows:

Step1: First distribute m tubes into [m/4] or [m/4] +1groups. That is quaternion. Second, for each group, put sub-set $S^{(1)},S^{(2)},S^{(3)},S^{(4)}$ of S into four different tubes, respectively. Third, put *m* different probes into the *m* different tubes T_1, T_2, …, T_m, and refrigeration to $4°C$ simultaneity. Then the strands in all of test tubes occur hybridize reaction. The strands of S encoding truth assignments satisfying the corresponding clause are captured in the corresponding capture layers, while those strands encoding non-satisfying assignments corresponding clause in every tube then be removed out of the test tubes with the buffer.

Step2: First wash the tubes and then add new buffer. Next rise the temperature to $85°C$: The probe modules in the tubes will melt of the captured strands into buffer then remove the probe modules.

Step3: Integrate every group tube into one of the four tubes, then add T4 DNA ligase, and the strands in them occur link reaction under the function of T4 DNA ligase, then they will come into being duplex.

Step4: Keep those DNA strands which length is 4l+3s (l is the length of L_i, while s is length of sticker end) by polyacrylamide gel electrophoresis. That is, to apart the longest length strands.

$$5'-(L_i, D(h), L_i, D(h), L_i, D(h), L_i)-3'$$

Where D(h) denotes duplex of h sticker end.

Step5: Rise the temperature to $95°C$ for every group tube, so that the duplex strands in those tubes are denatured. As a result, these strands turn into single strand form. Then regroup the remain [m/4] or [m/4] +1 test tubes after the first recycle to [m/16] or [m/16] +1 groups;

Step6: Put four probes into corresponding tubes of each group respectively, the four different tubes corresponding four different probes as follows:

The four probes of each group will capture corresponding four different kinds forms of S respectively, while those strands not be captured will be removed out of each tube with the buffer.

Step7: Repeat step2 to step6 until there is only one test tube left. If there are only two or three tubes in one group in the course of one recycle, then random take any of two probes or three probes in step6. If there is just one tube in one group in the course of recycle, then do not dispose it, and put it into the next recycle directly;

Step8: Apart the longest DNA strands by polyacrylamide gel electrophoresis from the last test tube. That is the answer strands of the formula F.

Step9: First, rise the temperature of the last tube, then the strands captured by the probe modules denatured to form single-strands. Finally, remove the probe modules.

Step10: Extract all the answer strands from the last tube, and then use PCR-amplify technique [1] enlargement these strands so that we can “read” [2] the answer strands easily.

4 Conclusions

A novel Grouping algorithm model is proposed by making use of the quartation and grouping strategy. The improved grouping algorithm model designs four forms for one strand, so that we can implement self-assembly in the algorithm and make use of tube numbers reducing at the speed of $\log_4 m$. As a result, the algorithm can get constant time-complexity. The total use of tube numbers are m+4. We can say that it is an effective improvement algorithm model for SAT problem in DNA computing. We must to say that it is a hard work to meet the need of DNA computing.

References

[1] L.Adleman et al., “Molecular Computation of Solutions to Combinatarial Problems”, *Science*, vol.266, 1994, pp.1021-1024

[2] R.Lipton et al., “DNA Solution of Hard Computational Problems,” *Science*, 1995, pp. 542-545

[3] Ouyang, Q., Kaplan, P.D., Liu, S., Libchaber, “DNA solution of the maximal clique problem,” *Science*, 1997, pp.446–449

[4] Sakamoto, Yukoyama, Takashi, Hagiya, Molecular, et al., “Computation by DNA Hairpin Formation,” *Science,* 2000, pp.1223-1226

[5] Faullhammer, D., Cukras, A.R., Lipton, R.J., Landweber, L.F. “Molecular computation: RNA solutions to chess problems,” *Proc.Natl. Acad. Sci.* U.S.A., 2000, pp. 1385–1389

[6] L.Adleman et al., “Solution of a 20-Varuable 3-SAT Problem on a DNA Computer,” *Science*, 2002, pp.1026

[7] Liu,Q…, L.Wang, Frutos, A.G., et al. “DNA Computing on Surfaces, ” *Nature*, 2003, pp. 175-179

[8] Dafa Li, Hongtao Huang, Xinxin Li, Xiangrong Li. “Hairpin formation in DNA computation presents limits for large NP-complete problems,” *BioSystems,* 2003, pp.203–207

[9] Weng-Long Chang, Minyi Guo, “Solving the set cover problem and the problem of exact cover by 3-sets in the Adleman–Lipton model,” *BioSystems,* 2003, pp. 263–275

[10] Minyi Guo, Michael,Weng-Long Chang, “Fast parallel molecular solution to the dominating-set problem on massively parallel bio-computing,” *Parallel Computing*, 2004, pp. 1109–1125

[11] Weng-Long Chang, Minyi Guo, Michael Ho, “Towards solution of the set-splitting problem on gel-based DNA computing,” *Future Generation Computer Systems*, 2004, pp. 875–885

[12] Chia-Ning Yang, Chang-Biau Yang, “A DNA solution of SAT problem by a modified sticker model,” *BioSystems*, 2005, pp. 1–9

[13] Wei Liu,YingGuo, et.al. “Grouping parallel algorithm model of 3-SAT in DNA computing,” *Impulsive Dynamic Systems and Appliations*, 2006, pp.1533-1535

Neural Networks Based on Fuzzy Clustering and Its Application in Electrical Equipment's Fault Diagnosis

Long Zhou[1] Xuezhi Wang[1] Mianyun Chen[2]

1 Wuhan Polytechnic University，Wuhan, 430023, P. R. China
Email：zhoulong@whpu.edu.cn

2 Huazhong University of Science and Technology, Wuhan, 430074, P. R. China

Abstract

Combining the fuzzy sets theory and neural network to carry on the fault diagnosis is a most prosperous diagnosis method. This article put forward a sample processing method using fuzzy clustering and studied the application of fuzzy competition classification method in extracting contradiction samples, then advanced the diagnosis method of neural network based on fuzzy clustering, finally carried on the simulation research. The calculation results show that all the above-mentioned methods are quite practical. It is a better and prosperous way to combine neural network theory and fuzzy clustering to carry on insulation diagnosis.

Keywords：Neural network; fuzzy clustering; electrical equipment; fault diagnosis

1 Introduction

The distinguished advantage of fuzzy diagnosis lies in that it needn't a creation of precise mathematic model as long as subordinate function, fuzzy relation and fuzzy rules are applied properly, then carry on fuzzy deduction, and we can realize the intellectualization of fuzzy diagnosis. But as for the complex diagnosis system, creating the right fuzzy rules and subordinate functions is very difficult, and needs long time. Because of the complexity of the electronic equipment faults, the mapping relation from time or frequency domain fundamental space to fault pattern space is strongly nonlinear. At the same time the shape of the subordinate function is extremely irregular. Only the shape of normal subordinate function can be used to process approximately, but to do so limits the precision and possibility of processing the input system with wider range, and makes the results of the nonlinear system diagnosis dissatisfactory [1-3].

In order to solve these problems existing in fuzzy diagnosis, it is a better way to introduce neural network. As we all know, in the fault diagnosis of equipment exists a lot of disdeterminacy. Studying the disdeterminacy is the key to determine whether the diagnosis is right or not. Exiting fuzzy diagnosis methods usually use fuzzy relation matrice to transform the disdeterminacy, however, introducing neural network to fuzzy diagnosis is using network structure to transform the disdeterminacy. This is the nature of the fuzzy diagnosis methods based on neural network. This method has become the most prosperous one in electronic equipment fault diagnosis. Although many learners have carried on the research in this field, it is still in the probing phase regarding creating learning sample, input fashion and the convergency of the learning process. So, this article stressed the fuzzy processing methods of learning sample, and put forward the fuzzy diagnosis methods based on neural network, and carried on the simulation research. [4-5].

2 The method of neural network based on fuzzy clustering

Combining the fuzzy sets theory and neural network to carry on the fault diagnosis is a most prosperous diagnosis method. When the BP network is applied to the fuzzy diagnosis, the inputs of the network are the subordinate degree value after the transition of the diagnosis parameters. By the fuzzy mapping relation that has been trained by the network, we can get the output subordinate degree values of the fault reason that show the possibility of certain fault’s occurency. Article[2] carried on the research of the application of fuzzy sets theory, expert system and neural network in the analysis and diagnosis of the dissolved gas in transformer oil, then advanced an insulation diagnosis method combining fuzzy sets theory and neural network.

From the analysis above, we can see that when the BP network is applied to diagnosis, it has extremely high demands for the range, density, consistency and uniformity which is distributed in the space by the sample data, otherwise, it will directly affect the adptivity and precision of diagnosis model. This section processes the sample using methods based on fuzzy sets theory, it is also a respect of combining fuzzy sets theory and neural network to carry on diagnosis. [6-10]

2.1 Competition classification of the sample

In order to reduce the network complexity and degree of difficulty of study, we must carry on the competition classification to the sample. Suppose the sample space of historical data is *{X,Y}*,the data of sample is *<X$_1$,Y$_1$>*, *<X$_2$,Y$_2$>……<X$_n$,Y$_n$>*,then the cluster's criterion function is:

$$J_e = \sum_{i=1}^{k}\sum_{u\in\Gamma}\left\|u - C_i\right\|^2 \qquad (1)$$

In this posture, *‘k’* is the number of subset of the sample; C_i is the *‘i’* one of center sample subset; *‘u’* is the *‘i’* one of it’s element of the sample space. The purpose of the cluster is to minimize this function and classify the following algorithm.

(1) Standardizing of data. In order to dispel the influence of the difference of every indexes of sample different from order of magnitude, we should deal with the normalization to every index value, obtained value is:

$$X'_{ij} = (X_{ij} - \overline{X}_j)/\delta_j,$$

$$\overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}, \delta_j = [\frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \overline{X_j})^2]^{1/2} \qquad (2)$$

(2) Put into a seed of N in the sample space: *C= {C$_1$, C$_2$ ……C$_k$}*, as the centre of the initial subset. Then we calculate the distance of each sample reaches to concentrated heart*(D$_{ik}$ = / X$_i$ − C$_k$ /),* the seed of minimum distance is to the victor of competition of the sample, it means the sample X_i is members of the *‘k’* sample subset, receiving the number of sample subset M_k.

(3) Calculation of cluster's criterion function. If $J_e(t) \leqslant \varepsilon$, cluster's quality is smaller than giving the value definitely in advance , classify is over , otherwise transfer to the next step.

$$C_i(t+1) = C_i(t) + (C_i(t) - u)/(M_i - 1) \qquad (3)$$

(4) Choose a sample *‘$u \in \Gamma_i(M_i \neq 1)$’*,move the *‘u’* to the *‘ $\Gamma_j(i \neq j)$’*, calculate the new centre and criterion function.

$$C_j(t+1) = C_j(t) + (u - C_j(t))/(M_j - 1) \qquad (4)$$

$$\begin{aligned} J_{ei}(t+1) &= J_{ei}(t) - \Delta J_{ei}(T) = \\ & J_{ei}(t) - M_i\left\|u - C_i(t)\right\|^2/(M_i - 1) \end{aligned} \qquad (5)$$

$$\begin{aligned} J_{ej}(t+1) &= J_{ej}(t) - \Delta J_{ej}(T) = \\ & J_{ej}(t) - M_j\left\|u - C_j(t)\right\|^2/(M_j - 1) \end{aligned} \qquad (6)$$

If$\triangle J_{ej}(t) < \triangle J_{ei}(t)$, we can get: $J_{ej}(t+1) < J_{ei}(t+1)$ because collecting and turning a sample to another stature collection from a stature will not influence other kinds.

To each *($i \neq j$)*, make the minimum: $\min\triangle J_{ej}(t) = \triangle J_{em}(t)$, then move *‘u’* into Γ_k, get $J_e(t+1)$.

(5) If the number of the competition has already reached, its over, otherwise move to step(4).

Competed and revised many times like this, finally the seed steadied in the center of each of subset through competition, each sample is only one stature collection, thus divide the sample space into *K* subsets.

2.2 Studying the neural network studying

To different subsets, we should adopt the forward propagation algorithm to pool design and make the whole network(including its structure and algorithm), and separately adopt the corresponding sample data to train the network, got a set of network made up by K networks finally. Putting the measure samples into the network will get the results of K neural networks separately.

2.3 Companion of the result

For measure sample, we can't decide whether it belongs to an accurate subset; we can just confirm that to how great a degree it belongs to a certain classification, then average the results of each sample and obtain the final result finally.

$$\mu_i(X) = S_i(X)/\sum S_i(X),\quad S_i(X) = M_i/\left|X - C_i\right|^2 \quad (7)$$

In this posture, $S_i(X)$ is the distance of from measured sample to each sample's subset, $\mu_i(X)$ is the result of the jurisdiction of degree finally.

3 Tion research

In order to verify the fuzzy diagnosis method based on neural network put forward by this article. Still take the data of dissolved gas in transformer oil as the analysis and simulation object of insulation diagnosis. The following is the concrete steps:

(1) Create the initial sample

Allowing for the influence of such factors as the type, volume and operating environment of the transformer, according to the document at hand, we collected thirty transformer chromatogram test records and corresponding fault results which are produced by different manufacture factories and operating under different voltage levels and in different zones, in order to facilitate comparison, take 13 sets in article [2] as experimental analysis sample, in addition, draw 7 sets from 30 sets randomly, they together form the initial data sample ,then adopt the fuzzy processing method advanced in the former section to analyze the initial sample .Here still adopt the method in article [2] to quantitatively process the test data ,then get the initial sample data .

At present , when intelligently diagnosing and analyzing the dissolved gas in transformer oil , mostly adopt the gas content and total hydrocarbon value of CO、H_2、CH_4、 C_2H_2、 C_2H_4、 C_2H_6，as the input parameter , to diagnose and analyze . According the synthetic consideration and comparison , take the total hydrocarbon value Total and the ratios in total hydrocarbon value of C_2H_2、H_2 CH_4 and C_2H_2 as the parameter. Fault outputs are divided into general super-heating、serious super-heating、partial discharge、spark discharge and electric arc discharge .

(2) Create standard sample

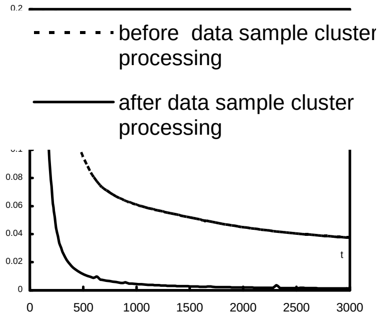


Figure1 he convergence comparison before and after data sample cluster processing

After creating the normal samples. According to the results of fuzzy cluster, choose the normal sample, the standard samples are twelve sorts, and then the standard sample can carry on the learning and training of neural network.

(3) Training and diagnosis of neural network

Figure 1 shows the convergence comparison before and after data sample cluster processing, the network ratio is η=0.75. This figure indicates that the learning convergence velocity of the standard sample after cluster processing increases largely, practical diagnosis results show that using standard sample after cluster processing to diagnose can meet the demand of diagnosis, and accurately diagnose.

4 Conclusions

This article studied the fuzzy method of samples optimization processing and put forward the diagnosis method of neural network based on fuzzy clustering. Some useful conclusions can be drawn as follows:

(1) It is a better and prosperous way to combine neural network theory and fuzzy clustering to carry on insulation diagnosis.

(2) When using neural network to carry on insulation diagnosis, it is important that processing method based on fuzzy clustering can well analyse the problems of single sample, sparse sample and contradiction sample existing in samples, removing contradiction sample can distinctly improve the accuracy of neural network diagnosis.

Acknowledgements

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] Zhao zhenyun, Xu yong mou,The basic and applications of fuzzy theory and neural networks，tsinghua university publish，1996

[2] Wang dazhong et al.，fuzzy theory ,expert systems, neural networks and their applications in fault diagnosis of the fourier transform，Proceedings of the Chinese society for electrical engineering ，Vol.16, No.5, pp.349-353, 1996

[3] S.K.Bhattacharyya, et al.,A neural network approach to transformer fault diagnosis using dissolved gas analysis data, NAPS'93, 1993

[4] Wang caisheng，The fault diagnosis method of BPNN in detecting for fourier transform，Proceedings of the Chinese society for electrical engineering，Vol.17, No.5, pp.322-325, 1997

[5] O.Baldi, Neural networks and principal component analysis:learning from examples and local minima, Neural Networks, 2:53-58, 2007

[6] Chen Ning, Chen An, Zhou Longxiang. Fuzzy K-prototypes algorithm for clustering mixednumeric and categorical valued data, Journal of Software,Vol.12, No.8, pp.1107-1119,2001

[7] zhoulong, et al., Recursive neural networks and its application in forecasting the state of metal oxide arrester, pp790-792,DCABES 2004 proceedings

[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

[10] Huazhong electrical power testing institute, The analysis on spoil reason of 500KV metal oxide arrester in Gezhouba.1994(11)

An Improved Genetic Algorithm for the Job-Shop Scheduling Problem

Hui Hong[1] Tianying Li[1] Hongtao Wang[2]

1 Modern Education and Technology Center, Shangqiu Medical College, Shangqiu, Henan, China

Email: sqyzhh@126.com

2 Wuhan University of Technology, Wuhan, Hubei, China

Email: waterfly2006@yahoo.com.cn

Abstract

As a class of typical production scheduling problems, job-shop scheduling problem is one of the strongly NP-complete combinatorial optimization problems, for which an improved genetic algorithm with search area adaptation is proposed in this paper. It has a capacity for adapting to the structure of solution space and controlling the tradeoffs balance between global and local searches. And our proposed method has an outstanding point which is that it does not need a crossover operator with an ability of characteristic inheritance ratio control.

Keywords: job-shop scheduling problem; an improved genetic algorithm; the tradeoffs balance between global and local searches

1 Introduction

As one aspect of operations research, classic job-shop scheduling problem has a very wide and well-developed engineering background, and it can be described as follows. A set of m machines and a set of n jobs are given; each job consists of a set of operations that have to be processed in a specified sequence; each operation has to be processed on a definite machine and has a processing time which is deterministically known. A schedule defines the time intervals in which the operations are processed, but it is feasible only if it complies with the following constraints: each machine can process only one operation at a time and the operation sequence is respected for every job. The objective is to find the optimal schedule, the operation order and starting time on each machine such that the makespan is minimal.

Stochastic optimization methods, such as the Genetic Algorithm (GA), have been widely applied to this problem. GA has shown a good performance regarding its ability to search globally. It starts searching with population and multiple points in the search space. It has an operator called crossover, which enables to search over wide region. However, for the proper tradeoffs balance between global and local search abilities, adjusting parameters, such as crossover rate and mutation rate, is necessary for GA.

Considering the roles of adjusting parameters, a genetic algorithm with search area adaptation (GSA) was proposed. It is developed from GA. GSA has capacity for adapting to the structure of solution space and controlling the tradeoffs balance between global and local search abilities dynamically. It had been confirmed that GSA shows good performance. But, GSA needs a crossover operator with an ability of characteristic inheritance ratio control.

In this paper, we propose the improved GSA (iGSA) for solving the job-shop scheduling problem that does not need such crossover operator. It is shown that this method has better performance than existing GAs.

2 GSA

GSA searches worthy regions in a solution space adaptively for good solutions. One cycle of generation is divided into the following two phases, crossover phase and mutation phase. The crossover search phase consists of three parts, selection for reproduction, crossover, and selection for survival. The mutation search phase also consists of three parts, selection for reproduction, mutation, and selection for survival.

2.1 Crossover search phase

The role of this phase is to find a region that has a high probability of containing good solutions between two parents. The procedure of searching a region depends on the topology of the search space. In searching this space, the region might be on the line segment connecting two parents. This phase has a characteristic of 'finding a small region that is worth of intensive search between two parents'. For global search, a middle region between two parents is selected if possible. If a child that is better than the worst in the population has not appeared until a certain stop condition is satisfied, the best child in this phase is selected as the eldest daughter. GSA already has excellent abilities of preservation and inheritance ratio control. This GSA procedure controls searching region dynamically. A child that is worse than its parents is allowed to survive. Thus, the parents can get out of local minima.

2.2 Mutation search phase

This phase consists of three parts, selection for reproduction, mutation, and selection for survival. An individual chosen by mutation is called Mutant. The role of this phase is to search the region, which is selected in the crossover search phase, for good solutions. This phase is given the role of 'generating good individuals by searching around a single individual intensively', under the constraint of not searching toward a worse solution.

3 Improved GSA for job-shop scheduling problem

3.1 Designing chromosomes

A chromosome has gene information for solving the problem in GA. The coding method currently most used to solve the scheduling problem is the permutation encoding. Thus, we apply this permutation encoding that is easy to use. This coding method can create an active schedule at every time. The chromosome shows the order of job-number. If the number of jobs is n and the number of machines is m, the chromosome consists of n×m genes. Each job will appear m times exactly; it shall be depending on an order relation. Thus, each chromosome represents a feasible solution. Each gene is given priority, and a left gene has the priority higher than a right gene. In other words, each gene is ordinarily scheduled from left side. The scheduled gene is positioned at the best feasible solution. For example:

Table1 Example of JSP

job	(machine number, processing time)		
J1	(1,3)	(2,3)	(3,2)
J2	(1,1)	(3,5)	(2,3)
J3	(2,3)	(1,2)	(3,3)

Table2 Chromosome

3	2	2	1	1	2	3	1	3

An example of a three-job three-machine (3×3) JSP is presented in Table 1.And its chromosome based on the rule that is mentioned above is showed in Table 2

3.2 Designing crossover and mutation operations

We use the following symbols in our crossover operation.

PA and PB: parent chromosomes.

CA and CB: child chromosomes.

cp1 and cp2: crossover points.

Ap and Bp: partial chromosomes between cp1 and cp2 in each parent.

Ac and Bc: partial chromosomes between cp1 and cp2 in each child.

The crossover operation selects genes from parent chromosomes and creates a new offspring. The crossover operation for the job-shop scheduling problem is modified as follows:

Step 1: Select PA and PB as parent chromosomes. Create empty string CA and CB to host the resulting chromosomes (children chromosome) of the crossover operation.

Step 2: Two crossover points (cp1 and cp2) are selected randomly. Let the partial chromosomes between cp1 and cp2 in each parent be Ap and Bp, respectively.

Step 3: All the genes except Ap and Bp are moved to CA and CB.

Step 4: The following processes are performed.

Step 4a: If the same gene exists the gene of Ap as compared with Bp, mutual position-information will be exchanged and it will move to AC and BC, respectively.

Step 4b: Order information is saved and the remaining genes are moved.

An example of this crossover operation is illustrated below.

Chromosomes

Ap	1	6	5	7	1	0
Bp	6	4	8	7	0	1

Step 4a move same gene

Ap			5	1
Bp		4	8	
Ac	6	6		
Bc	1			

Step 4b move remaining gene

Ac	6	5	1	7	0	1
Bc	1	6	4	7	8	0

In the mutation operation, we use the swap mutation operator that simply selects two mutation points randomly and swap each gene.

3.3 Crossover search phase

Our crossover search phase is modified on the basis of crossover operation as mentioned above and the crossover search phase on GSA.

Step 1: Chose three individuals from the population randomly, and make three pairs.

Step 2: For each pair, do the following step 3–step 6.

Step 3: Two crossover points are selected randomly, and create two children using the crossover operation.

Step 4: Generate a certain number of children that inherit from each parent using the crossover operation.

Step 5: If the best child produced through the crossover operation is worse than the worst individual in the population or the iteration number j is less than Rc (Rc is parameter), go back to Step 3.

Step 6: Select the best child as the eldest daughter.

3.4 Mutation search phase

Our mutation search phase is modified on the basis of mutation operation as mentioned above and the mutation search phase on GSA.

Step 1: For each eldest daughter, do the following step2–step6.

Step 2: Selected children are set to mutants B and P.

Step 3: Generate mutant C from mutant P using the mutation operation (reproduction-M).

Step 4: If mutant C is better than mutant B, mutant C is set to mutant B.

Step 5: If the iteration number j is less than Rm (Rm is parameter), go back to Step 3.

Step 6: If the best mutant is better than the worst individual in the population, it is selected as an offspring. Otherwise, at first, make a set including all individuals in the population and the best of the overall individuals appeared. Then, the chromosome of the individual chosen from the set randomly is copied to an offspring.

Step 7: Replace all parents with all offspring.

4 Experiments and results

In order to test the iGSA approach for job-shop scheduling problem in this study, we apply iGSA and GA to ft10×10 and ft20×5 respectively. For this numerical experiment, we follow the same environment given in. In our experiment, to be fair, each method is allowed to generate the new solutions of about 3,000,000 in each trial. The population size (Pop) is 400, the crossover ratio and the mutation ratio on GA are 0.8 and 0.2, and the parameters Rc and Rm on iGSA are 2 and 25 respectively.

The experimental results are shown in TABLE. Here, the terms 'Best', 'Worst' and 'Ave' mean the best, the worst and the average solutions in the numerical experiment respectively.

Table 3 Experimental Results

		Best	Worst	Ave	time
ft10×10	GA	932	1058	992	55
	iGSA	892	972	937	20
ft20×5	GA	1230	1355	1286	69
	iGSA	1263	1302	1271	25

In ft10×10, the makespan of optimal solution is 930, and in ft20×5, it is 1165. The programming language is Visual C++, and it runs on Windows XP with Pentium IV (1.6 GHz).

5 Conclusion

In this paper, a novel scheduling algorithm named iGSA has been applied to solve the job-shop scheduling problem. The iGSA approach combines the GSA immune theory, which can improve the global exploration performance of GA.

We observed the improved performance of iGSA in comparison with GA. In the future, we shall choose the instances from which the structure of the solution space differs clearly, and shall conduct more numerical experiments.

Acknowledgments

This research is completed under the aid of professor Xue. Many thanks to Dai Feng of the Department of Computer Science, Huazhong University of Science and Technology for his helpful comments. We would also like to thank our anonymous reviewers for their constructive comments.

References

[1] M. Vroblefski, E.C. Brown, A grouping genetic algorithm for registration area planning, Omega-International Journal of Management Science 34 (2006) 220–230

[2] A. S. Jain and S. Meeran, "Deterministic job-shop scheduling". European Journal of Operational Research, 113, pp. 390–434, 2005

[3] S. G. Ponnambalam, P. Aravindan and S. V. Rajesh, "A search algorithm for job shop scheduling", International Journal of Advanced Manufacturing Technology, 16, pp. 765–771, 2002

[4] Someya, H., & Yamamura, M. (2004). Genetic algorithm with search area adaptation for the function optimization and its experimental analysis. Proceedings of IEEE, CEC2001, 933–940

[5] S. Kobayashi, I. Ono, and M. Yamamura, "An efficient genetic algorithm for job shop scheduling problems," in Proceedings of the Sixth International Conference on Genetic Algorithms, 2006, pp. 506–511

[6] R.Bruns, "Direct chromosome representation and advanced genetic operators for production scheduling," in Proceedings of the Fifth International Conference on Genetic Algorithms, 2005, pp. 352–359

[7] T.Yamada and R. Nakano, "Job-shop scheduling," Genetic Algorithms in Engineering Systems, Chap. 7, IEE control Engineering Series, vol. 55, 1997, pp. 134–160

[8] Norman, B.A., Bean, J.C., 2001. A genetic algorithm methodology for complex scheduling problems. Naval Research Logistics 46, 199–211

[9] Lin, S.-C., Goodman, E.D., Punch, W.F., 1997. A genetic algorithm approach to dynamic job shop scheduling problems. Proceedings of the 7th International Conference on Genetic Algorithm. Morgan Kaufmann, San Francisco, CA

[10] Zhang H-P, Gen M. Multistage-based genetic algorithm for flexible job-shop scheduling problem. Journal of Complexity International 2005;11:223–232

Research on Evaluation Model and Applications of DSM Project Benefit

Yanhui Wang[1]　Xuebin Zhang[2]

1 School of Business and Administration, North China Electric Power University, Baoding,
Hebei , 071003, China
E-mail: wyanhui@126.com

2 School of Business and Administration, North China Electric Power University,
Baoding, Hebei, 071003, China

Abstract

Many industrial and commercial companies has implemented power DSM project. How to appraise benefit of DSM project is very important. DSM project benefit includes two aspects. One is power consumer side, the other is electric power system side. The former mostly considers economic income by saving electricity. And the latter mostly considers financial and social benefit from it. The thesis sets up a fuzzy comprehensive evaluation model based on fuzzy set theory to assess DSM project benefit from input and income of the company who has implemented power DSM project. The validity of the model is proved with practical examples.

Keywords: DSM; post evaluation; fuzzy set; AHP1 Introduction

The electric power supply and demand pressure in China has never been solved these years. With the rapid development of economy, the imbalance between supply and demand is conspicuous day by day. In order to solve the imbalance fundamentally, on one hand, power plants are constructed in a more cost-effective manner, on the other hand, the mechanism of electric power DSM must be explored to set up at the same time for the future.

The electric power DSM is an integrated systematic plan. A successful DSM project should be the plan from which all of government, power grid enterprise and the enterprise who implemented DSM project benefited. For enterprises who implemented DSM project, in order to provide more overall quantized theoretical foundation, to set up a post evaluation model DSM project benefit is very necessary. It is quite important for promoting DSM project implementation in China.

Some factors of evaluating DSM project benefit are not confirmed. So we adopt fuzzy comprehensive evaluation method.

2　Post Evaluation Model of DSM Project Benefit

2.1　To Determine the Evaluation Factor Set

Input and income are chosen to be the first layer evaluation factor set according to consultation from DSM project experts and based on repeated investigations. As table 1 shows.

$$U=\{\text{Input，Income}\}=\{U_1,U_2\}$$

2.2　To Determine the Weight of Each Factor by AHP

AHP is a kind of method that combines quantitative analysis with qualitative analysis by matrix calculation based on structure model. For the basic principle of AHP, first is to find out relative factors about a complicated environmental problem, and to make sure of their hierarchies, then to make certain of their comparative significance by comparing these factors each other, and finally give their weights.

Table 1 Evaluation Factor Set

Aim layer A	First layer B	Second layer C
DSM project benefit	Input U_1	Initial investment cost C_1
		Operation and maintenance expenses added C_2
	Income U_2	Cost reduced after DSM project being implemented C_3
		Power factor adjustment expenses reduced after DSM project being implemented C_4
		Cost avoided of power being limited by force C_5
		Power consumption avoided C_6
		Fund subsidy C_7

The procedure of determining the weight with AHP method is as follows:

1) to set up hierarchy structure model

The hierarchy structure is as table 1 shows.

2) to construct judgment matrix.

3) to figure out the weight for each hierarchy.

4) to make statistical test about uniformity.

5) to calculate the compositive weight.

The thesis invited 20 correlative experts. They presented the relative quantitative significance for each index. The presented quantity is first averaged and calculated by above five steps to determine the weight of each index. The weight of each index is shown in table 2.

Table 2 the Weight of Each Factor

Aim layer A	First layer B	Second layer C
DSM project benefit	U_1 0.4000	C_1 0.3333
		C_2 0.6667
	U_2 0.6000	C_3 0.2325
		C_4 0.1236
		C_5 0.2180
		C_6 0.3534
		C_7 0.0725

$C_t = (0.3333, 0.6667)$

$C_s = (0.2325, 0.1236, 0.2180, 0.3534, 0.0725)$

$A = (0.4, 0.6)$

2.3 To Determine the Comment Set

The thesis adopts the graduation law of 5 grades. In other words, the comment set of each evaluation factor V={best, better, good, worse, worst}.

2.4 Fuzzy Comprehensive Evaluation of the First Grade

According to the comment set determined, to appraise each factor in the second layer, the membership fuzzy set of each factor in the second layer compared with the corresponding factor in the first layer can be determined.

$$B=(b_{ki1}, b_{ki2}, b_{ki3}, b_{ki4}, b_{ki5})$$

Therein: $i = 1,2$ when k=1

$i = 1,2,\cdots 5$ when k=2

The number value of $(b_{ki1}, b_{ki2}, b_{ki3}, b_{ki4}, b_{ki5})$ can be determined by investigation questionnaire to experts and key staff in the enterprise after DSM implemented.

Considering the influence of every factor synthetically and by the formula (1), the appraisement membership fuzzy set of each factor in the first layer can be confirmed.

$$R = C \circ B \tag{1}$$

2.5 Fuzzy Comprehensive Evaluation of the Second Grade

Based on the fuzzy comprehensive appraisal of the first grade and using the formula (2), the fuzzy comprehensive appraisal result $E = (E_1, E_2, \cdots E_5)$ of DSM project benefit can be determined.

$$E = A \circ R \tag{2}$$

2.6 To Calculate Quantization Score of DSM

Project Benefit

According to the principle that the biggest membership functions, the post evaluation result of DSM project benefit can be confirmed. But for comparing with each other easily, the result can also be changed into the score of hundred-mark system. If best equals to 100, better equals to 80, good equals to 60, worse equals to 40 and worst equals to 20, then the quantization result Q=100 E_1+80 E_2 +60 E_3+40 E_4 + 20 E_5。

3 Application

One enterprise gets the following data through questionnaire investigation after DSM project transformed,

$$B_t = \begin{bmatrix} 0.4 & 0.2 & 0.1 & 0.2 & 0.1 \\ 0.3 & 0.3 & 0.1 & 0.1 & 0.2 \end{bmatrix}$$

$$B_s = \begin{bmatrix} 0.2 & 0.2 & 0.55 & 0 & 0.05 \\ 0.2 & 0.3 & 0.2 & 0.2 & 0.1 \\ 0.3 & 0.2 & 0.2 & 0.2 & 0.1 \\ 0.5 & 0.4 & 0.1 & 0 & 0 \\ 0.05 & 0.05 & 0.5 & 0.4 & 0 \end{bmatrix}$$

According to the formula (1),

$$R_t = C_t \circ B_t = (0.33, 0.3, 0.1, 0.2, 0.2)$$

$$R_s = C_s \circ B_s = (0.3534, 0.3534, 0.2325, 0.2, 0.1)$$

$$R = \begin{bmatrix} 0.33 & 0.3 & 0.1 & 0.2 & 0.2 \\ 0.3534 & 0.3534 & 0.2325 & 0.2 & 0.1 \end{bmatrix}$$

According to formula (2), $E = (E_1, E_2, E_3, E_4, E_5) =$ $(0.3534, 0.3534, 0.2325, 0.2, 0.2)$

To calculate the quantization score, Q=100E_1+ 80 E_2+60 E_3+40E_4+20E_5=89.562$\approx$90。

The post evaluation result of DSM project benefit of this enterprise is lying between better and best from the quantization score. The assessment result accords with actual conditions. The experience is worth popularizing.

4 Conclusions

Enterprises can adopt the comprehensive evaluation model to evaluate its benefit after DSM project implemented. The validity of the model was proved with practical example in the thesis. The evaluation result can offer quantization basis for relevant policymakers.

Acknowledgement

This research was supported by Scientific Technology Research and Development Plan of Hebei Province under Grant 054472261D and the Scientific Research Foundation of the Young Teacher of North China Electric Power University under Grant 200611035.

References

[1] Zhang Fei-Liang, “Comprehensive post-evaluation of fund raising scheme of railway construction project based on fuzzy set theory”, MATHEMATICS IN ECONOMICS, 2005,22(2):154-161(in Chinese)

[2] Yu Rui-feng, Wang Yong-xian, Chen Hai-shou, Peng Hai, “Control-Response Compatibility: Fuzzy Clustering Analysis and a Comparison Among Populations of Different Regions”, Tsinghua Science and Technology, 2004,9(6): pp. 635-642

[3] Wang Jing-min, Wang Zheng-qi, Zhou Feng-hua, “The research for the customers analyzing and evaluating on the power supply enterprise”, Electric Power.,2004,37(4): pp.66-70(in Chinese)

[4] Zeng Ming, Sun Xin, Zhang Qi-Ping, Lv Zhao-Min, Li Tao, “The stimulating system and policy analysis and research on PDSM”, Power DSM. 2003,5(2):3-6(in Chinese)

[5] Blattberg, R. C. and J. Deighton, “Manage marketing by the customer equity test”, Harvard Business Review, 1996, 74(Jul—Aug): pp.136-144

[6] IBM Tokyo Research Laboratory. Aglet software development kit [EB OL]. http://sourceforge.net/projects/aglets, 2004-02-20

[7] N.Borselius, “Moblie Agent Security”, Electronic & Communication Engineering Journal, 2002，10：pp.211～218

[8] Antonio Corradi, “Mobile Agents Integrity Protection for Electronic Commerce Application”, Information System，1999, 24(6)：pp.519～533

[9] Pan Feng, Cheng Hao-zhong, Yang Jingfei, “Power system short-term load forecasting based on SVM”. Power system Technology, 2004, 28(21): pp.39~42

[10] Xian Xiaodong, Xiong Qing-yu,, “A model for university student credit fuzzy evaluation based on neural network”, Computer science, 2007,34(9): pp.203~205

The Improved Load Forecasting Model of BP Neural Network

Yanmei Li

School of Business and Administration, North China Electric Power University, Baoding, Hebei 071003, China
E-mail: liyanmei28@yahoo.com.cn

Abstract

The traditional BP Neural Network forecasting model influenced clearly by the complexity of network frame and sample, which lead to the over-learning or low-extensive ability. The paper puts forward a new model to improve the BP Neural Network load forecasting model, which use attribution reduction algorithm of rough sets to reduce the various historical data related to the load and eliminate the property not related to the decision-making information. It was tested that this method reduces the input variables of the BP Neural Network, so shortens the training time of the BP Neural Network load forecasting model. At last the forecasting capability is improved.

Keywords: BP Neural Network; Load Forecasting Model; Rough sets; Attribution Reduction

1 Introduction

It's been a long time that many electric operators commit themselves to the investigation of electric system load forecast technique and have obtain some production to a certainty, for example time sequence, regress, analytical method, gray theory, artificially NN and so on. The artificially NN has powerful collateral disposal mechanism, imminent capability of discretionary function, learning capability and self-organize and self-adapt capability. And it is capable of considering the impact of variable factors such as weather, temperature and so on. So it has been widely applied in the field of load forecast and decision-making. But on the other hand, the NN also has some limitation and shortage: because the learning speed is fixed, the net convergence speed is slow and it needs quite training period; and for the complicated problems, the net's input nodes are too much, and it obviously lead to the complexity and training period of the net. And in this way, even we can convergence the power data to some value, it would not be the complete least value of the error plane, and possibly the least value of the part plane.

In 1982, the Poland scientist proposed the Roughness Sets theory (RS), the characteristic of which is from the description gather of the ready problems directly, and find the internal rule. It can dispose the uncertainty and redundancy information preferably. This paper will introduce several Attribute Reduction Algorithm at first, then apply these methods to reduction the factors of various historical data related to the load, and eliminate the property irrelevant to the decision-making, and predigest the input variable of the BP NN. To obtain the representative samples by through elapsing the redundancy information, and training network through the typical samples, and refining study samples, then confirm the input variables layer neurons number, and optimization of the hidden layer neurons, and in this way we can improve the training speed and precision. It is proofed that rough intensive reduction algorithm with improved model is better than the traditional neural network model prediction.

2 Rough Sets Reducton Algorithm

2.1 Use the Dicernibility Matrix to Get the Least Reduction

The Discernibility Matrix is proposed by the Poland Warsaw famous mathematician Skowron.

Utilizing this tool, we can give expression to the whole cannot distinguish relations which consist in the complicated information system. One presupposition to use the dicernibility matrix to get the least reduction is to deal with the antipathic register at first at the data in table pretreatment stage, that is, the discermibility matrix not deal with the incompatibility record. With the pretreatment methods, if the records of the conflict by dividing the number of the total number of records, we can get a measurement of the roughness and this measurement can be a character of the data sheet.

Suppose $S=\left(U,R,V,f\right)$ is a information system, R=CYD is property collection, subset $C=\left\{a_i/i=1,2,\cdots,m\right\}$ and $D=\{d\}$ is Conditional attribute set and decision-making attribute set, $U=\left\{x_1,x_2,\cdots,x_n\right\}$ is on domain, $a_k(x_j)$ is sample, x_j is the value on the a_k。Definition system get the dicernibility matrix is $M\left(S\right)=\left[m_{ij}\right]_{n\times n}$, it's i and j the list element is

$$m_{ij}=\begin{cases}a_k\in C,a_k\left(x_i\right)\neq a_k\left(x_j\right)\wedge D\left(x_i\right)\neq D\left(x_j\right)\\ \phi,D\left(x_i\right)=D\left(x_j\right) \qquad i,j=1,2,\cdots,n\end{cases}$$

So in the dicernibility matrix the element m_{ij} is capable of distinct the whole collection to the x_i and x_j ; But if x_i and x_j belong to the same decision-making class,the value of the m_{ij} is ϕ in the dicernibility matrix。

For every dicernibility matrix *M*(*S*) corresponding the only distinguish function $f_{M(S)}\left(a_1,a_2,\cdots,a_m\right)=\wedge\left\{\vee m_{ij},1\leq j<i\leq n,m_{ij}\neq\phi\right\}$。

Through the dicernibility matrix we can get the R attribute and a nuclear reduction expediently, and as the nuclear attributes the starting point. Then seek the smallest disjunctive of the different function, and the disjunctive computing can be greatly simplified. At last disjunction each component corresponding to a reduction, so we will be able to receive the smallest reduction.

2.2 Reduction Algorithm Based on the Attribute Dependability

Strike all about a history of NP-hard problem, so enlightening information to use simplified calculation to identify the optimal or sub-optimal reduction can be achieved is obviously a method.

Many basic steps of heuristic algorithms are incepted by the nuclear of the information systems or table, then some measure of the importance of attributes, choose the most important attributes to the nuclear, until the termination conditions, and we have the information systems or a decision table reduction. To be more exact, includes a reduction of the attribute set.

An information system for all the attributes of the decision-making is not as important as in the rough set theory, attribute importance of the dependencies can be reflected.

Decision-making property D to the property R (R belongs to C), the dependencies $\gamma\left(R,D\right)$ defined:

$$\gamma\left(R,D\right)=\frac{card\left(POS_R\left(D\right)\right)}{card\left(POS_C\left(D\right)\right)}$$

Obviously, $0\leq\gamma\left(R,D\right)\leq1,\gamma\left(R,D\right)$ give a measure of the D decision-making on the attributes of the dependencies. It reflect the importance of the property R to D. Under the premise known conditions R, a property $a\in C-R$ to the decision-making D's importance $SGF\left(a,R,D\right)$ may be defined as follows:

$$SGF=\left(a,R,D\right)=\gamma\left(R+\left\{a\right\},D\right)-\gamma\left(R,D\right)$$

$SGF=\left(a,R,D\right)$ reflect that after property a plus R, the rise degree of the dependence between R and D. And in fact, the stronger for the impact of a property attribute R and D dependencies, the greater value of $SGF=\left(a,R,D\right)$ is.

2.3 Reduction Algorithm based on Conditional Information Entropy

This reduction algorithm does not require distinction matrix, directly from the point of view of information theory, the idea of entropy calculated using various attributes the importance of the load. Reduction act and algebra is the same in the information entropy reduction in the importance of also need to set a threshold value greater than the threshold value of the property that is affected relatively large load factors, and less than the threshold value of the attribute is considered to be redundant and could be deleted.

In this method for the calculation of the importance of attributes as: $I(a_i, D) = \left[H\left[\frac{R_D}{R_{C-A}}\right] - H\left[\frac{R_D}{R_C}\right]\right] / H\left[\frac{R_D}{R_C}\right]$, C represent attribute set, that is, the various factors affecting load; D represent decision-making set, that is, the load of the forecast day; RC,RD represent separately the equivalence relation in the decision-making domain of the attribute set conditions C and D; $H\left[\frac{R_D}{R_C}\right]$ represent RD to RC's condition entropy.

2.4 Attribute Reduction based on Genetic Algorithm

Adaptive genetic algorithm is a method of random search, the search methods, not by a unitary structure or direction, it will more than individual as a possible solution and to consider the search space within the scope of the overall sample, thus leading to the possibility of greater convergence Global Optimal Solutions, Therefore, it was the introduction of the genetic algorithm Rough Set Attribute Reduction. Algorithm by using computer simulation of biological evolution, groups continues to be optimized, in the process of change and to identify the optimal solution. In the genetic algorithm, the design of fitness function is the core of the entire GA algorithm steps, as several genetic operator rely on the adaptation of the chromosome and therefore fitness function of design objectives, determines to a large extent iteration convergence direction, but the rough set of attribute reduction is to achieve the smallest reduction attribute set. In this way, the attribute set to meet in ensuring the accuracy of certain circumstances, at least to the number of attributes that ultimately the results that are needed to meet the classification requirements of the attribute set. Therefore, the design function to the ultimate goal should contain the following two objective function: ① classification must meet quality requirements usually must be about SR. ② covered by the reduction of the number of properties to minimize.

Then provisions for the fitness function: $score(r) = \frac{m - l_r}{m} + k$

R for which chromosome the corresponding attribute set, D for decision-making attributes, the attribute set dependence on R. This function will attribute dependence on the introduction of fitness function, and rely on the attributes of that decision attributes of the corresponding attribute set chromosome depended reflects the attributes of taxonomic capacity.

3 BP Traditional Neural Network Prediction Model

The paper title (on the first page) should begin 1.38 inches (35 mm) from the top edge of the page, centered, completely capitalized, and in Times 14-point, boldface type. The authors' name(s) and affiliation(s) appear below the title in capital and lower case letters. Papers with multiple authors and affiliations may require two or more lines for this information.

BP forward neural network is classified as a feed-forward neural network, it is very self-learning and self-organization capabilities, but also of non-linear, non-convex, and other characteristics. It reverse adjustment goals and the actual output to correct the error weights to achieve network prediction accuracy.

K samples with vector network input layer neurons for the number of middle layer neurons to p, the number of neurons in the output layer for m, the network input value $P_k = (x_1, x_2, \cdots, x_n)$, output value $Y_k = (y_1, y_2, \cdots, y_m)$, expected output value $T_k = (t_1, t_2, \cdots, t_m)$, wji is input layer to the middle layer of the connection weights, Vsj for the middle layer connected to the output layer to the middle layer θ j unit threshold, the output layer γ s unit threshold, and $i = 1, 2, \cdots, n, j = 1, 2, \cdots, p, s = 1, 2, \cdots, m$.

Step 1, the samples were normalized vector processing, data processing (0,1) between data value and the right to the

threshold to (-1,1) between the initial random, select a group of input and objectives of samples provided to the network.

Step 2, computing the hidden layer and output layer unit of input and the corresponding output.

$$u_j=\sum_{i=1}^{n} w_{ji}x_i+\theta_j, h_j=f\left(u_j\right), d_s=\sum v_{sj}h_j+\gamma_s, o_s=f\left(d_s\right)$$

Step 3, calculate the output layer error and hidden layer error according to the network output $\delta_s=\left(o_s-t_s\right)\left(1-o_s\right), \eta_j=\left[\sum_{s=1}^{m}\delta_{si}v_{sj}\right]h_j\left(1-h_j\right)$

Step 4, Error adjusted value of the use of layers weights and threshold adjustments

$$v_{sj}=v_{sj}+\alpha\delta_s h_j, w_{ji}=w_{ji}+\alpha\eta_j x_i, \gamma_s=\gamma_s+\alpha\delta_s, \theta_j=\theta_j+\alpha\eta_j$$

Step 5, select the sample vector to provide a learning network, to return to Step 2 until the global error E is less than a pre-determined value, the study concluded.

This paper first application of the above-mentioned Rough intensive SR BP neural network algorithm for the reduction of input variables, and then select the training vector samples will be available to the network, trained network load data projections.

4 Case Analysis

In this paper, a southern city of the 2005 data were analyzed to August 10, 2005 to August 19 points for the entire sample for the study and active load to August 20 for the entire load for the test samples were forecast in August on the 21st load. 55 selected by experience on the condition variable attributes, which is 12 load data, that is, on the 10th to the 19th day the whole point of load; The remaining 43 non-load data, including weather, the date type, sunshine duration, maximum temperature, minimum temperature, average temperature, the biggest humidity, humidity, such as minimum 43 factors, including the date and type of rest days (including weekends and the statutory rest), the weather conditions on the provision of meteorological information is divided into 17 types. The neural network input vector, using the above-mentioned were all rough intensive SR algorithm primaries to the impact of load reduction factors. The use of different matrix reduction, the remaining variable 37; rely on the use of attributes of a reduction, the remaining variable 28; if information entropy conditions for a reduction, the remaining variable 21; final by genetic algorithm reduction, the remaining 27 are variable. Which attributes Reduction Algorithm dependence on the importance of setting the threshold value of 0.06, conditions information entropy reduction algorithm set threshold of the importance of 0.6. Table 1 lists a variety of reduction algorithm results. Can be seen, after attribute reduction, the input vector be simplified.

Table 1 Forecasting results using single attribute reduction method

Method	Attribute reduction	Hidden layer units	Time/min	RSME/%
Reduction ago	43	13	55	2 7328
discernibility matrix	37	13	13	1 6546
attribute dependability	28	9	11	1 9976
conditional information entropy	21	15	5	1 3245
Genetic Algorism	27	14	9	1 8729

First of all input and output variables on the regulation of a processing, data processing interval [0,1] between the data. Regulation is one of the many ways here by the following formula:: $\hat{x}=\frac{x-x_{\min}}{x_{\max}-x_{\min}}$ 。

In a 3-layer BP neural network model to predict the hidden layer neurons of the first set to the number of $n_1=\sqrt{n+m}+a.$ Among them, m is the output neuron number, n is the number of input neurons, a [1,10] between the constants. Network layer neuron transfer function using S-type tansig tangent function, the output layer neuron transfer function using S-logarithmic function logsig, because the output function in the interval [0,1] between the output precisely meet the requirements. The learning algorithm, a faster convergence Levenberg-Marquardt dynamic numerical optimization algorithm. Training error indicators set to 0.01, taking into account the network structure more complicated, the number of training set for 1000, adopted by the end of the forecast reduction algorithm results such as shown in table 2.

Table 2 Comparation of load forcasting results

Time unit	Actual data	discernibility matrix		attribute dependability		conditional information entropy		Genetic Algorism	
		Forcasting data	Relative error	Forcasting data	Relative error	Forcasting data	Relative error	Forcasting data	Relative error
1	0.2119	0.2817	−0.0698	0.1813	0.0306	0.2104	0.0015	0.1576	0.0543
2	0.1215	0.1908	−0.0693	0.2025	−0.0810	0.1330	−0.0115	0.0698	0.0517
3	0.1612	0.2119	−0.0498	0.2718	−0.1097	0.1435	0.0186	0.2164	−0.0543
4	0.2161	0.3323	−0.1162	0.1001	0.1160	0.1750	0.0411	0.2858	−0.0697
5	0.6171	0.8916	−0.2745	0.4889	0.1282	0.5020	0.1151	0.8976	−0.2805
6	0.6159	0.5508	−0.0651	0.4992	0.1167	0.5389	0.0770	0.6679	−0.0520
7	0.7155	0.8839	−0.1684	0.5038	0.2117	0.6802	0.0353	0.7838	−0.0683
8	0.7201	0.6325	−0.0876	0.7424	−0.0223	0.9965	−0.2764	0.5625	0.1576
9	0.7243	0.5617	−0.1626	0.7808	−0.0565	0.6822	0.0421	0.6669	0.0574
10	0.7298	0.6636	−0.0653	0.7782	−0.0493	0.7323	−0.0025	0.6831	0.0458
11	0.8179	0.7845	−0.0334	0.8832	−0.0653	0.7871	0.0308	0.8879	−0.0700
12	0.8229	0.8928	−0.0699	0.8026	0.0203	0.7913	0.0316	0.8536	−0.0307

Integrated Table 1 and Table 2 shows that the use of information entropy reduction, the number of input variables reduced to 21 from 43, the computing time is 9 min, not only variable was the greatest degree of reduction, and the relative computing the shortest time, Operational results of the relative error minimum, this method in the region suitable for load forecasting samples; By contrast, the difference matrix, and the dependence on attributes such as genetic algorithms reduction in the variable method of reduction, and the computing time is not on the relative error for a sample of the region computing.

5 Conclusions

Rough intensive use of simple algorithm to affect the input variable load reduction, not only can choose from a greater impact on the load factor, but also to eliminate the correlation between these factors, reducing the number of input variables. This will not only reduce the complexity of the network, the network can also shorten training time and improve the accuracy of forecasts, with practical examples, closer to prove the reliability and validity.

Acknowledgement

This research was supported by the Scientific Research Foundation of the Young Teacher of North China Electric Power University.

References

[1] JIANG Yang-yi, "ulti-factor forecast models based on rough sets and BP neural network" computer engineering, 2007,33 (5): 154 ~ 155

[2] ZHANG Qing-Bao and the others, "hort-term load forecasting based on rough sets attribute reduction algorithm and Support Vector Machine", grid technology, 2006,30 (8): 56~60

[3] LIU Gang, GU Yu-gui, "Short-term load forecasting based on improved BP neural network" , DSM, 2005,7 (3): 24~27

[4] LIU Xiao-jie, CAO Li-ming,WANG Xiao-ping," The distribution center selection based on the fuzzy neural network model,Computer Applications and Software", 2007, 24(3): 15~17

[5] ZHU Wen-xi, SHAN Mi-yuan, "A neural network model of risk analysis for public construction project, Journal of Harbin Engineering Unicersity", 2006, 27:142~147

[6] XIAN Xiao-dong, XIONG Qing-yu,etc, "A model for university student credit fuzzy evaluation based on neural network", Computer science, 2007, 34(9):203~205

[7] CHEN Shu-zhen, "Several attributes reduction algorithm analysis based on rough set", Wuhan Institute of Industrial Journal, 2005,24 (3): 118-121

[8] CHEN Yao-wu,WANG Le-yu,LONG Hong-yu.,"Short-term load forecasting with modular neural networks" [J]. Proceedings of the CSEE, 2001, 21(4): 79~82

[9] PAN Feng, CHENG Hao-zhong,YANG Jing-fei, etc," Power system short-term load forecasting based on SVM" [J]. Power system Technology, 2004, 28(21):39~42

[10] XIE Hong, CHENG Hao-zhong, ZHANG Guo-li, "Load forecasting model of neural network based on rough set theory", China Electrical Engineering Journal, 2003,23 (11): 1~4

Study on Semi Supervised Clustering by Metric Learning

Xiuqin Jiang Shitong Wang

School of Information Technology, Jiangnan University, Wuxi, Jiangsu, China
Email:jxq8405@yahoo.com.cn

Abstract

Semi Supervised methods use a small amount of auxiliary information as a guide in the learning process in presence of unlabeled data. When using a clustering algorithm, the auxiliary information has the form of side information, that is a list of co-clustered points. The use of Semi Supervised methods may be useful especially in very difficult tasks, such as biological experiments. There are two frequently methods in Semi Supervised clustering: one is constraint-based methods that guide the clustering algorithm towards a better grouping of the data, the other is distance-based learning .In this paper, how to develop new metric learning fuzzy clustering-based is importantly discussed here.

Keywords: semi supervised learning, fuzzy clustering, labeled data, metric learning, data sets

1 Preface

With the development of the computer and the demand of the actual problem, the method of based on target function becomes to the main methods of clustering analysis.

Clustering is a process of unsupervised learning, it has no labeled data, so its uncertain is larger than the supervised learning. Between the supervised and unsupervised learning is semi supervised learning, It is the study mechanism that develops recently.

Semi supervised methods are methods that use a small amount of auxiliary information to guide the other techniques, providing the data analyst with a more flexible tool to use all the available a priori knowledge, so that it can get better result.

There are many reasons for considering semi supervised methods: often labeled data are expensive or impossible to obtain whereas unlabeled data are abundant and easy to collect; often some a priori information is available or easily obtainable from unlabeled data; often labeling is based on human experts judgments and so is prone to errors and subjectivity.

In this paper we will introduce semi supervised fuzzy clustering algorithms.

2 The Brief Introduction of the Main Clustering

Clustering analysis is widely used in the field of pattern recognition, data mining and fuzzy controlling, and as the development of the economy, also it gets much development in recent years, and people get much focus on the clustering analysis.

So far there are many clustering methods. Actual we use the fuzzy clustering based on the target function. Bezdek generalized the fuzzy clustering. From now on, the fuzzy clustering based on target function developed quickly, already becomes a huge system. These can be seen in many of the articles.

In this article, we main discuss fuzzy c-means clustering algorithm(FCM),which is widely used in many field, so it can be looked in the article easily.

FCM divides n vectors xi (i=1,2,…,n) into c fuzzy sets, and calculate the center of each clustering, make the no likeness index function value to the minimum. FCM uses fuzzy to divide, make each data use the value between 0 and 1 to assure the belonging. Belonging matrix U allows the value between 0 and 1.However, the sum of the belonging of a data set is 1:

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

Then，the target function of FCM is the following form:

$$J(U, c_1,...,c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j}^{n} u_{ij}^m d_{ij}^2 \tag{2}$$

Here u_{ij} is between 0 and 1; ci is the center of the fuzzy set i, $d_{ij} = \|c_i - x_j\|$ is the Euclidean metric of the ith clustering center and the jth data point; and $m \in [1, \infty)$ is a weighted index.

We can see design the above target function can get the essential condition to make(2)to get the minimum value,and the essential condition to make (2) to get the minimum value is :

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m} \tag{3}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \tag{4}$$

3 Semi Supervised Clustering

Already existed of the semi supervised methods are dividedinto two kinds of method，we call them based on constrained and metric learning.

The method of constrained based is talked about in many of the articles, and its objective function of a clustering algorithm is modified to include a penalty for wrongly classified points, In this paper, we main talk about metric learning.

In semi supervised clustering algorithm based on metric learning, we focus on our attention on metric learning, this is very recent and flexible in the choice of the metric function. so we use a suitable distance function to make the similar points more closer and to make the dissimilar points more away. This is very vivid to the method of the distance calculation. In[1] Mahalanobis distance is learned.This is a very effective approach, although severely limited in application to real data by the computational complexity O(d^6).where d is the dimension of data. In[2] the metric is learned considering pairs of samples belong to the same class and the computational complexity is reduced to O(d^3).Here we consider a different metric for each cluster and so allows for clusters of different shape. We use a two steps approach, separating metric learning from clustering.

3.1 Metric learning

The choice of a metric for a given experiment is not an easy task. Often a specific metric is suitable for some data and completely unsuitable for other. In a general framework, we call $Z = \{X_1,...,X_n\}$ the set of data consisting of n d-dimensional data points, $S \subset Z \times Z$ a set of pairs of similar points and $D \subset Z \times Z$ a set of pairs of dissimilar points. $(X_P, X_q) \in S$ means that the two vectors X_p and X_q are known to be in the same cluster, and vice versa $(X_P, X_q) \in D$ means that the two vectors X_p and X_q are known not to be in the same cluster,we look for a function f to respects. $f(X_p, X_q)_{(X_p, X_q) \in S}$ is minimized and $f(X_p, X_q)_{(X_p, X_q) \in D}$ maximized.

3.2 Side information

The auxiliary information can have many forms and should be modeled accordingly. For clustering, auxiliary information has generally the form of side information, That is Must-Link (ML) and Cannot-Link (CL): for each dataset, an expert declares the list of points that he believes should be co-clustered and the list of points that should not. We will assume, without loss in generality, that the set ML of co-clustered points and the set CL of non-co-clustered points have the following structure:

$$ML \subset \left\{ \left(X_{m_p}, X_{m_q} \right) | X_{m_p}, X_{m_q} \in Z; l_{m_p} = l_{m_q} \right\} \tag{5}$$

$$CL \subset \left\{ \left(X_{m_p}, X_{m_q} \right) | X_{m_p}, X_{m_q} \in Z; l_{m_p} \neq l_{m_q} \right\} \tag{6}$$

Where $m_p \in \{1,...n\}$ and l_{m_p} is the a priori known label of point X_{m_p}.

As we will learn a different metric for each cluster, we will need to subset the ML and CL matrices in order to have the specific side information relative to each cluster j:

$$ML_j \subset \left\{ \begin{matrix} \left(X_{m_p}, X_{m_q}\right) | \, X_{m_p}, X_{m_q} \in Z; \\ l_{m_p} = j, l_{m_q} = j \end{matrix} \right\} \tag{7}$$

$$CL_j \subset \left\{ \begin{matrix} \left(X_{m_p}, X_{m_q}\right) | \, X_{m_p}, X_{m_q} \in Z; \\ l_{m_p} \neq j, l_{m_q} \neq j \end{matrix} \right\} \tag{8}$$

3.3 Semi supervised fuzzy c-means clustering algorithm

The semi supervised fuzzy c-means clustering algorithm in this article has three points: the first one is we learn a specific metric for each cluster, the second is the learned metric are applied for the distance computation in the fuzzy c-means, the third is the weights optimization process is based on a stochastic search.

The method is realized in two steps: in the first step we use the a priori information to “tweak” the metric f_{A_j}

$$f_{A_j}\left(X_{p,} X_q\right) = \left[\left(X_p - X_q\right)^T A_j \left(X_p - X_q\right)\right]^{1/2} \tag{9}$$

To gain more generality and more flexibility, we considered a different matrix for each cluster A_j, j=1,…k. We demand that samples declared to be “similar” have small squared distance and samples declared to be “dissimilar” have high squared distance。 $ML_j$and CL_j respectively the sets of similar points and the set of dissimilar points in the jth cluster, then we pose a set of k constrained problems:

$$\min \sum_{\left(X_p, X_q\right) \in ML_j} \left\| X_p - X_q \right\|^2_{A_j} \tag{10}$$

$$\sum_{\left(X_p, X_q\right) \in CL_j} \left\| X_p - X_q \right\|_{A_j} \geq t \, , A_j \geq 0 \tag{11}$$

where j=1….k, t>0 is an arbitrary constant. If under the assumption of diagonal matrix A_j , to the minimization of the following k convex functions:

$$g\left(A_j\right) = \sum_{\left(X_p, X_q\right) \in ML_j} \left\| X_p - X_q \right\|^2_{A_j} - \log\left(\sum_{\left(X_p, X_q\right) \in CL_j} \left\| X_p - X_q \right\|_{A_j} \right) \tag{12}$$

Here we adopted a stochastic search based minimization algorithm based on the well-known Simulated Annealing (SA) method. A general schema of the algorithm in the formulation that we used follows:

L1: set starting temperature T0;

L2: compute energy function E;

L3: repeat until convergence or maximum number of iteration is reached:

--go to a neighbour state through a Gaussian perturbation of the current state;

--compute the energy variation EE ;

--if EE <0 accept the new state;

--if EE >=0 accept the new state with probability given by the Metropolis function;

$$\exp^{-EE/T}$$

--decrease temperature according to logarithmic cooling schedule

$$T(u) = T0/\log(u + alpha) \tag{13}$$

where u is the epoch counter and alpha is a free parameter.

In the second step, we calculate the fuzzy c-means clustering with more general distance metrics calculated previously：

$$J = \sum_i^n \sum_j^k f_{A_j}\left(X_i, m_j\right) U_{ij}^b \tag{14}$$

m_j is the center of the jth cluster.

4 Experiment

We tested the semi supervised fuzzy c-means versus conventional unsupervised fuzzy c-means to evaluate its efficacy. As our task was to validate and efficiency of the various methods and not to discuss cluster number issues, we considered the known true clusters number of each dataset. In the simulated annealing algorithm we used T0=1000 and alpha=5.The side info matrix ML has been generated in the fixed amount of 10% of all data choosing randomly .The side info matrix CL was generated choosing randomly half of the points from the ML matrix. The initial centroids were generated randomly and independently and the overlap index b was chosen to be 1.5 for both algorithms. The data for the experiments have been obtained from the UCI Machine Learning repository.

4.1 Iris data

The iris dataset IR is the famous dataset of Fisher that contains measurements of three species of iris plant. It has 150 instances,4 features,3 classes. From the experiments, we get the belonging of the data to the classes under the condition of the data belong to this class definitely (Figure 1 The Membership of Iris Data). The solid line stand for semi supervised fuzzy c-means clustering algorithm, the dot line stand for conventional fuzzy c-means clustering algorithm.

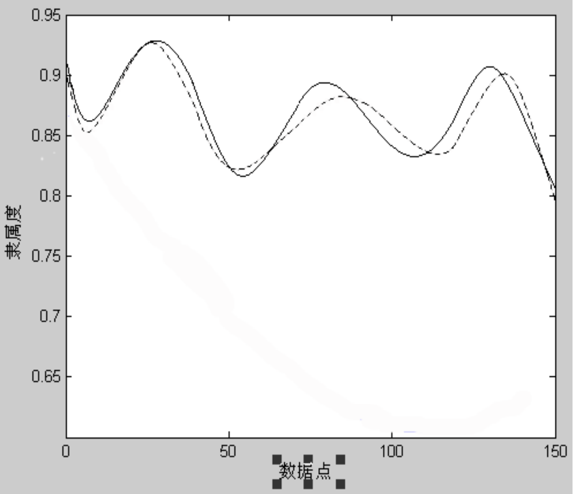


Figure 1 The Membership of Iris Data

The coefficient matrix A is:

Table 1 The coefficient of iris data

To the first class	[0.9829,0,0,0;0,1.1476,0,0; 0,0,1.0402,0;0,0,0,1.0048]
To the second class	[1.0953,0,0,0;0,0.9977,0,0; 0,0,1.0339,0;0,0,0,1.0119]
To the third class	[1.2019,0,0,0;0,0.5791,0,0; 0,0,1.0278,0;0,0,0,1.4108]

4.2 Thyroid data

We get the part of the thyroid data, it has 215 instances,5 features,3 classes. From the experiment, we get the belonging of the data to the classes under the condition of the data belong to this class definitely. (Figure 2 The Membership of Thyroid Data). The solid line stand for the semi supervised fuzzy c-means clustering algorithm, the dot line stand for the conventional fuzzy c-means clustering algorithm.

From the experiments, we can see that using the metric learning makes the most of the belonging of the data to the classes much higher. So, this appears the algorithm in this article valid.

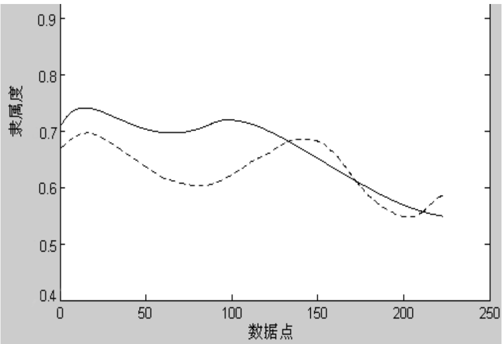


Figure 2 The Membership of Thyroid Data

The coefficient matrix A is:

Table 2 The coefficient of thyroid data

To the first class	[1.0804,0,0,0,0;0,0.8723,0,0,0;
	0,0,0.7676,0,0;0,0,0,0.7087,0;
	0,0,0,0,0.9881]

To the second class	[1.0199,0,0,0,0;0,1.0088,0,0,0;
	0,0,1.1346,0,0;0,0,0,0.6721,0;
	0,0,0,0,1.2317]

To the third class	[1.0316,0,0,0,0;0,0.9505,0,0,0;
	0,0,0.8002,0,0;0,0,0,1.0887,0;
	0,0,0,0,1.0454]

5 Conclusion

Semi supervised clustering technology is a new field. Semi supervised methods use a small amount of auxiliary information as a guide in the learning process in presence of unlabeled data. We use semi supervised method into fuzzy clustering, so semi supervised clustering is created. This algorithm is used widely now, such as picture partition, and we need to do further research on semi supervised clustering.

References

[1] E.P.Xing, A.Y.Ng,M.I.Jordan,S.Russell, Distance metric learning, with application to clustering with side-information, Advances in Neural Information Processing Systems 15 (2002)

[2] M.Bilenko, S.Basu,R.J.Mooney, Integrating constraints and metric learning in semi-supervised clustering, in: Proceedings of the 21st ICML,2004,pp.81-88

[3] Sinkkonen J, Kaski S.Clustering based on conditional distributions in an auxiliary space [J].Neural Computation, 2001(14) : 217- 239

[4] DEM IR IZ A, BENNETT KP, EMBRECHTSMJ. Semi supervised clustering using genetic algorithms [A]. Artificial Neural Networks in Engineering [C]. ANN IE299: 809 - 814

[5] M.Ceccarelli, A.Maratea, Semi-supervised fuzzy c-means for the analysis of biological data, Lecture Notes in Artificial Intelligence 3849 (2005) 259-266

[6] D.Demb, P.Kastner, Fuzzy c-means method for clustering microarray data, Bioinformatics 19 (2003) 973-980

[7] M.B.Eisen, P.T.Spellman, P.O.Brown,D.Botstein,Cluster analysis and display of genome-wide expression patterns, PNAS 95 (1998) 14863-14868

[8] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast Saccaromyces Cervesiae by microarray hybridization, Molecular Biology of the Cell 9 (1998) 3273 – 3297

[9] L. Zhengdong, T. Leen, Semi-supervised learning with penalized probabilistic clustering, in: Proceedings of Neural Information Processing Systems 2004, vol. 17, 2004

[10] M. Schultz, T. Joachims, Learning a distance metric from relative comparisons, in: Proceedings of Neural Information Processing Systems 2003, vol. 16, 2003

[11] T.R.Golub,D.K.Slonim,P.Tamayo,C.Huard,M.Gaasenbe ek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri,C.D.Bloomfield,E.S.Lander,Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531 – 537

An Efficient and Effective Outlier Detection Method Based on Voronoi Diagram*

Jilin Qu

School of Computer and Information Engineering, Shandong University of Finance, Jinan, Shandong, 250014, China
Email: qujilin@126.com

Abstract

Outlier detection is an important problem for many domains and has attracted much attention recently. The distance-based method is widely used in application. However, the complexity of the method is quadratic to size of the dataset, and it will be no effect when the data points exhibit different densities in different regions of the dataset. In this paper, we propose a new outlier detection method based on Voronoi diagram, called Voronoi based Outlier Detection (VOD), to provide highly-accurate outlier detection and reduces the time complexity from $O(n^2)$ to $O(n\log n)$.

Keywords: Data Mining, Outlier Detection, Distance-based, Voronoi Diagram, Algorithm

1 Introduction

Outlier detection has many important applications in financial surveillance, marketing and fraud detection. Mining outliers in database is to find exceptional objects that deviate from the rest of the data set 0. Methods for outlier detection in large data sets are drawing increasing attention. Various data mining algorithms for outlier detection has been proposed. The approaches can be classified into distribution-based, depth-based, clustering, distance-based, and density-based method 0.

Current techniques typically incorporate an explicit distance metric, which determines the degree to which an object is classified as an outlier. Distance-based outlier detection method was first proposed by Knorr and Ng 0. A point p in a data set S is a distance-based outlier (DB(ρ, r)-outlier) if at most a fraction ρ of the points in S lie within distance r from p. It has an intuitive explanation that an outlier is an observation that is sufficiently far from most other observations in the data set, and widely used in application. However, the naive complexity of this technique is quadratic due to pair wise comparison, and it will be no effect when the data set exhibit different densities in different regions of the data space or across time.

Bay & Schwabacher (2003) propose a simple pruning optimization that reduces complexity given a randomly ordered dataset 0. The authors report sub-quadratic time performance in the number of data points, but Ghoting et al. (2006) has proved it is unable to provide near-linear time performance 0. A Recursive Binning and Re-Projection (RBRP) algorithm was proposed in 0, which scales log-linearly as a function of the number of data points and linearly as a function of the number of dimensions. But the worst case time complexity of RBRP is O (n^2).

Tao et al. (2006) present the outlier detection method SNIF which is able to accommodate arbitrarily large datasets in three scans or the dataset through prioritized flushing 0. Priorities are assigned based on the likelihood that the object will be an outlier or a non-outlier with relatively few neighbors.

Angiulli & Pizzuti(2005) propose a new definition of distance-based outlier and an efficient algorithm, called HilOut, designed to detect the top n outliers of a large and high dimensional data set0, the temporal cost

* Supported by the Science and Technology Key Projects of Shandong Province under Grant No.2007GG3WZ10010; Doctoral Scientific Research Foundation of Shandong University of Finance under Grant No.06BSJJ09.

of the algorithm is $O(dn^2)$.

In this paper, we propose a new method based on Voronoi diagram, called Voronoi based Outlier Detection (VOD), to overcome the problems of distance -based outlier detection. The main contributions of this paper are as follows:

- We use the Voronoi nearest neighbor to calculate the outlier factor of a data point. With respect to the distance -based outlier detection method, the VOD performs better in identifying local outliers that deviate from the main patterns.
- The VOD outlier detection algorithm more efficient than the distance-based method. The running time of our algorithm is $O(n\log n)$, where n is the size of dataset.

The rest of this paper is organized as follows. In section 2, we introduce the basic properties of Voronoi diagram and describe our VOD method. Section 3 presents an experimental evaluation, and we conclude in Section 4.

2 Proposed Method

In this section we first introduce the basic properties of Voronoi diagram, and then describe our VOD method.

2.1 Preliminaries

Definition 2-1. (Voronoi diagram). Given a set S of n points $p_1, p_2, \ldots, p_n$ in the plane, the Voronoi diagram, denoted as *Vor(S),* is a subdivision of the plane into Voronoi cells. The Voronoi cell, denoted as $V(p_i)$ for p_i ,to be the set of points q that are closer or as close to p_i than to any other point in S . That is

$$V(p_i)=\{ q |\ dist(p_i, q) \le dist(p_j, q) , \forall\ j\neq i \}$$

where dist is the Euclidian distance function.

See Figure 1 for an example.

The Voronoi diagram decomposes the plane into n convex polygonal regions, one for each p_i. The vertices of the diagram are the *Voronoi vertices*, and the boundaries between two Voronoi cells are referred to as the *Voronoi edges*. The boundaries of a Voronoi cell $V(p_i)$ is a Voronoi polygon having no more than n-1 edges.

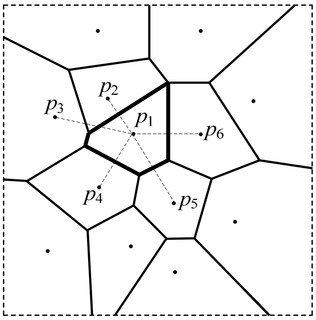


Figure 1 Voronoi diagram

Voronoi diagram contains all of the proximity information defined by the given set 0. It is one of the most important structures in computational geometry, and has been widely used in clustering, learning, graphics, and other applications. We focus on the properties of the Voronoi diagram related to the outlier detection.

Theorem 2-1. Every nearest neighbor of p_i defines an edge of the Voronoi polygon $V(p_i)$ 0.

Theorem 2-2. Every edge of the Voronoi polygon $V(p_i)$ defines a nearest neighbor of p_i0.

Theorem 2-3. For $n \ge 3$, A Voronoi diagram on n points has at most $2n$-5 vertices and $3n$-6 edges 0.

Theorem 2-4. The Voronoi diagram of a set of n points can be constructed in $O(nlogn)$ time and this is optimal 0.

There are four well-known algorithms for constructing the Voronoi diagrams, including divide-and-conquer, randomized incremental, plane sweep, and reduction to convex hulls.

Theorem 2-5. With the Voronoi diagram, nearest neighbor search can be performed in $O(logn)$ time, which is optimal 0.

2.2 The VOD method

The Voronoi diagram captures the proximity uniquely. We address the outlier detection by refining the concept of a neighborhood with the Voronoi diagram.

Given a data set S, the neighborhood relationship is the inherent properties of the data set. For a point $p_i \in S$, each edge of the Voronoi polygon $V(p_i)$ defines a nearest neighbor of p_i. The numbers of nearest neighbor vary for different points; it can't be of a fixed number k. Once the polygons are formed, it creates a periphery of the immediate neighborhood in the form of neighborhood. Therefore, the k nearest neighbor definition in the existing distance-based method is not reasonable, and results in a quadratic number of pair wise distance evaluations.

To solve the problems, we propose a VOD (Voronoi based Outlier Detection) method, to address the outlier detection by refining the concept of a neighborhood with the Voronoi diagram.

Definition 2-2. (Voronoi nearest neighbor). For a point p_i of set S, the nearest neighbors of p_i defined by the Voronoi polygon $V(p_i)$ are the Voronoi nearest neighbor of p_i, denoted as $V_{NN}(p_i)$.

In Figure 1, the Voronoi nearest neighbors of point p_1 are p_2, p_3, p_4, p_5 and p_6.

Definition 2-3. (Voronoi density). The Voronoi density of point p_i defined as

$$V_D(p_i) = 1 \Big/ \left(\sum_{o \in V_{NN}(p_i)} dist(p_i, o) \Big/ \left| V_{NN}(p_i) \right| \right) \tag{1}$$

where $|V_{NN}(p_i)|$ is the number of points in $V_{NN}(p_i)$.

Intuitively, the Voronoi density of point p_i is the inverse of the average distance based on the Voronoi nearest neighbors of p_i.

The VOD outlier detection algorithm based on the discussion is illustrated below.

Algorithm1.VOD outlier detection

Input. Data set S

Output. Outlier factor of the points in S, in descending order

1. Constructing Voronoi diagrams Vor(S) of data set S.

2. For each $p_i \in S$, compute Voronoi density $V_D(p_i)$

3. Sort the data by $V_D(p_i)$ in descending order

2.3 Complexity analysis

Given a set S of n points $p_1, p_2, \ldots, p_n$, computing the outlier factor of the data set in descending order involves the following steps:

The first step is to construct the Voronoi diagrams $Vor(S)$ of data set S. By theorem 2-4, the computational cost is $O(n\log n)$.

The second step is to compute the Voronoi density $V_D(p_i)$ for p_i, we need to find the Voronoi nearest neighbors of p_i and calculate the distance between them. By theorem 2-5, with the Voronoi diagram, a single nearest neighbor query can be performed in $O(\log n)$, the cost of all nearest neighbor query is $O(n\log n)$. By theorem 2-2, each edge of the Voronoi polygon $V(p_i)$ defines a nearest neighbor of p_i. Each edge shared between two polygons is an explicit representation of a neighborhood relation between two points. By theorem 2-3, a Voronoi diagram on n points has at most $3n$-6 edges. The times to calculate the distance between the points is at most $2(3n-6)$; the cost of computing the distance is $O(n)$. Thus, the cost of the second step is $O(n\log n)$.

Finally, we sort the data by $V_D(p_i)$ in descending order, for which the cost is $O(n\log n)$.

Thus, the final cost of the algorithm is $O(n\log n)$. Compared with the distance-based method, the time complexity is reduced from $O(n^2)$ to $O(n\log n)$.

3 Experimental Evaluation

In this section, we will perform an experimental evaluation to show that the proposed VOD outlier detection method can efficiently identify local outliers, and compare the performance of the proposed method with the distance-based method.

The experiments are executed on P4 2.0GHz CPU with 768Mb RAM running WIN XP. The algorithm is implemented by MATLAB 7.1.

3.1 Synthetic data

We start with a synthetic dataset generated based on multiple-Gaussian distribution, which contains 600 data points and has two clusters of non-uniform density. We add 10 outliers between the two clusters.

Figure 2 shows the outlier detection result from distance-based method. The 10 outliers are denoted by small circle, where 4 of them are not detected.

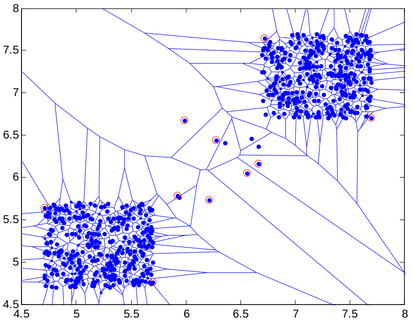


Figure 2 Outlier detection result from distance-based method in synthetic data

The result of VOD method showed in Figure 3 captures all the outliers correctly.

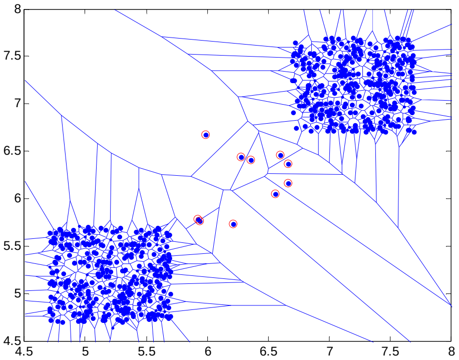


Figure 3 Outlier detection result from VOD in synthetic data

With the 600 data points, the running time of VOD is 2.976057 seconds while that of the distance-based method is 327.239739 seconds.

When the size of dataset vary from 200 to 1000, the running times of VOD and distance-based method are shown in Table 1, which shows that the VOD method is more efficient than distance-based method.

Table1 The running times of VOD and distance-based method

SIZE	VOD	DB
200	1.925527	40.091546
400	2.140383	128.129100
600	2.976057	327.239739
800	3.795783	492.061738
1000	4.428431	705.088370

3.2 Stock data

The second example uses the real-world dataset. Our experiment considered the outlier detection in Disney stock daily closing prices time series 0.

The time series contains 756 data points from 3/29/1996 to 3/29/1999. By piecewise-linear representation 0, the time series is transformed into 232 linear segments, which represented by points (length, slope), where slope is the volatility of the closing prices.

Figure 4 shows the outlier detection result from distance-based method. The 10 outliers are denoted by small circle, where 4 of them are not detected.

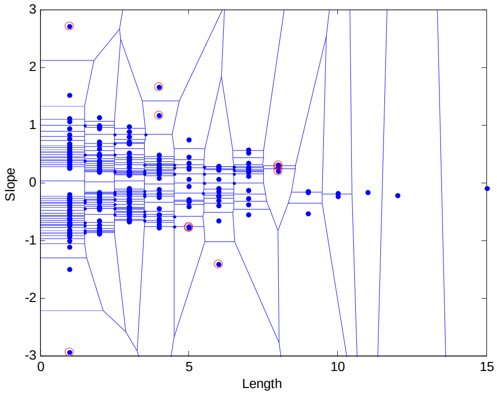


Figure 4 Outlier detection result from distance-based method in stock data

The result of VOD method showed in Figure 5 captures the outliers correctly.

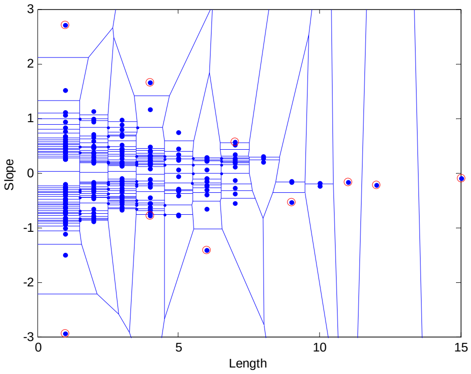


Figure 5 Outlier detection result from VOD in stock data

With the 232 data points, the running time of VOD is 1.161263 seconds while that of the distance-based method is 37.880471 seconds.

4 Conclusions

In this paper, we have proposed an efficient and effective VOD outlier detection method which uses the Voronoi nearest neighbor to calculate the outlier factor of a data point. With respect to the popular distance-based outlier detection, the proposed method performs better in identifying local outliers that deviate from the main patterns in a given dataset without parameter. The running time of our algorithm is $O(n\log n)$ where n is the size of dataset, which shows VOD method is more efficient.

In performing the VOD algorithms for high-dimensional dada set, to avoid the curse of dimensionality, we can construct an approximate Voronoi diagram of near linear size 0, which ensures the VOD method perfectly in the same way.

References

[1] D. Margineantu, S. Bay, P. Chan and T. Lane, "Data Mining Methods for Anomaly Detection KDD-2005 Workshop Report", ACM SIGKDD Explorations, Vol.7, No.2, Dec 2005, pp. 132-136

[2] E. M. Knorr, R.T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Datasets", In Proceedings of the 24th Conference on VLDB, New York: ACM Press, 1998, pp. 392-403

[3] S. D. Bay, M. Schwabacher, "Mining distance-based outliers in near linear time with randomization and a simple pruning rule", In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington: ACM Press, 2003, pp. 29-38

[4] A. Ghoting, S. Parthasarathy, and M. E. Otey, " Fast Mining of Distance-Based Outliers in High-Dimensional Datasets" , In Proceedings of the Sixth SIAM International Conference on Data Mining, Bethesda, MD, 2006, pp. 608-612

[5] Tao Yufei, Xiao Xiaokui, and Zhou Shuigeng, "Mining Distance-based Outliers from Large Databases in Any Metric Space", In Proceedings of the 12th ACM International Conference On Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 2006, pp. 394-403

[6] F. Angiulli, C. Pizzuti, "Outlier mining in large high-dimensional data sets", IEEE Transactions on Knowledge and Data Engineering, Vol.17, No.2, February 2005, pp.203-215

[7] F. P. Preparata, M. I. Shamos, Computational Geometry-An Introduction, New York: Springer-Verlay, 1985

[8] MathWorks Inc, Financial time series toolbox [EB/OL], http://www. mathworks. com/

[9] E. Fink, K. B. Pratt, "Indexing of Time Series by Major Minima and Maxima", In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Washington, DC, 2003, pp.2332-2335

[10] H-P. Sariel, "A Replacement for Voronoi Diagrams of Near Linear Size", In Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science, Las Vegas, Nevada, 2001, pp.94-103

An New Clustering Algorithm Based on QPSO and Simulated Annealing

Yong Wang   Wenbo Xu   Jun Sun

School of Information Technology, Jiang Nan University, Wuxi, 214122, China
Email:kfwy2006@163.com

Abstract

Particle Swarm Optimization (PSO) algorithm is a random population-based optimization technique which is simple and effective. Quantum-behaved Particle Swarm Optimization (QPSO)is a new algorithm model based on PSO. Simulated annealing is a computational intelligence algorithm which performances great and widely used in the solving nonlinear optimization problem. Clustering in data mining is similar with simulated annealing in essence, and simulated annealing can be used in data mining and clustering analysis. This article introduces the simulation annealing thought into QPSO algorithm which merges the hybrid and Gaussian variation, and proposes a new clustering algorithm based on QPSO and simulated annealing. This algorithm keeps the characteristic of QPSO which is simple and easy to achieve, and improves the ability of global searching and raises the convergence rate and stability. The experiment result indicates that this algorithm is better than other commen clustering algorithms.

Keywords: Cluster, Simulated, Annealing, QPSO, Hybridization, Variation

1 Introduction

Clustering analysis has been widely applied into data analysis, pattern recognition, image processing, etc.Particle Swarm Optimization (PSO) algorithm is a random population-based optimization algorithm. Inspired by the research of artificial life, Kennedy and Eberhartt proposed Particle Swarm Optimization (PSO) algorithm at 1995, and this algorithm has been widely used in function optimization, neural network training, pattern classification, fuzzy control system, as well as other areas of application. PSO is simple, easy to achieve, however, it is easy to fall into the local extreme point and has a slow convergence at later period. In order to overcome these shortcomings, there are lots of improved PSO algorithms, such as Quantum-behaved Particle Swarm Optimization algorithm. QPSO is a new algorithm model based on PSO which outperforms traditional PSO in search ability as well as having less parameter to control. Simulated annealing is a computational intelligence algorithm which performances great and widely used in the solving nonlinear optimization problem. Clustering in data mining is similar with simulated annealing in essence, and simulated annealing can be used in data mining and clustering analysis. This paper introduces simulated annealing into QPSO and proposes a new algorithm, which is simple and easy to achieve, and improves the ability of global searching and raises the convergence rate and stability. The experiment result indicates that this algorithm is better than other commen clustering algorithms.

2 Simulated Annealing

Simulated annealing (SA) is a Monte Carlo approach for minimizing multivariate functions.

SA is a numerical optimization technique based on the principles of thermodynamics. SA is motivated by an analogy to annealing in solids. The idea of SA comes from a paper published by Metropolis et al1. in 1953. The algorithm in this paper simulated the cooling of material in a heat bath. This is a process known as annealing.

If you heat a solid past melting point and then cool it , the structural properties of the solid depend on the rate of cooling. If the liquid is cooled slowly enough, large crystals will be formed. However, if the liquid is cooled quickly (quenched) the crystals will contain imperfections.

Metropolis's algorithm simulated the material a system of particles. The algorithm simulates the cooling process by gradually lowering the temperature of the system until it converges to a steady, frozen state.

It helps to visualize the problems presented by a system as a geographical terrain. For example, consider a mountain range, with two "parameters," e.g., along the North-South and East-West directions. We wish to find the lowest valley in this terrain. SA approaches this problem similar to using a bouncing ball that can bounce over mountains from valley to valley. We start at a high "temperature," where the temperature is an SA parameter that mimics the effect of a fast moving particle in a hot object like a hot molten metal, thereby permitting the ball to make very high bounces and being able to bounce over any mountain to access any valley, given enough bounces. As the temperature is made relatively colder, the ball can not bounce so high, and it also can settle to become trapped in relatively smaller ranges of valleys.

We imagine that our mountain range is aptly described by a "cost function." We define probability distributions of the two directional parameters, called generating distributions since they generate possible valleys or states we are to explore. We define another distribution, called the acceptance distribution, which depends on the difference of cost functions of the present generated valley we are to explore and the last saved lowest valley. The acceptance distribution decides probabilistically whether to stay in a new lower valley or to bounce out of it. All the generating and acceptance distributions depend on temperatures. The simulated annealing algorithm is as follows:

```
Initialization(Current_solution,Temperature)
Calculation of the Current_Cost
LOOP
    New_State
    Calculation of the new_Cost
  IF △(Current_cost*New_Cost)≤0 THEN
      Current_State=New_State
  ELSE
  IF Exp($\frac{Current\_cos t-New\_cos t}{Temperature}$) > Random(0,1)
  THEN
  -Accept
  Current_State=New_State
  ELSE
  -Reject
  Decrease the temperature
  EXIT When STOP_CRITERION
  END LOOP
```

3 Particle Swarm Optimization

Particle Swarm Optimization (PSO) algorithm is a population-based optimization technique originally introduced by Kennedy and Eberhart in 1995 (Kennedy and Eberhart, 1995). A PSO system simulates the knowledge evolvement of a social organism, in which individuals (particles) representing the candidate solutions to the problem at hand fly through a multidimensional search space to find out the optima or sub-optima. The particle evaluates its position to a goal (objective function) at every iteration, and particles in a local neighborhood share memories of their "best" positions, and then use those memories to adjust their own velocities, and thus subsequent positions. It has already shown that the PSO algorithm is comparable in performance with and may be considered as an alternative to the Genetic Algorithm (GA) (Angeline, 1998a; Eberhart and Shi, 1998).

In the original PSO with M individuals, each individual is treated as a volume-less particle in the D-dimensional space, with the position vector and velocity vector of particle i at k-th iteration represented as $X_i(k)=(X_{i1}(k),X_{i2}(k),\cdots,X_{iD}(k))$ and $V_i(k)=(V_{i1}(k),V_{i2}(k),\cdots,V_{iD}(k))$. The particle moves according to the following equations:

$$\begin{aligned}V_{ij}(k+1) = V_{ij}(k) + c_1 \cdot r_1 \cdot (P_{ij}(k) - X_{ij}(k)) + \\ c_2 \cdot r_2 \cdot (P_{gj}(k) - X_{ij}(k))\end{aligned} \quad (1)$$

$$X_{ij}(k+1) = X_{ij}(k) + V_{ij}(k+1) \quad (2)$$

for $i = 1, 2, \cdots M$; $j = 1, 2 \cdots, D$, where c_1 and c_2 are called acceleration coefficient. Vector P_i=(P_{i1},P_{i2}, ⋯P_{iD}) is the best previous position (the position giving the best objective function value) of particle i called personal best (pbest) position, and vector $P_g = (P_{g1}, P_{g2}, \cdots, P_{gD})$ is the position of the best particle among all the particles in the population and called global best (gbest) position. Take the following minimization problems into account,

$$\underset{X}{\text{Min}} f(X) \quad (3)$$

subject to

$$X \in \Omega \subseteq \Re^D$$

where $f(X)$ is an objective function and Ω is the feasible space. And thus the subscript g can be found by

$$g = \arg \min_{1 \le i \le M} (f(P_i(k))) \quad (4)$$

The parameters r_1 and r_2 are random numbers distributed uniformly in (0,1), that is $r_1, r_2 \sim U(0,1)$. Generally, the value of Vij is restricted in the interval $[-V_{\max}, V_{\max}]$.

4 Quantum-Behaved Particle Swarm Optimization

Quantum-behaved Particle Swarm Optimization (QPSO)was proposed[4][5] by Sun et al,it is a new algorithm model based on PSO. Quantum-behaved Particle Swarm Optimization algorithm that outperforms traditional PSOs in search ability as well as having less parameter to control.In QPSO algorithm.

Evolution formula of the particle is:

$$mbest = 1/M \sum_{i=1}^{M} P_i = (1/M \sum_{i=1}^{M} P_{i1,} \cdots, 1/M \sum_{i=1}^{M} P_{id}) \quad (5)$$

$$PP_{id} = \phi \times P_{id} + (1-\phi) \times P_{gd} \qquad \phi = rand \quad (6)$$

$$x_{id} = PP_{id} \pm \alpha \times |mbest_d - x_{id}| \times \ln(1/u) \quad u = rand \quad (7)$$

Where,mbest is the middle position of the particle swarm(pbest); PP_{id} is the random point between $P_{id}$and P_{gd}, *a* is the only parameter of the QPSO algorithm. Commonly let $\alpha = (1.0-0.5) \times (MAXITER - T) / MAXITER + 0.5$ Where *T* is the current number of iterations, *MAXITER* is the maximum number of iterations.

The QPSO algorithm is as follows:

(1) Set the iteration number T to zero .Initialize the swarm.

(2) Evaluate the performance $f(p_i^{(t)})$ *of each particle.*

(3) Evaluate new p_{id} *of each particle.*

(4) Evaluate new p_{gd} *.*

(5) luate mbest by Eq.(5).

(6) Evaluate the random point PP_{id} *of each particle by Eq.(6).*

(7) Move each particle to its new position , according to Eq.(7).

(8) Make T=T+1,go to step 2,and repeat until terminal condition satisfied.

5 An New Clustering Algorithm Based on QPSO and Simulated Annealing

The algorithm proposed in this paper takes the operating process of QPSO as the main process. In order to introduce the thought of simulated annealing into this algorithm, we use the hybrid operation in Hybrid-based Particle Swarm Optimization and the Gaussian variation operation in Variation-based Particle Swarm Optimization to further adjust and optimize the particle swarm. The basic operating process is that: first, initialize the swarm randomly. Second, start random search and evaluate a new swarm by Eq.(7). Third, carry on the hybrid operation and the Gaussian variation operation independently. Through the process of simulated annealing to every particle we get the result as the new swarm. In every evolution, the hybrid operation select some particles to put into a pond. The particles in the pond randomly hybridize in pairs and produce child particles with the same number. Then we replace the parents particles with the child particles so the number of the swarm is invariable. The positions of the child particles are counted by the arithmetic weighted sum of the positions of the parent particles:

$$child_1(x) = p * parent_1(x) + (1-p) * parent_2(x) \quad (8)$$

$$child_2(x) = p * parent_2(x) + (1-p) * parent_1(x) \qquad (9)$$

In the formulas, x is a D-dimensional position vector, and $child_k(x)$ and $parent_k(x)$, k=1,2 indicates the positions of the child and parent particles separately. P is a D-dimensional random vector which is averagely distributed, and each division is at [0,1].'*'means that the divisions of the vector multiply correspondingly.

At each evolution, the variation operation choose particles of appointed number to mutate with the Gaussian variation operator and replace the old particles with the ones mutated.

$$mutation(x) = x * (1 + Gaussian(\sigma)) \qquad (10)$$

The operating process of the algorithm includes two components: first, initialize a swarm with Eq.(7). Then use the hybrid operation and the Gaussian variation operation to further adjust these particles. The evolution process alternates repeatedly until some termination qualification is satisfied. The processes of the algorithm are as follows:

(1) initialize parameters: hybrid probability P_c, variation probability P_m, shrink-extend coefficient α, temperature cooling coefficient C, annealing initial temperature T;

(2) initialize the swarm;

(3) adjust the particles in the swarm by Eq.(7);

(4) choose particles from the swarm by P_c to form a sub-swarm and get a new swarm as follows:

select particles x_j, x_k in pairs randomly from the sub-swarm and hybridize the particles by Eq.(8) and (9) and get a new pair $x_j^{'}, x_k^{'}$.Count adaptation function value $f(x_j)$, $f(x_k)$, $f(x_j^{'})$, $f(x_k^{'})$.If $\min\{1, \exp(-(f(x_j^{'}) - f(x_j))/T)\} > random$,then replace x_j with $x_j^{'}$ .If $\min\{1, \exp(-(f(x_k^{'}) - f(x_k))/T)\} > random$,then replace x_k with $x_k^{'}$.Here random is a random number at [0,1];

(5) choose particles from the latest swarm by P_m to form a sub-swarm and get a new swarm as follows:

Select particle x_j randomly from the sub-swarm and mutate it by Eq.(10) and get $x_j^{'}$ .Count adaptation function value $f(x_j)$, $f(x_j^{'})$ .If $\min\{1, \exp(-(f(x_j^{'}) - f(x_j))/T)\} > random$, then replace x_j with $x_j^{'}$;

(6) Repeat until terminal condition satisfied and we get the global best value;

(7) If the evolution number is less than the predestinate biggest evolution number, then change the annealing temperature of the swarm, which means that $T \leftarrow CT$, and go to step(3).

6 Experimental Result

In order to verify the effectiveness of the method proposed in this paper, we make the emulation experiment in the environment of Celeron, 2.1G, 256M memory, WinXP, MATIAB7.0, using the iris, wine, breastercancer database and a group of remote sensing image data.

DataSet1: Iris sample is the most common database in the pattern recognition. The number of sample is 150, the number of characteristic is 4, the classification of clustering is 3.Every class is one kind of iris, which is linear with the other two, and the other two is non-linear.

DataSet2:Wine data is one group of MCI. The number of sample is 178, the number of characteristic is 13, the classification of clustering is 3.The sample number of class1,2,3 is 59,71 and 48.

DataSet3: breastcancer data is one group of MCI, too. The number of sample is 684, the number of characteristic is 9, the classification of clustering is 2.

DataSet4: The remote sensing image data uses the datapoints of Landsat-TM5, TM4, TM5 bands synthetic image. he number of sample is 360, the number of characteristic is 3,the classification of clustering are road, city, farmland, garden, woodland, territorial waters. The sample number of every class is 60.

In order to evaluate the algorithms in this paper, we propose the following three rules:

(1) the cluster adaptive function.

(2) the distance of the clusters, which means the distance between the centre vectors.

$$inter\_distance = \|z_i - z_j\|$$

(3) the distance in the cluster, which means the distance from all the feature vectors to the cluster center.

$$\operatorname{int} ra\_dis\tan ce = \frac{1}{|C_i|}\sum_{j\in C_i}\|x_j - z_i\|$$

To cluster, the smaller the object function(2) ,the better the result is. It means the result is accurate and the error is small. If the distance between the clusters is large, the similarity between the clusters is small. If the distance in the cluster is small, the inner-similarity in the cluster is large.

The number of particles is 20 and the simulating times and function evaluating times is 30 in the following experimental results.

Table1

Algorithm	Adaptition function	Inter_distance	Intra_distance
SA	60.5760	3.2945	184.9944
PSO+SA	60.5057	3.2963	190.7448
QPSO+SA	59.6809	3.2963	189.2801

Table2

Algorithm	Adaptition function	Inter_distance	Intra_distance
SA	0.0025	0.0028	0.1991
PSO+SA	0.0022	0.0075	0.1905
QPSO+SA	0.0022	0.0075	0.1866

Table3

Algorithm	Adaptition function	Inter_distance	Intra_distance
SA	0.0066	0.0003	0.3546
PSO+SA	0.0059	0.0077	0.4105
QPSO+SA	0.0059	0.0077	0.4229

Table4

Algorithm	Adaptition function	Inter_distance	Intra_distance
SA	302.86*10	172.1989	4628.00
PSO+SA	286.54*10	180.6204	5379.10
QPSO+SA	286.53*10	180.6903	4604.40

The results in the table4 indicate that the smaller the adaptation function is, the lager the inter_distance and the smaller the intra_distance becomes. The new clustering algorithm performances better in three conditions, and the run time is much shorter than the other algotithms.

7 Conclusion

Particle Swarm Optimization (PSO) algorithm is a simple random population-based optimization technique. Quantum-behaved Particle Swarm Optimization(QPSO) is a improved algorithm based on PSO. Simulated annealing is a computational intelligence algorithm which performances great and widely used in the solving nonlinear optimization problem. Clustering in data mining is similar with simulated annealing in essence, and simulated annealing can be used in data mining and clustering analysis. The algorithm introduced in this paper get a better clustering result than SA and PSO+SA. The prematurity doesn't happen and the convergence speed is faster.

References

[1] Kennedy J, Eberhart R, “Particle swarm optimization[C]”, In:IEEE Int’1 Conf on Neural Networks,Perth,Australia, 1995,pp.1942-1948

[2] Eberhart R,Kennedy J, ” A new optimizer using particle swarm theory[C]”, In:Proc of the sixth international symposium on MicroMachine and Human Science , Nagoya, Japan,1995,pp.39-43

[3] Shi Y, Eberhart R, ”C Fuzzy Adaptive particle swal'//l optimization[C], ”In:Proc of the Congresson Evolutionary Computation,Seoul Korea,200l,pp.67-74

[4] Sun, J. and Xu W.B.,”A Global Search Strategy of Quantum-behaved Particle Swarm Optimization[C]”, Proceedings of IEEE conference on Cybernetics and Intelligent Systems,2004,pp.111-116

[5] Sun, J. Feng B. and Xu W. B., ”Particle Swarm Optimization with Particles Having Quantum Behavior[C]”. Proceedings of 2004 Congress on Evolutionary Computation,2004,pp.325-331

[6] Natsuki Higasshi ,Hitoshi Iba, Particle swarm optimization with Gaussian mutation[C]”,In:Proc of the Congress on Evolutionary Computation,2003, pp.72～79

[7] Van den Bergh F,Engelbrecht A P,”A new locally convergent particle swarlTl optimizer[C]//2002”,IEEE International Conference on Systems,Man and Cybernetics, 2002, pp.101-110

[8] Van den Bergh F,”An analysis of particle swarlTl optimizers[D]”, University of Pretoria,Nov 2001,pp-77-88

[9] Kang.L.S. Non-numerical parallel algorithm--simulated annealing algorithm[M], BeiJing: Science Press, 1997

[10] Wang.X.M, Wang.Y.H, ”Integration of Simulated Annealing Algorithm and Genetic Algorithm”,Chinese Journal of Computers, 20(4),1997,pp.381-384

Solving Constrained Optimization Problems with Adaptive Quantum-Behaved Particle Swarm Optimization *

Yang Liu[1,2] Yan Ma[1] Baoxiang Cao[2] Deyun Yang[1]

1 Department of Information Science and Technology, Taishan University, Tai'an Shandong, 271021, China

2 College of Computer Science, Qufu Normal University, Rizhao Shandong, 276826, China
Email: 1 Willow-2001@hotmail.com

Abstract

In this paper we propose a new algorithm in solving constrained problem---Adaptive Quantum-behaved Particle Swarm Optimization (AQPSO).The AQPSO outperforms QPSO and PSO in global search ability and local search ability, because the adaptive method is more approximate to the learning process of social organism with high-level swarm intelligence and can make the population evolve persistently. We adopt a non-stationary multi-stage assignment penalty in solving constrained problem to improve the convergence and gain more accurate results. This approach is tested on several accredited benchmark functions and the experiment results show much advantage of AQPSO to QPSO and the traditional PSO.

Keywords: constrained; adaptive; quantum; PSO; multi-stage

1 Introduction

In general, a CO problem can be described as the following nonlinear programming problem:

$$\min_{x} f\left(x\right), x \in S \subset \Re^{n} \tag{1}$$

subject to the linear or nonlinear constraints:

$$\begin{aligned} & g_i\left(x\right) \le 0, && i = 1,2\ldots m \\ & \mathrm{h_j}\left(x\right) = 0, && j = 1,2\ldots k \\ & a\left(i\right) \le x_i \le b_i,\ 1 \le i \le n. \end{aligned} \tag{2}$$

where $f\left(x\right)$ is an objective function, $g_i\left(x\right)$ and $h_j\left(x\right)$ are equality and inequality constrained functions respectively, $a(i)$ and $b(i)$ are the search space up-bound and low-bound for x_i .The formulation of the constraints in (2) is not restrictive, since an inequality constraint of the form $g_i(x) \ge 0$ can also be represented as $-g_i(x) \ge 0$, and the equality constraint $g_i(x) = 0$ is equal to two inequality constraints $g_i\left(x\right) \ge 0$ and $g_i\left(x\right) \le 0$.

In this paper, a non-stationary function multi-stage assignment penalty function is adopted. The penalty values are dynamically modified according to equality constraints $g_i(x)$ and inequality $h_j(x)$ constraints. A penalty function is generally defined as:

$$F(x) = f(x) + h(k)H(x), \qquad x \in S \subset \Re^{n} \tag{3}$$

where $f\left(x\right)$ is the original objective function of the CO problem in Eq. (1); $h(k)$ is a dynamically modified penalty value, where k is the algorithm's current iteration number; and $H(x)$ is a penalty factor, defined as $H(x) = \sum_{i=1}^{m} \theta(q_i(x)) q_i(x)^{\gamma(q_i(x))}$ where $q_i(x) =$ max $\{0, g_i(x)\}$. $i = 1\ldots\ldots m$. The function $q_i(x)$ is a relative violated function of the constraints; $\theta(q_i(x))$ is a multi-stage assignment function; $\gamma(q_i(x))$ is the power of the penalty function; and $g_i(x)$ are the constraints described in Eq. (2).

The function $h(\cdot)$, $\theta(\cdot)$ and $\gamma(\cdot)$ are problem dependent. Details of the penalty function used in this study are given in Section 4.

* This work was supported by the National Natural Science Foundation of China under Grant No. 60572113 and Taishan University Scientific Research Projects under Grant No.Y07-2-12.

In this paper, we consider to solve the Constrained Optimization (CO) Problems by AQPSO algorithm. The CO problem is tackled through the minimization of a non-stationary multi-stage assignment penalty function. In the next section, the Particle Swarm Optimization (PSO) and the Quantum-Behaved Particle Swarm Optimization (QPSO) is briefly described. In section 3, the Adaptive Quantum-Behaved Particle Swarm Optimization (AQPSO) is reported .The test problems and the results of experiments are reported in Section 4. The paper ends with the conclusion and ideas for future research in Section 5.

2 PSO and QPSO

Particle Swarm Optimization (PSO), [1] originally proposed by Kennedy and Eberhart in 1995, is a population-based evolutionary computation technique, which differs from other evolution-motivated evolutionary computation in that it is motivated from the simulation of social behavior. In a PSO system, a particle corresponds to individual of the organism, which depicted by its position vector $\vec{x}$ and its velocity vector $\vec{v}$, is a candidate solution to the problem. That is the trajectory of the particle is determined. Then the optimal solution of the probability of moving out the trajectory is ignored. Therefore, in general, PSO can obtain good solutions in high-dimensional spaces but the ignorance of optimal solution does exist and PSO stumbles on local minima.

Keeping to the philosophy of PSO, we proposed a Delta potential well model of PSO in quantum world (QPSO) [4]. Because $\vec{x}$ and $\vec{v}$ of a particle are not determined simultaneously according to uncertainty principle, the term trajectory is meaningless in quantum world [5, 6].

2.1 Dynamics of classical PSO

In a classical PSO system proposed by Kennedy and Eberhart, the particles are manipulated according to the following equation:

$$v_i(t+1) = wv_i(t) + \varphi_1(p_i - x_i(t)) + \varphi_2(p_g - x_i(t)) \tag{4}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{5}$$
$$i = (1,2,\cdots,M)$$

Where x and v denotes the position and velocity of particle i among the population correspondingly, $\varphi_1{}^T$ and $\varphi_2{}^T$ are two random vectors in the range [0,1].

In Eq.(4), vector p_i is the best position (the position giving the best fitness value) of the particle i and vector p_g is the position of the best particle among all the particles in the population. Parameter w is the inertia weight [2], which does not appear in the original version of PSO [1]. In [3], M. Clerc and J. Kennedy analyze the trajectory and prove that, whichever model is employed in the PSO algorithm, each particle i in the PSO system converges to its local point p, whose coordinates are $p_d = (\varphi_{1d} p_{id} + \varphi_{2d} p_{gd}) / (\varphi_{1d} + \varphi_{2d})$ so that the best previous position of all particles will converge to an exclusive global position with $t \to \infty$.

2.2 Dynamics of quantum PSO

In Quantum-Behaved Particle Swarm Optimization (QPSO)[1, 4], the particles move according to the following equation:

$$\begin{aligned} mbest &= \frac{1}{M}\sum_{i=1}^{M} p_i \\ &= \left(\frac{1}{M}\sum_{i=1}^{M} p_{i1}, \frac{1}{M}\sum\nolimits_{i=1}^{M} p_{i2}, \ldots, \frac{1}{M}\sum\nolimits_{i=1}^{M} p_{id} \right) \end{aligned} \tag{6}$$

$$p_{id} = \varphi * p_{id} + (1-\varphi) * p_{gd}, \varphi = rand() \tag{7}$$

$$X_{id} = p_{id} \pm \alpha * \left| mbest_d - X_{id} \right| * \ln(\frac{1}{u}), u = rand() \tag{8}$$

where mbest is the mean best position among the particles. φ and u are a random number distributed uniformly on [0,1] respectively and α is the only parameter in QPSO algorithm.

When Quantum-behaved Particle Swarm Optimization algorithm is applied to practical problems, there are several control methods for parameter α .A simple one is that α is set to be a fixed value when the algorithm is running. But this approach is lack of robustness [7, 8].

Another efficient method is linear-decreasing method that decreasing the value of α linearly as the algorithm is running. That is, the value of α is determined by

$$\alpha = (\alpha_1 - \alpha_2) \times \frac{(MAXITER - t)}{MAXITER} + \alpha_2 \qquad (9)$$

The Quantum-behaved Particle Swarm Optimization (QPSO) Algorithm in is described as follows:

1) Initialize an array of particles with random position and velocities inside the problem space

2) Determine the mean best position among the particles by Eq. (6)

3) Evaluate the desired objective function (for example minimization) for each particle and compare with the particle's previous best values: If the current value is less than the previous best value, then set the best value to the current value. That is, if $f(x_i) < f(p_i)$ then $x_i = p_i$.

4) Determine the current global position minimum among the particle's best positions. That is:

$g = \arg\min_{1 \le i \le M} f(p(i))$ (M is the population size).

5) Compare the current global position to the previous global: if the current global position is less than the previous global position; then set the global position to the current global.

6) For each dimension of the particle, get a stochastic point between p_{id} and p_{gd}

$$p_{id} = \varphi * p_{id} + (1-\varphi) * p_{gd}, \varphi = rand()$$

7) Attain the new position by stochastic Eq. (6)

8) Repeat steps 2)-7) until a stop criterion is satisfied OR a pre-specified number of iterations are completed.

3 AQPSO

Like many other evolutionary algorithms, the major problem confronts Quantum Particle Swarm Optimization Algorithm is premature convergence, which results in great performance loss and sub-optimal solutions. With QPSOs the fast information flow between particles seems to be the reason for clustering of particles. Diversity declines rapidly, leaving the QPSO algorithm with great difficulties of escaping local optima. Another problem with QPSO in multi-modal optimization is computational cost. With the dimension of the optimization problem increasing, the population size must be enlarged to ensure the algorithm have a good performance, which, however, makes the algorithm computationally expensive.

To solving the aforementioned problems, we propose in this paper a Adaptive Quantum Particle Swarm Optimization (AQPSO) Algorithm.

In QPSO, Contraction-Expansion Coefficient is a vital parameter to the convergence of the individual particle in QPSO, and therefore exerts significant influence on convergence of the algorithm[2]. Mathematically, there are many forms of convergence of stochastic process, and different forms of convergence have different conditions that the parameter must satisfy. In this paper, we do not mean to analyze theoretically the convergence process of the individual particle in QPSO, but implement stochastic simulation to discover the knowledge about convergence of the particle.

For simplicity, we consider the evolution Eq. (8) of QDPSO [3] in one-dimensional space. The p_{id} is denoted as point *p*. In practice, when $t \to \infty$, the point *p* of the individual particle and the Mean Best Position point *mbest* will converge to the same point, and consequently, the particle in QDPSO and that in QPSO have the same convergence condition for parameter α except that they have different convergence rate.

From the results of stochastic simulation, we can conclude that when $\alpha \le 1.7$, the particle will converge to the point p, and when $\alpha \ge 1.8$, it will diverge. Therefore there must be such a threshold value $\alpha_0 \in [1.7, 1.8]$ that the particle converges if $\alpha \le \alpha_0$, and diverges otherwise. To get more precise value of α_0 we need to do simulation experiment with α set to be the value between 1.7 and 1.8 by more times.

However, for practical application of QPSO, the knowledge about parameter α we acquired so far is adequate [9].

The better parameter control method is to use adaptive mechanism. Firstly, we introduce the following error function

$$\Delta F = (F_{pbest} - F_{gbest}) / MIN(ABS(F_{pbest}), ABS(F_{gbest})) \quad (10)$$

where *Fi* is the fitness of the *i*th particle, *Fgbest* is the fitness of *gbest*, ABS(x) gests the absolute value of *x*, and *MIN*(x_1,x_2) gets the minimum value between x_1 and x_2.

Let $z = \log(\Delta F)$, then the function is

$$\alpha(z) = \begin{cases} 0.6 & z > 0 \\ 0.7 & -2 < z \le 0 \\ 0.6 + 0.1 \times k & -k-1 < z \le -k \ (k = 2,3,4) \\ 1.0 + 0.2 \times (k-4) & -k-1 < z \le -k \ (k = 5,6,7) \\ 1.8 & z \le -8 \end{cases} \quad (11)$$

The QPSO employing the above adaptive function is called Adaptive Quantum-behaved Particle Swarm Optimization (AQPSO).

4 Test Problems and Experimental Results

In our experiments, the population of the swarm was set equal to 100. We recorded mean best fitness values for 10 runs of each test problem, and the PAQPSO algorithm run for 1000 iterations in each case. A violation tolerance was used for the constraints. Thus, a constraint $g_i(x)$ assumed to be violated, only if $g_i(x) > 10^{-5}$. Regarding the penalty parameters, the same values as the values reported in [10] were used, to obtain results comparable to the results obtained using different EA. Specially, if $q_i(x) < 1$, then $\gamma(q_i(x)) = 1$,otherwise $\gamma(q_i(x)) = 2$ .Moreover, if $q_i(x) < 0.001$,then $\theta(q_i(x)) = 10$, else if $q_i(x) \le 0.1$ the $\theta(q_i(x)) = 20$, else if $q_i(x) \le 1$ then $\theta(q_i(x)) = 100$, otherwise $\theta(q_i(x)) = 300$. Regarding the function $h(\cdot)$, it was set to $h(k) = \sqrt{k}$ for Test Problem 1 and $h(k) = k\sqrt{k}$ for the rest problems.

The test problems are defined immediately below:

Table 1 Test functions

F	Mathematical Representation	Subjection	Best solution
f1	$f(x) = (x_1 - 2)^2 + (x_2 - 1)^2$	$x_1 = 2x_2 - 1, \quad x_1^2/4 + x_2^2 - 1 \le 0$	$f^* = 1.3934651$
f2	$f(x) = (x_1 - 10)^3 + (x_2 - 20)^3$	$100 - (x_1 - 5)^2 - (x_2 - 5)^2 \le 0, (x_1 - 6)^2 + (x_2 - 5)^2$ $-82.81 \le 0, 13 \le x_1 \le 100, 0 \le x_2 \le 100$	$f^* = -6961.81381$
f3	$f(x) = (x_1 - 10)^2 + 5(x_2 - 12)^2 + x_3^4$ $+3(x_4 - 11)^2 + 10x_5^6 + 7x_6^2 + x_7^4$ $-4x_6x_7 - 10x_6 - 8x_7$	$-127 + 2x_1^2 + 3x_2^4 + x_3 + 4x_4^2 + 5x_5 \le 0,$ $-282 + 7x_1 + 3x_2 + 10x_3^2 + x_4 - x_5 \le 0,$ $-196 + 23x_1 + x_2^2 + 6x_6^2 - 8x_7 \le 0, 4x_1^2 + x_2^2 - 3x_1x_2 + 2x_3^2 + 5x_6 - 11x_7 \le 0,$ $-10 \le x_i \le 10, i = 1.....7$	$f^* = 680.630057$
f4	$f(x) = 5.3578547x_3^2 + 0.8356891x_1x_5$ $+37.293239x_1 - 40792.141$	$0 \le 85.334407 + 0.0056858T_1 + T_2x_1x_4 - 0.0022053x_3x_5 \le 92,$ $90 \le 80.51249 + 0.0071317x_2x_5 + 0.0029955x_1x_2 + 0.0021813x_3^2 \le 110$ $20 \le 9.300961 + 0.0047026x_2x_5 + 0.0012547x_1x_3 + 0.0019085x_3x_4 \le 25,$ $78 \le x_1 \le 102, 33 \le x_2 \le 45, 27 \le x_i \le 45, i = 3,4,5,$ $where\ T_1 = x_2x_5 and\ T_2 = 0.0006262$	$f^* = -30665.538$
f5	$f(x) = 5.3578547x_3^2 + 0.8356891x_1x_5$ $+37.293239x_1 - 40792.141$	$0 \le 85.334407 + 0.0056858T_1 + T_2x_1x_4 - 0.0022053x_3x_5 \le 92,$ $90 \le 80.51249 + 0.0071317x_2x_5 + 0.0029955x_1x_2 + 0.0021813x_3^2 \le 110$ $20 \le 9.300961 + 0.0047026x_2x_5 + 0.0012547x_1x_3 + 0.0019085x_3x_4 \le 25,$ $78 \le x_1 \le 102, 33 \le x_2 \le 45, 27 \le x_i \le 45, i = 3,4,5,$ $T_1 = x_2x_{3,}\ T_2 = 0.00026$	unknown
f6	$f(x,y) = -10.5x_1 - 7.5x_2 - 3.5x_3$ $-2.5x_4 - 1.5x_5 - 10y - 0.5\sum_{i=1}^{5} x_i^2$	$6x_1 + 3x_2 + 3x_3 + 2x_4 + x_5 - 6.5 \le 0, 10x_1$ $+10x_3 + y \le 20, 0 \le x_i \le 1, i = 1,.....5, 0 \le y.$	$f^* = -213.0$

For each test problem, the mean and the best solution obtained in all 10 runs were recorded. The results for all test problems are reported in Table2.. The mean run-time for 10 runs of each functions in Table 3. In most cases AQPSO outperformed the results reported in [10] for other QPSO and PSO algorithm. Especially for test problem 4 the result can reach around its theoretical value. So the result of test problem 5 obtained from AQPSO algorithm may be even closer to its unknown theoretical value. Proper fine-tuning parameters of AQPSO may result in better solutions.

Table 2 Mean and the best solution for each test problem in 10 runs

F	Method	Mean	Best Solution
f1	PSO-In	1.394006	1.393431
	PSO-Co	1.393431	1.393431
	QPSO	1.39346498	1.39346498
	AQPSO	1.393438	1.39346382
f2	PSO-In	-6960.866	-6961.798
	PSO-Co	-6961.836	-6961.837
	QPSO	-6961.7274	-6961.80434
	AQPSO	-6961.7938	-6961.8120
f3	PSO-In	680.671	680.639
	PSO-Co	680.663	680.635
	QPSO	680.646034	680.635235
	AQPSO	680.637	680.634521
f4	PSO-In	-31526.304	-31543.484
	PSO-Co	-31528.289	-31542.578
	QPSO	-30665.535	-30665.5382
	AQPSO	-30665.5387	-30665.5381
f5	PSO-In	-31523.859	-31544.036
	PSO-Co	-31526.308	-31543.312
	QPSO	-31026.428	-31026.4277
	AQPSO	-31045.426	-31026.3652
f6	PSO-In	-213.0	-213.0
	PSO-Co	-213.0	-213.0
	QPSO	-213.0	-213.0
	AQPSO	-213.0	-213.0

5 Conclusion

According to the experimental results, in most cases the performance of the AQPSO method in coping with constrained optimization problems is better than QPSO and PSO algorithm [10], and than those obtained through other EAs.

Future work will focus on approaches of solving large-scale problems or apply it to problems with the run-time demands strictly.

References

[1] Others: J. Kennedy and R. Eberhart, “Particle Swarm Optimization”, Proc. IEEE Conf. On Neural Network, 1942-1948 (1995)

[2] Others: Y. Shi and R. Eberhart, “Empirical study of particle swarm optimization,” Proc. Congress on Evolutionary Computation, 1945-1950 (1999)

[3] Others: J. Sun and Wenbo Xu, Parameter “Selection of Quantum-behaved Particle Optimization”. ICNC 2005, LNCS 3612, pp. 543 – 552, 2005.© Springer-Verlag Berlin Heidelberg 2005

[4] Others: Jun Sun and Wenbo Xu, “Particle Swarm Optimization with Particles Having Quantum Behaviour” IEEE Congress. Evolutionary Computation (2004)

[5] Others: Angeline,P. J. Tracking exterma in dynamic environments. Proc. Evolutionary Programming V1. Indianapolis, IN.pp.335.345,1998

[6] Others: Back, T. On the behavior of evolutionary algorithms in dynamic environments.Proc. Int.Conf. on Evolutionary Computation.Piscataway, NJ:IEEE Press,pp.446-451,1998

[7] Others: D.W.Boeringer and D. H. Werner, “Particle Swarm Optimization Versus Genetic Algorithms for Phased Array Synthesis,” IEEE Trans. Antennas Propagat., in Press

[8] Others: L.Piegl and W. Tiller, The NURBS Book. Berlin: Springer, 1997

[9] Others: E. S. Peer, F. van den Bergh, A. P. Engelbrecht, Using Neighborhoods with the Guaranteed Convergence PSO,2003,IEEE,pp.235-242

[10] Others: K.E. Parsopoulos and M.N. Vrahatis, “Particle swarm optimization method for constrained optimization problems”, Intelligent Technologies-Theory and Application, 214-220 (2002)

Survey on the Multiple Sequence Alignment Based on Hidden Markov Model

Wenjuan Ji Jun Sun

School of Information Technology, Jiangnan University, Wuxi, 214122, China
Email: Jiwenjuan831209@yahoo.com.cn

Abstract

In computational biology, Multiple sequence alignment(MSA) is one of the basic problems. Realistic problem instances of MSA are computationally intractable for exact algorithms. One way to tackle MSA is to use Hidden Markov Models (HMMs), which are known to be very powerful in the related problem domain of speech recognition. The HMMs can be trained with different methods. And we can use a GA for optimizing the HMM structure, as well as hybrid GA-HMM training. In this paper, we will introduce several different methods to train the HMMs, and will show the advantages and disadvantages of each method, making comprehensive analysis to them. Analysis of the behavior of the algorithm sheds light on possible improvement.

Keywords: Hidden Markov Models, GA-HMM training, Hybrid Generic Algorithm

1 Introduction

Hidden Markov Models (HMMs) are a class of probabilistic models that are generally applicable to time series or linear sequences. HMMs have been most widely applied to recognizing words in digitized sequences of the acoustics of human speech[1]. HMMs were introduced into computational biology in the late 1980s, and have been a preferred choice of method when solving problems. Profile HMM is a particular type of HMM well suited to modeling multiple alignments. Profile HMM can be used to detect potential membership in a family by obtaining significant matches of a sequence to the profile HMM, to give an alignment of a sequence to the family or more precisely to add it into the multiple sequence alignment of the family, and to classify protein families. The standard algorithm for training HMM from initially unaligned example sequences are hill-climbing algorithms, such as gradient descent or Baum-Welch expectation maximization, which are iterative algorithms in which the likelihood (or the posterior probability) increases in each iteration. A serious problem with any hill-climbing optimization technique is that it often ends up in a local maximum. Genetic algorithm was also used to estimate HMM. In this paper, we present several methods to deal with the problem, such as a hybrid genetic algorithm for training profile HMM, the GA-HMM, etc. and producing multiple alignments to test the applicability of the introduced method.

2 Hidden Markov Models for MSA

In this section, we briefly describe the structure of the HMM and the involved algorithms used in this study. See Rabiner (1989) [2] and Krogh et al[3]. (1994) for a detailed introduction to HMMs, their associated algorithms and the computation of alignments. The HMM structure used in this study is the standard topology for the MSA problem originally suggested by Krogh et al. (1994). Figure 1 shows a simple topology example as a directed graph.

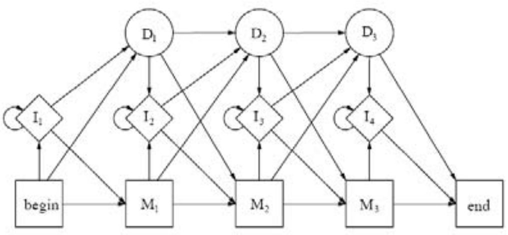


Figure 1 An example of a simple HMM of length 3 for MSA

The HMM model for MSA consists of a set of n states ($S1$,…… Sn) that are divided into three groups: match (M), insert (I) and delete (D). States are connected to each other by directed transitions that have an associated transition probability a_{ij}. The sum of the probabilities of all transitions going out of a state is 1. A match or insert state (Sj) emits an observable (a symbol) (v_k) from an output alphabet Σ with a probability $bj(k)$. The sum of all emission probabilities in each state is 1. Delete states do not emit observables and are called silent states. Starting in a special *begin* state and ending in a special *end* state, the HMM generates sequences (strings of observables) by making nondeterministic walks that randomly go from state to state according to the transitions. Each walk yields a path $\pi = (\pi 1, \pi 2, . . . , \pi p)$ of visited states and a sequence consisting of emitted observables along the path. When applying HMMs to MSA, the sequence of observables is given in the form of an unaligned sequence of amino acids. The goal is to find a path π that yields the best alignment. Given a sequence (o) and a HMM (λ) there are effective algorithms (e.g. the *forward* and *Viterbi* algorithms described in Rabiner, 1989) for determining the probability of o being generated by λ, i.e. $P(o|\lambda)$ and for deriving the path π with the maximal probability of generating o.

3 Methods and Algorithm

3.1 Training HMM structure with genetic algorithm

3.1.1 Generic Algorithm for HMM

To discover if GAs are potentially useful for evolving HMMs we implemented a standard GA where a population of HMMs are evolved from one generation to the next. At each generation some proportion of the HMMs are trained with Baum-Welch on a test set. The fitness of the HMMs are measured on a validation set and the fitter members are selected. Finally the members are mutated and crossed-over to form then next generation. This procedure is shown in Figure 2. The state labeled genetic operations include selection, mutation and crossover.

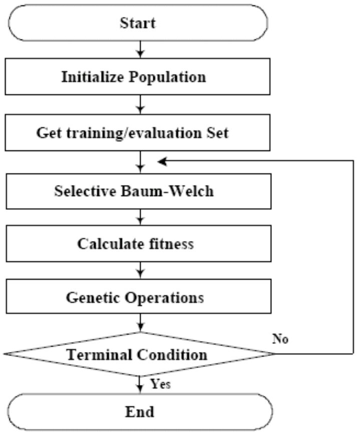


Figure 2 The GA-HMM algorithm. Baum-Welch training is combined with selection, mutation and crossover to evolve HMMs

In the experiments, the initial population consists of HMMs with just two states. The number of states will change due to state insertion and deletion mutations and through crossover. Also as part of the initialization stage the training data is divided into a set used for training with Baum-Welch and a set used for evaluating the fitness. The algorithm terminates when there is no significant change in the structural model.

3.1.2 Genetic Operations for GA-HMM

The genetic operations consist of selection, mutation and crossover in that order. Selection uses proportional selection with stochastic universal sampling (Baker, 1987) to reduce genetic drift. In both stochastic universal sampling and the more traditional roulette wheel selection, the sampling process can be visualized as assigning the pocket sizes of a roulette wheel to be proportional to the probability of selecting an individual. In roulette wheel selection P games are played independently to select P individuals. In stochastic universal sampling the roulette wheel is spun once and P individuals are selected at equally spaced intervals around the wheel. For a mutation to be useful it should make changes which cause minimal disruption so that the new HMM has a high probability of having a fitness close to that of the unmutated HMM. We considered mutations that only change either the number of states or the number of transitions by one. This gave us four mutation operators; insert state, delete state, insert transition and delete transition, which are shown in Figure 3. Insertion of a state can happen between any two states or at either end of the chain. When a state is inserted, the states on its right hand side shift by one as shown in Figure 3 (a). The emission probabilities of the new state are set to randomly selected values. If a state is deleted, all its transitions are removed. Insertion of a transition can happen between any two states and deletion of a transition happens at any state if the state has more than one outgoing transition. Although, these mutations allow highly interconnected HMMs they provide a certain bias towards chain structures because of the state insertion operator.

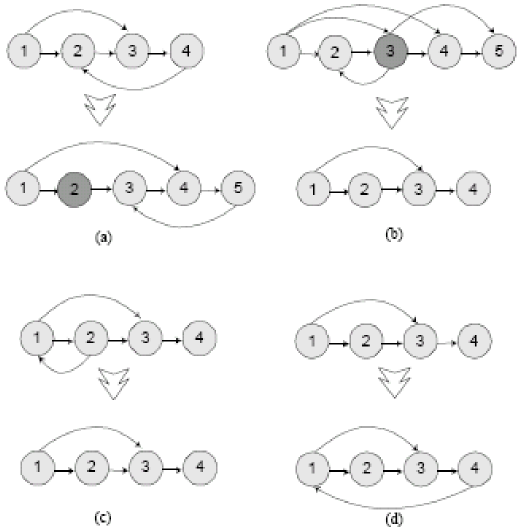


Figure 3 Four types of mutations

(a) insert state (inserting a state in the second position), (b) delete state (delete the third state), (c) delete transition, (d) insert transition.

Crossover takes place between two HMMs and exchanges states. A number of successive states can be crossed over in one operation. Only outgoing transitions from a state are exchanged during crossover. An example of crossover is shown in Figure 3.

3.1.3 About the GA-HMM method

This approach was used to get compact HMM architecture by merging states, it does not have the splitting state operation which is useful in dealing with new data. State splitting methods as well as deleting negligible states and transitions were used to find an optimal HMM topology (Y.Fujiwara et al., 1995). They used the transition ambiguity and the expected observation differences to split state and applied this to find a HMM structure of a leucine zipper motif. Those statistical approaches can be used to find a particular pattern like motif. However, those algorithms do not seem to find a structural model shown in the Simulation II. Our experiments were carried out on short sections of an HMM. It seems very unlikely in the current state of development that a GA would be able to find large HMM structures ab initio that are competitive with hand designed architectures. Nevertheless, even in the short term GAs may be able to ‘tune’ a hand designed HMM especially in areas where the biological significance of a region is poorly understood.

3.2 A novel optimization of profile HMM by a hybrid genetic algorithm

3.2.1 The Hybrid Genetic Algorithm

Since GA has a global search ability and heuristics have a local search ability, their hybridization will possibly form a more powerful search. In the hybrid GA, Baum-Welch algorithm is employed. Crossover is applied with probability Pc using one of the three mentioned crossover operators (the choice is made randomly with equal probability for all three crossover operators).

Mutation is applied with probability. P’m using one of the three mentioned mutation operators (the choice is made randomly with equal probability for all three mutation operators)[9]. P’m is decreased with current generation g increased using the formula: P’m = Pm(1.0 − g/g_{max}), where g_{max} is the predefined maximal generation. An offspring replaces the parent only if it is fitter. The offspring is refined using Baum-Welch algorithm, which re-estimated all the parameters (p(x|mk) and T(r|q)). To avoid over-fitting, Dirichlet mixture priors for regularization is used, for more details on Dirichlet mixture priors. The algorithm terminates either after a predefined number of generations g_{max} or after a predefined number of non-improved generations $g_{unimproved}$. All the steps can be summarized by the following pseudo-code:

```
Procedure Hybrid GA-HMM
BEGIN
  ①. Initialize population.
  Set models’ length to $M_{avg}$;
  $P(\bullet|\bullet)$ = random();
  $T(\bullet|\bullet)$ = random();
  ②. Evaluate.
  Calculate fmi , i=1...populationsize;
  WHILE ($g_{current} \leqslant g_{max}$ and $g_{currentunimproved} \leqslant g_{unimproved}$)
  BEGIN
    WHILE ($p_{currentsize} \leqslant$ populationsize)
    BEGIN
      ③. Tournament selection.
      ④. Crossover.
      IF( random() $\leqslant$ Pc)
        Select one of the three crossover operators with equal probability;
        Apply the selected crossover operator on parents;
      ⑤. Mutation.
      IF( random() $\leqslant$ Pm)
        Select one of the three mutation operators with equal probability;
        Apply the selected mutation operator on offsprings;
      ⑥. Apply Baum-Welch on offsprings.
      Calculate $P(\bullet|\bullet)$, $T(\bullet|\bullet)$;
      ⑦. Apply regularization on offsprings.
      ⑧. Evaluate.
    END
  END
END
```

3.2.2 Analyzing hybrid GA-HMM methods

The modeling was first tested on the globins, a large family of heme-containing proteins involved in the storage and transport of oxygen that have different oligo medic states and overall architecture. The globin protein sequences used for the training set were taken from the file globin50.fa of HMMER1.8.4, which contains 50 randomly selected unaligned globin sequences. We validated this model from the alignments it produced by the Viterbi algorithm, and compared to the performance of the hmmt program of HMMER1.8.4, hmmt can build an HMM from initially unaligned training sequences, and allows a choice of approaches, simulated annealing (SA), the Viterbi approximation of the Bauw-Welch (Viterbi), and the full Bauw-Welch (BW) implementations. The alignment accuracy was assessed by the sum-of-pairs score (SPS) and the column score (CS), SPS indicates the ratio of pairs correctly aligned while CS shows the ratio of columns correctly aligned, for how to calculate them see Thompson et al. The alignment of seven representative globins from Bashford et al. The BAliScore program was used to calculate SPS and CS score. In the test, we used Dirichlet mixture priors for regularization, a ten-component mixture Dirichlet prior for match emissions, and single component Dirichlet priors for insert emissions and transitions, the data was taken from the file BrownHaussler.pri of HMMER1.8.4. All the tests were performed on the same precondition.

We built a profile HMM from the alignment of seven representative globins from Bashford et al. Using the MAP construction algorithm, and encoded this model as an individual for the population initialization, the performance of hybrid GA is improved. For complicated HMMs, the parameter space may be complex, with many spurious local optima that can trap a training algorithm. Large collections of protein structural alignments are now available, so build the model from a multiple sequence alignment, and further refining it using one of these methods (BW, SA, GA etc) can greatly improve the accuracy of the model.

3.3 Training HMM by a particle swarm optimization-evolutionary algorithm hybrid [13]

3.3.1 The PSO–EA hybrid

We kept the length of the HMM constant during training and only optimized the parameters of the HMM, i.e. the transition and emission probabilities. We represented the candidate solution for a HMM as the position vector of a particle procedure train HMM begin initialize population P insert SA and BW solutions into population P

```
    while (not termination-condition) do
    begin
    copy population P  →P'
    normalize copy population P'
    evaluate P' regarding HMM training
    calculate new velocity vectors for P
    move P
    breed P
    end
    end
```

With real encoding of n transitions and m emission probabilities, which spanned a search space of n + m dimensions[20]. The structure of the algorithm for training HMMs with the PSO–EA hybrid is shown .First, a random initial population is created into which two seed solutions, one found by the BW algorithm and one found by a SA algorithm, are inserted. The initialization process is further described in the following section. During each iteration of the PSO, a copy of the population is created. All particles in this copy are normalized such that the constraints on the transition and emission probabilities mentioned are satisfied. Each particle in the copy population is then evaluated either according to the log-odds or the sum-of-pairs score as the objective function. If the log-odds score is chosen, the PSO–EA hybrid tries to maximize the probability that the HMM generates the given unaligned sequences of amino acids. For the sum-of-pairs score, the PSO–EA hybrid tries to maximize the quality of the alignment produced by the HMMs encoded in the particles. Afterwards, new velocity vectors for the original population (not the copy) are calculated. The particles are then moved and bred as described in the previous section.

The HMM training with the PSO–EA hybrid is based on the idea of prior seeding with SA and BW solutions, which means that any derived solution is at least as good as the SA and BW solutions. In this respect, the PSO–EA hybrid can be considered as a refinement method for BW and SA training. Even a few iterations of the PSO–EA hybrid can yield rapid further improvements. From a practitioner's point of view, this allows a user to stop the PSO–EA hybrid anytime when results are required, while a longer runtime has the prospect of further improvements if needed.

4 Discussion

We have described three algorithms, for alignment of protein sequences with repeated and shuffled domains. Analysis of the behavior of the algorithm sheds light on possible improvement[29].

From these new algorithms in recent years, we can see the alignment of protein sequences algorithms are no longer remaining in the early traditional algorithms. People pay more attention to new algorithms which are more suitable for the alignment of protein sequences. The main purpose to research the methods of alignment of protein sequences is to provide a tool that used to mine protein sequences for biological researchers. So, the flexibility and easy using of algorithm are the key issues to be considered.

References

[1] Kwong,S., Chau,C.W, “Analysis of parallel genetic algorithms on HMM based speech recognition system”, IEEE Transactions on Consumer Electronics, 43, 1997, pp.1229-1233

[2] Lipman,D.J., Altschul,S.F. and Kececioglu,J.D, “A tool for multiple sequence alignment”, Proc. Natl Acad. Sci. USA, 86,1989,pp.4412–4415

[3] Thompson,J.D., Higgins,D.G. and Gibson,T.J, “CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice”, Nucleic Acids Res., 22,1994, pp.4673–4680

[4] Thompson,J.D., Plewniak,F. and Poch,O, “A comprehensive comparison of multiple sequence alignment programs”, Nucleic Acids Res, 27,1999, pp.2682–2690

[5] Edgar,R.C, “MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res”, 32,2004, pp.1792–1797

[6] Katoh,K., Kuma,K., Toh,H. and Miyata,T. “MAFFT version

5: improvement in accuracy of multiple sequence alignment", Nucleic Acids Res., 33, 2005,pp.511–518

[7] Notredame,C., Higgins,D.G. and Heringa,J,"T-Coffee:a novel method for fast and accurate multiple sequence alignment", J. Mol. Biol., 302, 2000, pp.205–217

[8] Thomsen.R, "Evolving the Topology of Hidden Markov Models using Evolutionary Algorithms", Proceedings of the Seventh International Conference on Parallel Problem Solving from Nature - PPSN VII, 2002, pp. 861-870

[9] Lifang Liu, Hongwei Huo and Baoshu Wang, "A Novel Optimization of Profile HMM by a Hybrid Genetic Algorithm", Computational Intelligence and Bioinspired Systems, 3512 ,2005,pp.734-741

[10] Durbin,R.,Eddy,S.R.,Krogh,A.and Mitchison,G.J,(1998) "Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids", Cambridge University Press, Cambridge, UK

[11] O'Sullivan,O., Suhre,K., Abergel,C., Higgins,D.G. and Notredame,C, "3DCoffee: combining protein sequences and structures within multiple sequence alignments", J. Mol. Biol., 340,2004, pp.385–395

[12] Pei,J. and Grishin,N.V, "PROMALS: towards accurate multiple sequence alignments of distantly related sequences", Bioinformatics doi: 10.1093/bioinformatics/ btm017

[13] Zhou,H. and Zhou,Y, "SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures", Bioinformatics, 21, 2005, pp.3615–3621

[14] Pei,J., Sadreyev,R. and Grishin,N.V, "PCMA: fast and accurate multiple sequence alignment based on profile consistency", Bioinformatics, 19, 2003,pp.427–428

[15] Henikoff,S. and Henikoff,J.G, "Amino acid substitution matrices from protein blocks", Proc. Natl Acad. Sci. USA, 89,1992, pp.10915–10919

[16] Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res, 25,1997, pp.3389–3402

[17] Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker, W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H. et al, "The Universal Protein Resource (UniProt): an expanding universe of protein information", Nucleic Acids Res., 34,2006, D187–D191

[18] Ari L¨oytynoja and Michel C. Milinkovitch, "A hidden Markov model for progressive multiple alignment", Bioinformatics,19, 2003,pp. 1505–1513

[19] Baldi,P.,Chauvin,Y.,Hunkapiler,T., and McClure,M.A, "Hidden Markov modes of biological primary sequence information", Proc. Natl. Acad. Sci. USA, 91, 1994, pp.1059-1063

[20] Eddy,S, "Multiple alignment using hidden Markov models", Proc. Int. Conf. on Intelligent Systems for Molecular Biology, Cambridge, England: AAAI/MIT Press , 1995, pp.114-120

[21] Brown,M.P, Hughey,R., Krogh,A., Mian,I.S.,Sj olander, K., and Haussler,D, "Using Dirichlet mixture priors to derive hidden Markov models for protein families", Proc. of First Int. Conf. on Intelligent Systems for Molecular Biology, Menlo Park, CA:AAAI/MIT Press , 1993, pp.47-55

[22] Pei,J. and Grishin,N.V, "AL2CO: calculation of positional conservation in a protein sequence alignment", Bioinformatics, 17,2001, pp.700–712

[23] Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S, Khanna,A., Marshall,M., Moxon,S. et al, (2004) The Pfam protein families database. Nucleic Acids Res, 32, D138–D141.

[24] Thompson,J.D., Plewniak,F. and Poch,O, "A comprehensive comparison of multiple sequence alignment programs", Nucleic Acids Research, 27(13), 1999, pp. 2682-2690

[25] Bashford,D.,Chothia,C. and Lesk,A.M, "Determinants of a protein fold:unique features of the globin amino axid sequence", Journal of Molecular Biology,196,1987, pp.199-216

[26] Thompson,J.D.,Plewniak,F. and Poch,O, "BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs", Bioinformatics,15,1999,pp. 87-88

[27] Kyoung-Jae Won1 , Adam Prügel-Bennett and Anders Krogh, "Training HMM Structure with Genetic Algorithm for Biological Sequence Analysis", K.J. Won et al,2002,pp.1-7

[28] Thomas Kie,Rasmussen, Thiemo Krink, "Improved Hidden Markov Model training for multiple sequence alignment by a particle swarm optimization—evolutionary algorithm

hybrid", BioSystems ,72 ,2003, pp. 5–17

[29] Rabiner,L.R, "A turorial on hidden Markov models and selected applications in speech recognition", Proc. IEEE, 77,1989,pp.257-286

[30] Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D, "Hidden Markov models in computational biology: Applications to protein modeling", Journal of Molecular Biology, 235,1994, pp.1501-1531

[31] Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou, S, "ProbCons: Probabilistic consistency-based multiple sequence alignment", Genome Res., 15,2005, pp. 330–340

[32] Pei,J. and Grishin,N.V, "MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information". Nucleic Acids Res., 34, 2006,pp.4364–4374

[33] Simossis,V.A. and Heringa,J, "PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information",2005, Nucleic Acids Res., 33, W289–W294

[34] Thompson,J.D., Plewniak,F., Thierry,J. and Poch,O, "DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches", Nucleic Acids Res., 28,2000, pp.2919–2926

[35] Jones,D.T. "Protein secondary structure prediction based on position-specific scoring matrices", J. Mol. Biol., 292,1999, pp.195–202

[36] Van Walle,I., Lasters,I. and Wyns,L, "SABmark – a benchmark for sequence alignment that covers the entire known fold space", Bioinformatics, 21,2005,pp.1267–1268

[37] Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E, (2004)" The ASTRAL Compendium in 2004. Nucleic Acids Res., 32, D189–D192

[38] Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C, " SCOP: a structural classification of proteins database for the investigation of sequences and structures", J. Mol. Biol., 247,1995,pp.536–540

Application of Bayesian Network Learning Methods in E-learning *

Qing Yang[1,2] Lianfa Zhang[2] Zhufeng Huang[2] Xueping Wang[2]

1 Center for Language& Language Education, Central China Normal University, Wuhan, Hubei 430079, P.R.China
Email: yangq@mail.ccnu.edu.cn

2 Department of Computer Science, Central China Normal University, Wuhan, Hubei 430079, P.R.China

Abstract

E-learning is becoming one of the most important educational means. As more and more organizations and institutions are moving towards the e-learning strategy, self-learning model becomes a big challenge. Background knowledge and learning objectives of various groups of students on the network are very different. Self-learning system, which uses different learning programs for different students, can enhance the efficiency of learning process. In a self-learning system, the algorithm dealing with uncertainty factors of Self-learning model is very important. Bayesian network artifice is a very effective one within various methods dealing with uncertainty. In this paper, we applied Bayesian network method to self-learning model; designed Bayesian network structure in a self-learning model; assigned the local probability distribution and discussed the way to acquire and propagate related evidences. The practice has proven Bayesian network approach for self-learning model is a very effective method.

Keywords: Bayesian Network, Learning Methods, self-learning, e-learning, DAG

1 Introduction

As more and more educational organizations are moving into e-learning, some organizations are pushing a uniform standard to facilitations are pushing a uniform standard to facilitate e-learning implementation. There are four main e-learning standard organizations: AICC, IEEE Learning Technology standards Committee, IMS Globe Consortium and ADL [1]. E-learning is aimed at maximizing the efficiency of teaching and learning by means of a self-learning method which enables individual learners to choose the direction of their learning befitting their level and ability.[8][9][10]

Self-learning system must acquire and continually updated the recognition of students' mastery of knowledge and learning ability. The system has to record the learning information for each student: the individual learning environment, his dynamic learning situation and even detail information such as if the student respond to the teacher in time. The system quantizes this information and provides appropriate teaching content based on each self-learning character. This approach is implemented through Self-learning model.

Self-learning model is the core of the adaptive distance-education and also the most difficulty part to implement [5]. Not only does it need to acquire the learning condition of the student but also it needs to have certain reasoning ability. Bayesian network has a powerful ability for reasoning and semantic representation, which combined with qualitative analysis and quantitative analysis, with prior knowledge and observed data, and provides an effective way to deal with prediction, classification and clustering [4]. Bayesian networks are graphical representations of a multivariate joint probability distribution that exploits the dependency structure of distributions [2]. Bayesian

* Supported by State 211 Project and National Society Science foundation of China (No. 07BYY033)

networks are Directed Acyclic Graphs (DAG), where the nodes are random variables, and the arcs specify the independency assumptions that must be held between the random variables [3]. It has the ability to express cause-effect relationship and is widely used for uncertain knowledge expression and reasoning. This paper focused on application of Bayesian network learning methods to Self-learning model. Based on expert knowledge, prior information and the given dataset, an evaluation model was constructed. All of the work would be helpful for Self-learning model.

This paper is organized as follows. Section 1 presents an overview of Self-learning model and its characteristics. Section 2 discusses Bayesian network and how to construct a Bayesian network model from a given dataset. Section 3 constructs a Bayesian network model for Self-learning model. Finally, some conclusions and suggestions are presented in Section 4.

2 Construction Bayesian Network

2.1 Bayesian network

Bayesian Network is a Directed Acyclic Graph with probability independency assumptions. A Bayesian network is denoted by B=(S, P), where S is a DAG. The set of nodes represents a set of random variables X={X_1,X_2,…,X_n}, and the P= p(x_i|Pa_i) (i=1,2…,n) is the conditional probability of node i. Where X_i is the node, Pa_i is the parent of the node i, so the joint probability distribution over X which admits the following joint probability distribution decomposition [2]:

$$p(x) = n\prod_{i=1}^{n} p(x_i \mid Pa_i) \qquad (1)$$

The probability distribution is subjective if it is built only according to the historical experience while it is objective if it is learned from the data.

2.2 Construction bayesian network

Learning a Bayesian network from data involves two tasks: Estimating the probabilities for the conditional probability tables (learning parameters) and deriving the structure of the network. There are two main approaches to structural learning [6]: Bayesian and constraint-based. In the Bayesian approach, the user first constructs a BN with which she encodes her knowledge of the subject and her confidence in this network. This prior network is then combined with data to find the most likely model structure. This can be computationally very demanding. The constraint-based algorithms search for conditional dependences between each pair of variables, and build the model structure based on them. They are computationally easier, and therefore more common. Constraint-based learning requires no prior knowledge or input from the user.

In this section, we propose an approach to Bayesian network structure based on information theoretical method and Bayesian method. It firstly constructs the undirected graph by analyzing dependency relationships among nodes (respond to variables). The dependency relationships are measured by conditional mutual information. Then it defines the direction and evaluates the model by Bayesian scoring method. By this approach, an undirected graph is obtained through conditional independence (CI) tests, and the most probable directed structure is selected that maximizes the aposteriori probability of the model given the data. Consequently, it reduces the search space of possible structures and improves the efficiency of learning.

Given dataset D, the objective of structure learning is to identify the best network structure G that best matches D. For search-and-scoring approaches, how goodness the structure matches data set is measured by adopted scoring metric.

Because of the decomposition characteristic of Bayesian network shown in Eq. (1), commonly used scoring metric such as Bayesian score, BIC score and MDL score could be decomposed into summation of sub-scores of each variable given the states of their parents[7],

$$Score(G) = \sum_{x_i \in X} Score(X_i \mid \Pi_{x_i}) \qquad (2)$$

If a score metric follows.Eq.(2), it is “decomposable”. For any decomposable score metrics, it is easy to see that, the best network structure G_{best} could be obtained by searching the best combination of parents Π_{x_i} for every variable X_i in X.

In implementation, the network structure G is usually represented as an n×n adjacent matrix, where n is the number of nodes in G. In G, if node i is a parent of node j, then the element in the matrix $g_{i,j}=1$, otherwise ,$g_{i,j}=0$. This paper use G to indicate the adjacent matrix which represents the network structure. It can be easily identified in context, whether G represents a network structure or a matrix.

Let vector $g_i=[g_{1,i},g_{2,i},\ldots,g_{n,i}]$, we call g_i the parents vector of node i. g_i represents the edges from other nodes to node i in the network structure. From g_i we could easily know parents ΠX_i for variable Xi. Therefore G could be denoted by $G=[g_1,g_2,\ldots,g_n]$, and Eq. (2) could be rewritten,

$$\text{Score(G)} = \sum_{i=1}^{n} Score(g_i) \tag{3}$$

Therefore the problem of structure learning could be converted to an optimization problem.The objective function is illustrated,

$$\max(\text{Score(G)}) = \sum_{i=1}^{n} \max(Score(g_i)) \tag{4}$$

Searching the best structure G could be decomposed into n sub-processes. In each sub-process, the best parents vector g_i is searched. The process of searching should be subjected to the constraints that the resulted network structure G can not contain any cycles, i.e.G must be a DAG.

3 Construction Bayesian Network for the Self-Learning Model

Collecting and structuring expert knowledge While Bayesian models are a useful way to the model expert knowledge, it may prove difficult to get the knowledge out of the experts in a form that can be converted into probability distributions.

There are two main reasons for this. Firstly, many ecology researchers are used to working with real sampling or experimental data, and may find it exceedingly difficult to provide any numbers without relying on data. Secondly, they may be used to classical statistical analyses and feel uncertain when trying to think about their knowledge in terms of distributions rather than point estimates and confidence intervals.

This uncertainty together with only superficial knowledge about the methodology may also lead to distrust towards the BNs, which easily leads to reluctance to provide the estimates. The task of estimating probabilities, especially those of rare events, is a difficult one, and people naturally rely on a set of heuristic procedures that often do serve them well but may also result in biased outcomes. Studies of estimation processes have also revealed that regardless of elicitation technique, human estimators are prone to over confidence, that is, giving estimates that are too near to zero or one. On the other hand, experts' judgments tend to be rather under- than overconfident.

This section is to build up the network topology between the self-learning model and teaching resource knowledge item. Every node in Bayesian Network has to be given conditional probability, which can be obtained by samples learning or evaluating by the expert. A node without parent should be given prior probability. If there aren't enough samples, it has to evaluate the conditional probability in Bayesian network. Conditional probability table in Bayesian network is built to be component parts for evaluation, and the date in which is refined and experienced mostly.

(1) Design and analysis

The data set contains 500 cases. The domain problem has x variables; each of them has several attributes. And one variable responds to one node in the model respectively. The variables and their implications describe as follows. For instance, some knowledge key is listed in database courses ①field. ②record. ③table. ④database and so on.

Figure 1 is the local teaching resource network topology, each node has four discrete states as above: A, B, C, and D. when design the network, we have to assign the conditional probability table for each node. In Figure 1 we need to obtain the probability as following: P(data type), P(variable), P(Definition of

One-dimensional array), P(definition of Two-dimensional array).

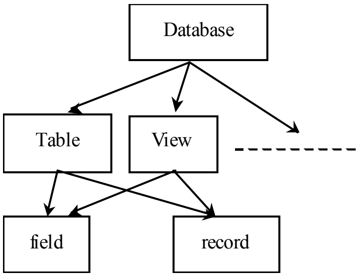


Figure 1 Construction of local teaching resource

(2) Sample collection

By analyzing the result of the test of the chapter or unit, we can obtain the influence degree between the knowledge items; and then we can determine the conditional probability function between the items in the self-learning model using Bayesian Network. We classify the students' result into four groups as following: 100-85, 84-70, 69-60, and blow 60, each of which is respond to the four state of the node.

(3) Learning conditional probability table

The brief thought of the learning conditional probability algorithm is as following: First, according to the network topology sequence, we get a random number n by the random number generator. Second, we take n papers at random in each grade group to assign values to the state of the network nodes and then get the one sample of the network when the value of the nodes is assigned. Repeat the above steps we get S_n samples, that is, a stochastic sample serials. Third, we get the approximate inference results of network. When the amount of sample serials is enough, margin and condition statistic quantity are close to margin and condition probability of node respectively.

The steps of the learning conditional probability algorithm:

a. Divide the continuously variable into some subsections and in each of them get n discrete variable;

b. Change the graphic network to numerical expression;

c. Find out the topology sequence of the network node and sequential sample the node according to it;

d. Get the sampling result according to the stochastic discrete variable;

e. Repeat step d until all node is sampled, then get a sample group;

f. Repeat step d and step e until all the node is sampled m times, so get m sample groups, called sample serial S, where m is the number of sample groups, that is, the sampling times of the network.

The node' s complete probability is obtained by making the statistic of the samples serials by column singly, while the joint probability and conditional probability between the nodes are obtained by making the joint statistic of the samples serials by multi-column.

① Joint statistic of the sample series

Form table 1, the sample series are obtained as above. The next work is to make statistic of the samples, and get the approximate inference results of network.

Table 1 Sample series S

node time	Field N1	Record N2	Table N3	Database N4
1	1	0	1	0
……	……	……	……	……
m	0	0	0	0

Note: where 1 is the correct answer, 0 is the wrong answer.

② Complete probability of the nodes

The complete probability of node Ni (i =1,……, n) is the result of make statistic of the correspond column singly. For example, the sum of "1" in the column "field" of Table 1 divided by sampling times m is approximate to the probability that students mastering the "field" is Know: P (field=Know), that is:

$$P(field = Known) = \frac{\sum_{i=k}^{m} a_i}{n} \qquad (5)$$

$$P(field = UnKnown) = 1 - P(field = Know) \qquad (6)$$

where a_i is the number of right answer, n is the total of the papers.

③ Conditional probability and joint probability of the nodes

The condition probability of the node Ni (i =1,……, n) is obtain by making joint statistic of the correspond column and the linked columns. For example, the way to calculate the probability of the "Database" while the node "Table" is Known is as following:

First, select the samples in column "Table" in the Table 1 and get new sample series S0, the number of group is n0. Second, calculate the sum of the samples in column "Database" in the Table 1, and then divided by n0, the result can be approximate considered as the probability of the "Database" while the node "Table" is Known. Similarly, the way to calculate the joint probability of some nodes is: obtain the sample groups by selecting the samples that the node states are satisfying certain conditions, and divided by the all sample groups m.

Bayesian Network built according to the user's prior experience is called prior Bayesian Network, the Bayesian Network joining prior Bayesian Network and data is called posterior Bayesian Network. The process from prior Bayesian Network to posterior Bayesian Network is called Bayesian Network learning, which correct the prior experience using data. The learning process is durative and the posterior Bayesian Network learned can be considered as prior Bayesian Network for next time.

Given the probability of the item X, that is,

P(X= Known), and the parent node of X is P_a , the probability of the P_a is:

$$P(p_a(X)\,|\,X) = \frac{P(X\,|\,p_a(X)) * P(p_a(X))}{P(X)} \tag{7}$$

Based on the Bayesian network model for the adaptive teaching system, and learning the parameters of each node with the dataset by using Eq.(7) and Bayes criterion, the complete network including conditional probability distributions was got. We selected 500 cases to test the validity of the model. Table 2 shows the experiment result, the evaluation grade of 455 cases is the same their actual grade, so the evaluation accuracy is 85.1%. The experimental results validate the practical viability of the proposed approach for the Self-learning model.

Table 2 The practical data of Bayesian network for "database"

Evaluation Grades	Actual Grades		
	Known	UnKnown	Total
Known	387	4	391
UnKnown	68	41	109
Total	455	45	500
Evaluation accuracy %	85.1%	91.1%	

4 Conclusions

Self-learning system is the system which uses different learning programs for different students. It can enhance the efficiency of learning process. In a self-learning system, the algorithm dealing with uncertainty factors of Self-learning model is very important. We use Bayesian network Learning Methods to Self-learning Model. In order to determine whether the student master the detail knowledge item in the adaptive teaching system, there are two ways to obtain proof: (1) the student answer to the questions of the system; (2) test the student after learning the correspond chapter or unit, and then update the Bayesian Network. The student provide feedback to the system by: A mastering the knowledge item, B not mastering the knowledge item, C not determining whether master the knowledge item (which need to have a test) after the student learned the knowledge items on the web page. A and B mean that the network has obtained the proof while C needs to have a test.

Adaptive teaching system improves the learning efficiently according to the characteristics of the students. However, it is complex to describe the cause-effect relationship of the knowledge items in the course correctly, and meanwhile it has directed influence on the construction of the Bayesian Network. In addition, the high computation in the Bayesian Network lowers the efficiency of the system. For the problems mentioned above exist, further research work should by done.

References

[1] Jianming Yong(2007),”Security modeling for e-Learning”, Proceedings of the 2007 1st International Symposium on Information Technologies and Applications In Education, IEEE PRESS,pp.1-5

[2] Zhongzhi Shi(2006),”The Advanced Artificial Intelligence”, the Science Press,pp. 184-185

[3] Cristina C, Abigail G, Kurt V(2002),”Using Bayesian Networks to Manage Uncertainty in Student Modeling”, User Modeling and User-Adapted Interaction, 12,pp. 371-417

[4] Li De-Yi, Liu Chang-Yu, Du-He, Han-Xu,”Artificial Intelligence with Uncertainty”, Journal of Software 2004, 15(11),pp.:1583-1594

[5] Skaanning C, “Jensen F V(2000). Using Bayesian Networks Industrial and Engineering Application of Artificial Intelligence and Expert Systems”, New Orleans, USA.M. King, B. Zhu, and S. Tang, “Optimal path planning,” Mobile Robots, vol. 8, no. 2, March 2001, pp. 520-531

[6] Laura ,” Advantages and challenges of Bayesian networks in environmental modelling”, ecological modelling 203(2007), pp. 312–318

[7] DU Tao, ZHANG Shen-sheng, WANG Zong-jiang, “Learning Bayesian Networks from Data by Particle Swarm Optimization”, Journal of Shanghai Jiaotong University(Science),Vol.E-11,No.4,2006,pp. 423-429

[8] Se bastien George, Herve Labas , “E-learning standards as a basis for contextual forums design”, Computers in Human Behavior 24 (2008) ,PP.138 – 152

[9] Javier Andrade, Juan Ares, Rafael Garca, Santiago Rodrguez, Mara Seoane, Sonia Suarez, “Guidelines for the development of e-learning systems”, Computers & Education (2008) , PP.1-13 , www.elsevier.com/ locate/ compedu

[10] Anatoly Gladun, Julia Rogushina, Francisco Garcıa-Sanchez, Rodrigo Martnez-Bejar, Jesualdo Tomas Fernandez-Breis, “An application of intelligent techniques and semantic web technologies in e-learning environments”, Expert Systems with Applications, www.elsevier.com/locate/eswa

An Optimized Genetic Algorithm for TSP

Dongling Bai[1] Qingping Guo[2]

1 Department of Computer Science and Technology, Wuhan University of Technology Wuhan, Hubei 430063.P.R.China

Email: 1 bdling123@163.com; 2 qpguo@mail.whut.edu.cn

Abstract

It is difficult to find out a precise answer to TSP (Traveling Salesman Problem). But GA (Genetic Algorithm) can find a better answer to it. This paper introduce GA and its essence. Then TSP, which is based on GA, is presented. In this paper main operators in GA, such as crossover operators and mutation operators, are analyzed and compared. By employing heuristic crossover and inversion mutation, a new method based on genetic algorithm for solving TSP is presented. The experimental results simulated on TSP show that this algorithm is feasible and effective to solve TSP. Employing heuristic crossover and inversion mutation can prevent premature convergence and ensure that the population is diverse, like what happens in the nature.

Keywords：Genetic Algorithm, Selection, Crossover, Mutation, TSP

1 Introduction

Genetic Algorithm (GA) is first proposed by Holland, a professor of Michigan University in American, in the 1970s and developed it. GA is based on Darwin' organic evolution theory, which is called "survival of the fitness", and Mendel's genetic theory [1]. It is very difficult for traditional methods to deal with complex and nonlinear problems; however, GA is very good at them. After more than 20 years of development, now GA has already successfully applied to many fields such as combinatorial optimization and artificial intelligence (AI). GA consists of four parts: coding mechanism, controlling parameters, fitness function, and genetic operator. There are a lot of methods to encode, for example binary encoding, coded-decimal notation. There are also many parameters in GA, among which the size of the population, the times of the interaction (generation), Pc (crossover probability), Pm (mutation probability) are more important. Fitness function reflects individual adapts to the nature or not. Selection, crossover and mutation are the three important genetic operators.

2 The Theory of Genetic Algorithm

There are two types of explanations for GA: First, traditional schemata theorem; Second, finite-state Markov chain model, which develops after 1990s.

(1) Schemata theorem, which is created by Holland, consists of schemata theorem, inner parallelism and building block hypothesis. The so-called schemata theorem is a set of symbols used string {0, 1, *}, among which * can be either 0 or 1. For example: H = *1* 00** 1 is a schemata. The total number of 0 or 1 in the model is called exponent, symbolizing o (H). The distance between the first figure and the final figure in the schemata is seen as schemata length, symbolizing δ (H). So the above schemata H, o (H) = 4, δ (H) = 6. The contents of schemata theorem, which is created by Professor Holland, is that schemata is a set of all the coding, among which values is fixed in certain positions in feasible zone. Schemata theorem considers that genetic algorithms are schemata calculation in essence. The shorter the alphabet coding, the more schemata are which the algorithm impliedly deals with a certain population. When an algorithm uses binary code, the

highest efficiency is achieved. If one population includes N individuals, it can also handle o (N^3) schemata at one time. The nature that genetic algorithms calculate such a small amount of code but deals with a large number of schemata is seen as inner parallelism. Clearly, traditional genetic algorithm has its inherent characteristics of parallel processing.

(2) Finite-state Markov chain model: because of lots of flaws of schemata theorem, researchers start to study genetic algorithm by means of finite-state Markov chain model. Regarding to the genetic algorithm which has m feasible solutions to an objective function and the population has N individuals, N individuals altogether has $\begin{pmatrix} N+m-1 \\ m-1 \end{pmatrix}$ methods, so does Markov model. Actually the quantity m of feasible solutions of optimization problems and population N is very considerable. Markov model thinks that different regions of search space are samples and fitness in different ones are estimated in order to figure out the probability that optimal solutions lie in different regions. Therefore you can change the samples in different regions so that optimal solutions with great precision are gotten. Obviously, as tallies with the reality take the neighborhood structure based on the division equal kind of Markov model, whose neighborhoods structure are based on equivalence partitioning, are more conform to the fact, and well manifest the essence of optimal solutions[2].

3 Describe TSP

TSP (Traveling Salesman Problem) is a typical combinatorial optimization problems and is also an NP Complete hard problem. It is very simple to describe TSP, but it is very difficult to solve it. It can be described like this: find out a shortest way to travel n cities without duplication path. In mathematics, it can describe as follows, the city set is (v_1, v2…v_n). The distance between any two cities v_i and v_j, d (v_i, v_j) =d (v_j, v_i)(1≤i, j≤N). Now find a city order, $c_{n(1)}$,$c_{n(2)}$…$c_{n(N)}$(1≤n≤N), making the path shortest: $TP = \min (\sum_{i=1}^{n} d(v_i, v_{i+1})), v_{n+1} = v_n$.

3.1 The procedure of GA

Generally genetic algorithm's main steps are as follows:

(1) Produces the initial population at random which is composed of the character string whose length is fixed.

(2) As for the string, we repeat to carry out the following step 1) and the step 2), conditioned the termination condition.

1) Compute adaptive value of each individual in the population.

2) Produce the next generation by using genetic operators such as selection, crossover, and mutation.

(3) Assign the best individuals which appears in the descendant to be the result of GA. The procedure can be described as figure 1[3].

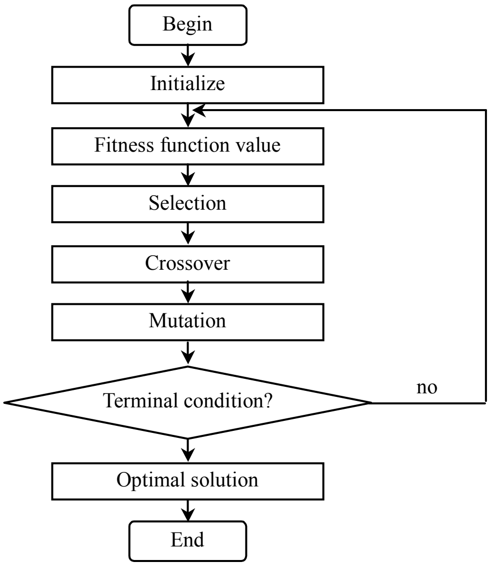


Figure 1 the flow process chart of GA

3.2 The steps an optimized genetic algorithm for TSP

3.2.1 Produce initial population

Randomly generate n initial population. The method of encoding in this paper is code-decimal notation. For example, the string 12345678 means you can travel from 1 city, to 2345678 by order, and in the end return to 1.

3.2.2 Fitness function

Fitness function is also called evaluation function. It is used to measure how well each chromosome adapts to the nature. Fitness function determines that each chromosome is good or not, reflects “survival of the fittest” in the nature. To optimization problems, fitness function is the objective function. The goal of TSP is finding out the shortest path, so the path length on the TSP can be used as fitness function.

$$f(v_1,v_2,...v_n)=1\Big/\sum_{i=1}^{n} d(v_i,v_{i+1}) \text{ ,} \quad v_{n+1}=v_n \quad (1)$$

It is necessary that fitness function can effectively reflect the gap between every chromosome and the optimum solution. If the gap is small, fitness function’s difference is also small. On the contrary, if the gap is big, the difference is also big [4].

For the minimum, cost function g(x) and the relationship between the appropriate function:

$$f(x)=\begin{cases} c_{\max}-g(x), g(x)<c_{\max} \\ 0, \qquad g(x)\geq c_{\max} \end{cases} \quad (2)$$

$c_{\max}$ stands for the largest factor. It can be equal to the maximum of g(x) in the process of organism evolution, so be the maximum of g(x) in the current group [5].

3.2.3 Selection

Selection operator also be called reproduction operator. Its principle is: fitness function value determines that the individual is eliminated or reproduced. That is to say, the higher the fitness value is, the more possible the chromos is selected and the higher frequency of replication, on the contrary if the fitness value f_i is very small, it is impossible for the individual to survive and reproduce and will be eliminated in the end. After the initial population is generated, you can choose different selection algorithm to calculate each individual’s fitness value. In this study expected value model is adapted, which can be described as follows:

1) Firstly calculate the mathematical exception f_i of sufficiency,

$$\overline{f_i}=\frac{1}{n}\sum_{i=1}^{n} f_i$$

2) Secondly calculated the mathematical exception R_i of each individual in the group to survive in the next generation. $\overline{R_i}=f_i\Big/\overline{f_i}$.

3) At last according to round-up in principle, change r_i to round numbers, which are the times of individual being selected. If R_i = 0 (that is to say, R_i<0.5), the individual i will be eliminated.

3.2.4 Crossover operator

In the nature, the combination of mother and father makes their respective genes reorganized, form the new chromosome and new individuals will be born. Similarly GA imitates in the same way the crossover and mutation which take place in the nature. GA produces a new generation through crossover and mutation.

In GA, crossover operator plays an important part. On one hand, it makes the original group of the fine characteristics of the individual to be maintained, which means the offspring and the parents are very similar. On the other hand, it allows GA can explore new gene space, so that the new population can be diverse. At the same time when thinking about crossover, you would better deal with encoding together so that the two can cooperate [6].

The process of crossover can be described as follows:

First, according to Pc (Crossover Probability), select randomly individuals from the group.

Next, randomly select crossover points from paired individuals, and exchange corresponding parts. Below the methods are used. In resolving TSP, PMX (partial matched crossover), OX (order crossover), CX (cycle crossover) and HX (heuristics crossover) are often used. Here we mainly introduce CX and HX.

1) The method CX is: according to its father’s

characters, each city is reunited under constraint conditions.

For example, parent p1 and p2,

P1=9 8 2 1 7 4 5 0 6 3,

P2=1 2 3 4 5 6 7 8 9 0,

First start from the left city

P1'=9 * * * * * * * * *,

P2'=1 * * * * * * * * *,

Then start from another city to find the next city

P1'=9 * * 1 * * * * * *,

P2'=1 * * * * * * * 9 *,

And then goes on like this, at last you can get

P1'=9 2 3 1 5 4 7 8 6 0,

P2'=1 8 2 4 7 6 5 0 9 3,

2) Set another example HX

Here are HX steps:

①At random choose a child as the initial city that the expression of the parents born.

②Choose the edges that are shortest and can not make a circle. If two edges constitute a cycle, a city is chosen randomly that can go on with the tour.

③If TSP cycle is completed, tour ends, or to ②[5].

3.2.5 Mutation

In the nature, because of some accidental factors replication errors may occur so that it is possible that certain genes may change, resulting in new genes and new organisms, which makes the offspring and the parents are different. This phenomenon can also take place in GA. Mutation can change certain individuals of the population in small probability at random. It will introduce variability to population, thus increasing the diversity of population and providing a means escaping from locally optimal solution and make sure population can go on with evolution.

Regarding to TSP, the below methods of mutation are used: inversion mutation, exchanged mutation, insertion mutation, shifted mutation, displacement Mutation, position-based mutation and so on. Here the first two are illustrated.

1) Inversion mutation

Reverse the code number within two points, for example

A =987 | 456 | 321,

After the operation into reverse

A'=987 | 654 | 321.

2) Exchanged mutation

Choose two crossover points random, and exchange their numbers in the sites. For example,

A = 12345678, of which the 4th and the 7th are crossover points, after exchanged mutation,

A'= 123756489.

3) Insertion mutation

Choose a city at random from the string, and insert it into a place at random. For example,

A = 123456789,

Suppose that choose the 8th city at random, and then insert it into the 4th at random. After insertion,

A'= 123845679.

In a word, it is more flexible to design mutation operation than crossover operation [7]. Anyone can be mutation operator on condition that it searches in the local field.

Among several methods of mutation, all of them do not take adjacency of edge into account except inversion mutation; therefore they can't retain the syntopy of edges which are created originally. The result is that the next generation can't inherit fine performances that the last one got in the tour and also cannot raise the optimization speed.

3.2.6 Simulation

According to the steps introduced above, now let's carry out an experiment on TSP.

(1) Code-decimal notation

(2) $f(x) = c_{max} - g(x)$, you can refer to ② to understand $f(x)$, c_{max} and $g(x)$.

(3) Excepted selection

(4) Heuristic crossover

(5) Inversion mutation

(6) Reference [8], n=1000, generation=150;

Reference [9], n=500, generation=4329, P_c =0.90,

P_m=0.10; in this paper n=100, Pc=095, P_m=0.003, generation=4 000.

Compare the other methods with the one in this paper. You can find out the shortest length in the Table 1, and the method in this paper is better.

Table 1 The results of different methods

methods	the shortest length
binary tree	428.90
heuristic search	436.01
reference [8]	424.86
reference [9]	424.8693
reference [10]	424.86929
the method in this paper	423.74

Applying the method in this paper, we get the path just as Figure 2 .

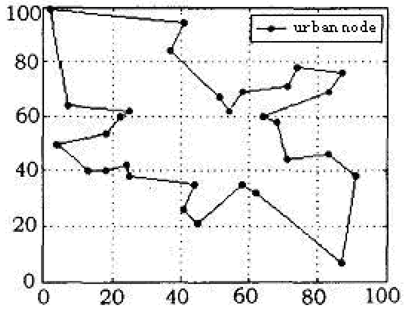


Figure 2 The Path of Oliver 30 TSP

4 Conclusion

Heuristic crossover and inversion mutation are used in order to prevent premature convergence, guarantee the diversity of population and make sure that effective genes can not be lost. This paper gives us more value of reference for our future study on methods of crossover operator and mutation operator. So in the future, we should go on with our further research on genetic operators.

References

[1] Fu Xuefang, “An Overview of the Modern Optimization Computing Method—Genetic Algorithms”, Journal, Changhuai college press, Changhuai, pp.42-45, Apr, 2001

[2] Wang Xiaoliang and Li Qiang “Application and Research on Parallel Genetic Algorithm”, Journal, softwore space-time, Guangzhou, pp.45-48, March,2007

[3] Cai Zixing, Artificial Intelligence Controlling, Beijing: chemical industry press, Beijing, pp.78-82, 2005

[4] Ma Shaoping, Zhu Xiaoyan, Artificial Intelligence, Tsinghua University Press, Beijing, pp.300-319. 2004

[5] Zhu Jianying, intelligent system non-classical mathematical methods, Huazhong University of Science and Technology Publishing House, Wuhan, pp.295-301, 1999

[6] Duan Yuqing, and He Jiali, “genetic algorithm and its modification”, Journal, China academic journal electric publishing house, Tianjin, pp.39-44, March 1998

[7] Jin Cong, Dai Shangping, Guo Jinglei, Zhang Wei, Artificial Intelligence Tutorial, Tsinghua university press, Beijing, pp.200-204 , 2007

[8] Xie Shengli, Zhang Yangu, and Li Guang, “solution for TSP based on GA”, journal, Wenzhou normal university, Wenzhou, pp.7-10, June 2002

[9] Gao Jingwei, Zhang Xu,Li Feng, and ZhaoHui, “Application of TSP”, journal , Wuxi college press, Wuxi, pp.19-21, Feb 2004

[10] Ao Youyun,and Chi Hongqin, “A Method Based on Genetic Algorithm for Solving TSP”, journal, Computer and digital, pp.52-54

Fuzzy Cognitive Maps Learning Based on Genetic Algorithm

Lifeng Yu[1] Jie Zhang[2] Jing Zhao[2]

1 School of International Education, Jiangnan University, Wuxi, Jiangsu, P. R. China

2 Center of Modern Education Technology, Shandong Institute of Light Industry, Jinan, Shandong, P. R. China

Email: yulifeng@jiangnan.edu.cn; zhangj@sdili.edu.cn; zj@sdili.edu.cn

Abstract

Fuzzy Cognitive Maps is a combination of fuzzy Logic and neural network, and it has got extensive applications in many fields. In this paper a new technique based on Genetic Algorithm for Fuzzy Cognitive Maps learning is introduced. The proposed approach is used for the detection of proper weight matrices that lead the Fuzzy Cognitive Map to desired steady states. For this purpose a properly defined objective function is constructed and minimized. The application of the proposed methodology to an industrial control problem supports the claim that the proposed technique is efficient and robust.

Keywords： fuzzy cognitive maps, GA algorithm; weight matrices, objective function

1 Introduction

Tloman originally proposed Cognitive Maps in 1948 [1]. Cognitive Map(CM)[1] is a useful model to represent and inference concepts' causal-effect relations in system. CM is directed graph, The concepts are represented as nodes, and the causal relationships between these concepts are represented as *edges*. Kosko [2] enhanced the power of cognitive maps considering fuzzy values for concepts of the cognitive map and fuzzy degrees of interrelationships between concepts. Then Fuzzy Cognitive Maps (FCM) were introduced in 1986 as signed directed graphs for representing causal reasoning and computational inference processing.

FCMs have been applied in knowledge representation reasoning and artificial intelligence, including modeling of intelligent system, fault checking, decision-making analysis, geography information systems, negotiable securities paper business, playing chess, controlling, and so on [3,4,5].

A few learning algorithms have been proposed in literature [6,7,8,9,10]. However, established algorithms are heavily dependent on the initial weight matrix approximation, which is provided by the experts. it need stronger mathematical justification, and further testing on systems of higher complexity. Moreover, the elimination of deficiencies, such as the Abstract estimation of the initial weight matrix and the dependence on the subjective reasoning of experts' knowledge, will significantly improve the performance of FCMs.

In this paper, an approach for FCMs learning, based on GA (Genetic Algorithm) algorithm, is presented. GA is applied to optimize the weight values of the FCM, and found the proper weight matrices for the system through defining the proper objective function, so the FCM leads to a desired steady state.

2 Fuzzy Cognitive Maps

FCM which amalgamate fuzzy theory and neural network is introduced by Kosko. Fuzzy Cognitive Map (FCM) is a soft computing tool, its knowledge denotation and reasoning ability is better. Concepts are pictured different aspects of the system and their behavior, and the dynamics of the system are represented by the interaction of concepts [2]. An FCM is signed and directed graph. An FCM models consists of nodes–concepts, C_i (i=1,2…N), where N is the total number of concepts. Each node–concept represents one primary factor of the system and it is depicted by a value $A_i \in [0,1]$(i=1,2…N). The weight, W_{ij}, indicates the causality between two concepts. The back node-concept value is influenced by front node-concept value and W_{ij}. The weight, W_{ij}, can descript different fuzzy information, the values of weights are in continuum [-1, 1]. $W_{ij}>0$, expresses positive causality; $W_{ij}<0$, expresses negative causality; $W_{ij}=0$, expresses no relation. A simple FCM with five nodes and ten weighted arcs range is illustrated in Figure 1

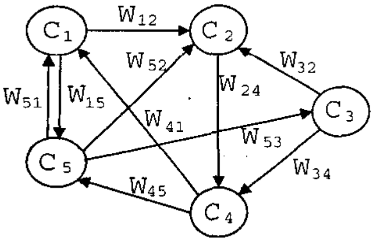


Figure 1 A simple Fuzzy Cognitive Map

The initial value A_i of a concept C_i, and W_{initia} are obtained by experts in the design process. $W_{initial}=[W_{ij}]$(i,j=1,2,…,N), $W_{ii}=0$(i=0,1,…,N). Then the FCM is let to converge to a steady state through the interaction subsequently described.

At each step, the value A_i of a concept C_i is influenced by the values of concepts–nodes connected to it, and is updated according to the equation [11]:

$$A_i(k+1) = f(A_i(k) + \sum_{\substack{j=1 \\ j\neq i}}^{n} W_{ji} A_j(k)) \qquad (1)$$

where k stands for the iteration counter; and W_{ji} is the weight of the arc connecting concept C_j to concept C_i. The function f is the sigmoid function:

$$f(x) = \frac{1}{1+e^{-\lambda x}} \qquad (2)$$

$\lambda > 0$. In our method the value is set to 1. This function can restrict the values A_i of the concepts within [0,1]. The interaction of the FCM results after a few iterations in a steady state, i.e. the values of the concepts are not modified further.

The two most significant weaknesses of FCMs are heavy dependence on the experts' opinion and the convergence to undesired steady states. However we can avoid the undesired steady states by amending the rights of FCM. The new learning algorithm is proposed to resolve it.

3 The New Learning Approach Basedon Genetic Algorithm

Genetic algorithm(GA) is a global searching algorithm used to simulate biology's genetic and evolutionary process and introduced by J.Holland in 1973[12]. It comes of nature and artificial adaptive system research. GA is a high-parallel, stochastic and adaptive optimization algorithm based on "Survival of the fittest". GA adopts simple coding technique to denote complicated structure and confirm searching direction by genetic operator and survival of the fittest.

GA is applied to update the weight values of the FCM, and determined the proper weight matrices for the system, so that leads the FCM to a desired steady state. The learning procedure is, to some extent, similar to that of neural network training. Set output concepts keep in strict bounds:

$$A_{out_i}^{\min} \le A_{out_i} \le A_{out_i}^{\max}, \quad i=1,2,\ldots,m$$

GA is applied to update the weight values of the FCM, and determined the proper weight matrices for the system, so that leads the FCM to a desired steady state. The learning procedure is, to some extent, similar to that of neural network training. Set output concepts keep in strict bounds:

$$A_{out_i}^{\min} \le A_{out_i} \le A_{out_i}^{\max}, \quad i=1,2,\ldots,m$$

U (i=1,2,…,m) is the steady state value of output concepts, which is obtained through the application of Eq.（1）. Obviously, the global minimization of the objective function F is weight matrices that lead the FCM to a desired steady state.

The learning process is introduced as follows:

1. initialize swarm, i.e. coding swarm. This step can randomly bring a swarm S={X_1，…，X_M}, M is the

size;

2. repeat

For i=1:MaxIt

（1）decode binary number to decimal number and calculate function value;

（2）calculate fitness function F. If the ending condition is reached, the algorithm stops. Otherwise continue next step;

（3）calculate the best good individual and the best bad individual;

（4）use select, crossover and mutation to generate next generation.

The coding and decoding approach:

If the value of X is in continuum [-1, 1] and precision is set 10^{-4}, then the continuum at least divide into $(b-a)*10^4$ parts. Set the binary number lengths are m, so

$$2^{m-1}< (b-a)*10^4 \leqslant 2^m-1$$

The formula of decoding binary number to decimal number is:

$$X=a+decimal(substring)*(b-a)/ 2^{m-1}$$

where decimal(substring) stands for the decimal value of variable X, X is a binary number substring.

4 Simulation Results

4.1 An industrial Process Control Problem

The proposed learning algorithm previously described, is applied on a simple industrial process control problem [13]. This process is consisted of one tank, three valves and one sensor, as illustrated in Figure 2 . Valve 1 and Valve 2 pour two different liquids into the tank. During the mixing of the two liquids, a chemical reaction takes place in the tank, and a new liquid is produced. When the new liquid produced reaches a specific level, valve 3 opens and empties the tank. A sensor is placed inside the tank to measure the specific gravity of the produced liquid. The objective of the system are firstly to keep the height of liquid between some limits, an upper limit H_{max} and a low limit H_{min} ; secondly the specific gravity of the liquid should be kept an upper limit G_{max} and a low limit G_{min}. When the specific gravity G lies in a range [G_{min}, G_{max}], the desired liquid has been produced.

The control objective is to keep values of these variables in the following range of values:

$$H_{min} \leqslant H \leqslant H_{max}$$
$$G_{min} \leqslant G \leqslant G_{max}$$

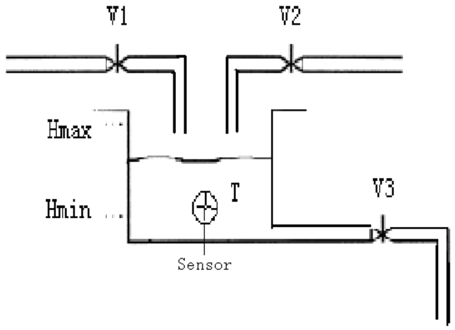


Figure 2 The illustration of a process problem

Fuzzy cognitive map that models and controls this system is depicted on Figure 3 . It consists of twelve concepts that are defined as:

- Concept 1——the amount of the liquid in tank 1.It depends on the operational state of Valve 1, 2, and 3;
- Concept 2——the state of Valve 1(closed, open or partially opened);
- Concept 3 the state of Valve 2(closed, open or partially opened);
- Concept 4——the state of Valve 3(closed, open or partially opened);
- Concept 5——the specific gravity of the liquid in the tank;

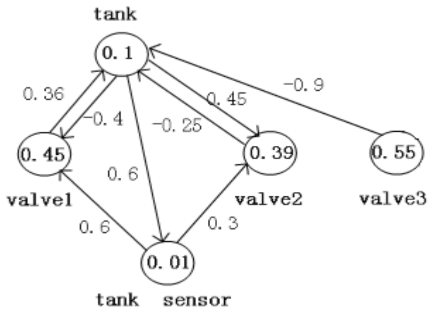


Figure 3 The FCM that corresponds to the problem of Figure 2

The ranges of the weights implied by the fuzzy regions are:

$-0.50 \leqslant W_{12} \leqslant -0.30$ $-0.40 \leqslant W_{13} \leqslant -0.20$

$0.20 \leqslant W_{15} \leqslant 0.40$ $0.30 \leqslant W_{21} \leqslant 0.40$

$0.40 \leqslant W_{31} \leqslant 0.50$ $-1.0 \leqslant W_{41} \leqslant -0.80$

$0.50 \leqslant W_{52} \leqslant 0.70$ $0.20 \leqslant W_{54} \leqslant 0.40$

the nonzero weight values of initial weight matrix is:

$$W_{initial} = \begin{bmatrix} 0 & -0.4 & -0.25 & 0 & 0.3 \\ 0.36 & 0 & 0 & 0 & 0 \\ 0.45 & 0 & 0 & 0 & 0 \\ -0.90 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.3 & 0 \end{bmatrix}$$

QPSO is applied to update eight nonzero weight values of the FCMs. The bounds [-1,0] or [0,1] implied by the directions of the corresponding arcs of the FCM, are imposed on each weight.

The output concepts regions are:

$0.60 \leqslant C_1 \leqslant 0.70$ $0.70 \leqslant C_5 \leqslant 0.75$

4.2 Simulation Results

In our experiments, a total of 100 independent experiments have been performed using GA. The parameters of GA are defined as follows: the swarm size is set to 30, crossover probability pc=0.8, mutation probability pm=0.1, most repeat times MaxIt=500. The accuracy for the determination of the minimized objective function has been equal to 10^{-12}.

The obtained sub-optimal matrices using GA is the following:

$$W = \begin{bmatrix} 0 & -0.4723 & -0.2157 & 0 & 0.3038 \\ 0.3900 & 0 & 0 & 0 & 0 \\ 0.4580 & 0 & 0 & 0 & 0 \\ -0.9595 & 0 & 0 & 0 & 0 \\ 0 & 0.3038 & 0 & 0.3038 & 0 \end{bmatrix}$$

Which leads the FCM to the desired steady state:

C_1=0.6120, C_2=0.6374, C_3=0.6195, C_4=0.7178, C_5=0.7101.

The convergent process of objective function F is follow.

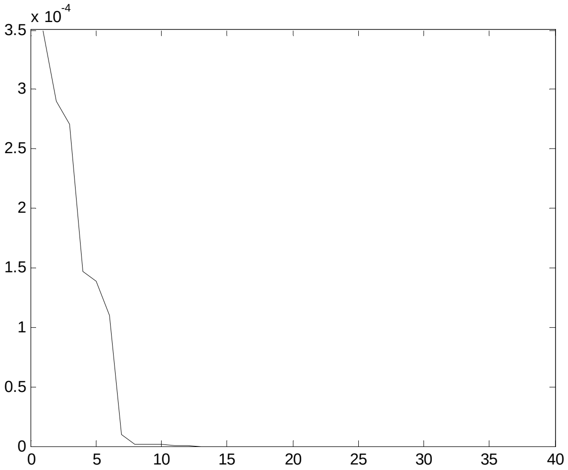


Figure 4 The convergence graph of objective function F

5 Conclusion

A new learning algorithm, which is based on GA, is introduced. The proposed approach is used for determining suboptimal weight matrix for Fuzzy Cognitive Maps with fixed structures, in order to lead the Fuzzy Cognitive Map to desired steady states. The workings of the approach are applied to an industrial control problem. The results support the claim that the proposed approach is a promising method for Fuzzy Cognitive Maps learning, and the method is effective and efficient.

References

[1] B. Chaib-draa, J.desharnais, “A Relational Model of Cognitive Maps”, International Journal of Human-Computer Studies, Vol.49, August 1988, pp. 181-200

[2] B. Kosko, “Fuzzy Cognitive Maps”, International Journal of Man–Machine Studies, Vol.24, January 1986, pp. 65-75

[3] J.S. Jang, C.T. Sun, E. Mizutani, Neuro–Fuzzy and Soft Computing, New York: Prentice Hall Pub., 1997

[4] M. Shamim Khan, Alex Chong, Tom Gedeon, “A Methodology for Developing Adaptive Fuzzy Cognitive Maps for Decision Support ”, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.4, No.6, 2000, pp. 403-407

[5] P. Craiger, M.D. Coovert, “Modelling Dynamic Social And Psychological Processes With Fuzzy Cognitive Maps”,The 3rd IEEE Conf. Fuzzy Systems, Vol. 3, June 1994, pp. 1873-1877

[6] Vazquez A. "A Balanced Differential Learning algorithm in Fuzzy Cognitive Maps", The Sixteenth International Workshop on Qualitative Reasoning, Barcelona, Spain, (2002)

[7] Elpiniki Papageorgiou, Chrysostomos Stylios, Peter Groumpos, Fuzzy Cognitive Map Learning Based on Nonlinear Hebbian Rule, Berlin: Springer Berlin / Heidelberg Pub., 2004

[8] Elpiniki I. Papageorgiou, Konstantinos E. Parsopoulos et al, “Fuzzy Cognitive Maps Learning Using Particle Swarm Optimization”, Journal of Intelligent Information Systems, Vol.25, July 1994, pp. 95-121

[9] Parsopoulos K.E, Papageorgiou E.I. et al “A First Study of Fuzzy Cognitive Maps Learning Using Particle Swarm Optimization”, IEEE Congress on Evolutionary Computation 2003, Vol.2,December 2003, pp. 1440-1447

[10] E. Papageorgiou , C. D. Stylios and P.P. Groumpos, “Activation Hebbian Learning Rule For Fuzzy Cognitive Maps”, 15th IFAC World Congress Barcelona, Spain, July 21-26, 2002

[11] B. Kosko, Fuzzy Engineering, New York: Prentice Hall Pub., 1997

[12] John H.Holland, “Genetic Algorithms and the Optimal Allocations of Trials”, SIAM Journal of Computing, Vol.2, 1973, pp. 88-105

[13] Chrysostomos D. Stylios, Peter P. Groumpos, " The Challenge of Modelling Supervisory Systems Using Fuzzy Cognitive Maps ", Journal of Intelligent Manufacturing Vol. 9, 1998, pp. 339-345

Research on the Calculation Method for Weight of the Feature Weighted Fuzzy Clustering Algorithm

Xiaojun Tong[1,2] Qin Jiang[1] Haitao Gan[1] Shan Zeng[1] Kai Zhao[1]

1 Department of Mathematics and Physics, Wuhan Polytechnic University, Wuhan 430074, China

2 Department of Control Science and Engineering, Huazhong University of Science and Technology
Wuhan 430074, China

E-mail: tongxiaojun1998 @yahoo.com.cn

Abstract

In the traditional fuzzy C-Means 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 blindness, 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 seriously. In this paper, we get a method for calculating the feature weight based on the feature contribution balance principle and the most separate degree principle of intra-cluster. By the IRIS example, we find that the calculation method for weight can not only enhance the calculation speed, and also make the clustering result better than the existing result.

Keywords: FCM; the feature weigh; the center of clustering

1 Introduction

The first people of studying the fuzzy clustering systemly is Ruspini [1], he defined fuzzy partition data set in 1969. At the same time, Zadeh [2] and Tarmura also bring forward the clustering methods based on the semblable relationship and the fuzzy relationship. But these methods are not suitable for large data set, so few works are gone on in the aspects. For solving the fuzzy clustering question, some people have done all kinds of tries, for example, in virtue of graph theory, convex decomposition of data set, dynamic programming and based on the difficult distinguished relationship and so on. However, owing to various reasons, these methods are not feasible. At present, the popular method is the clustering method based on objective function, this method is simple, and can solve many problems, furthermore, it also can be come down to an optimization problem. Hence, with the development of computer, this method turns into the main measure for fuzzy clustering analysis.

The fuzzy clustering method based on the objective function was brought forward by Ruspini [1], but Dunn [3] gave the most valid algorithm –fuzzy c-means algorithm. Afterward, Bezdek [4,5] extended the algorithm and established the fuzzy clustering theory. Thence, the fuzzy clustering method was developed vividly, and has formed an enormous system. The FCM algorithm is based on the objective function, and the study about the algorithm focus on the bellowing sides mainly: reforming the objective function [6,7], confirming the number of clustering parameter of c [8], studying the weight index of m [9,10], research on the cluster validity [11,12] and so on. Aiming at these aspects, the paper [13] gave an optimization algorithm based on evolution strategy using the feature weighted FCM algorithm. Contrasting the traditional FCM algorithm, the method can decrease the number of misclassified data points, furthermore, it also make the center of every class closer to the actual. All show that the feature weight is necessary. But the algorithm has some disadvantage: using search method to calculate the feature weight, and adopting random search, so the search direction is eyeless. Specially, we can not get the best weight when there are many features and calculation is complex. So we will construct a new method for calculating the weight based on the feature contribution balance principle and the most separate degree principle of different sorts. In the method, eyeless search is not necessary. Finally, contrasting the IRIS example, we find that the calculation method on weight not only enhance the calculation speed, and also make the clustering result better than the existing result [13].

2 The Calculation of Feature Weight

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: $p_1, p_2, \ldots, p_c$, $p_i = \{p_{i1}, p_{i2}, \ldots, p_{is}\}$. 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.

For the given original data, their features’ units are different, so can not be compared, but the features’ size express a sample’s characteristic in the corresponding feature. For the feature contribution balance principle, namely each feature contribution for clustering is important coequally. Suppose the balance coefficient be *rj*, we write it down as follows:

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

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:

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

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 $d_j = \sqrt{\sum_{i=1}^{c}(r_j p_{ij} - r_j \overline{p}_j)^2} \quad j = 1,2,\cdots,s$, lastly, we can get the feature weight *w*.

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

$$w_j = d_j \times r_j \quad j = 1,2,\cdots,s \qquad （3）$$

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 *dj* a power of *m*, thus we have the formula （4）:

$$w_j = (d_j)^{m'} \times r_j = \left(\sqrt{\sum_{i=1}^{c}(p_{ij} - \overline{p}_j)^2}\right)^{m'} \times r_j^{1+m'}$$

$$j = 1,2,\cdots,s \qquad （4）$$

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:

$$d_j = \sqrt{\sum_{i=1}^{c} (p_{ij} - \overline{p}_j)^2 \times r_j} \quad j = 1, 2, \cdots, s \qquad （5）$$

and

$$w_j = \sqrt{\sum_{i=1}^{c} (p_{ij} - \overline{p}_j)^2 \times r_j^2} \quad j = 1, 2, \cdots, s \qquad （6）$$

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:

$$w_j = \left(\sqrt{\sum_{i=1}^{c} (p_{ij} - \overline{p}_j)^2} \right)^{m_2} \times r_j^{m_1} \quad j = 1, 2, \cdots, s \qquad （7）$$

3 Example 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 and the new FCM algorithm to process the IRIS data. According to their misclassification number, we can evaluate their capability.

From the traditional FCM algorithm and the feature weighted FCM algorithm, IRIS data can be classed. Existing experiment tell us the optimal value of m is 1.8 [13], so we let m be 1.8. Then we will get the optimal c by computer program. The following shows the corresponding clustering validity function of $FP(U;C)$ while c choose different value.

c=2, fp=5.221414e-003, i=14

c=3, fp=3.176109e-003, i=36

c=4, fp=2.575802e-002, i=100

c=5, fp=4.497380e-002, i=67

c=6, fp=5.666194e-002, i=96

c=7, fp=1.440522e-002, i=57

c=8, fp=1.557129e-002, i=44

The above tell us that the clustering validity function $FP(U;C)$ arrives at the minimal when $c = 3$. It also proves the new FCM algorithm is valid and feasible.

After getting the optimal value of C and the corresponding clustering center of P ,we will process IRIS data by the traditional FCM algorithm and the feature weighted FCM algorithm, then process the fuzzy partition matrix by eliminating fuzzy (i.e. classify the samples to the corresponding class with the maximal membership grade), at last, the classification result is showed as the following:

Table1　the clustering result from the traditional FCM algorithm and the new FCM algorithm

clustering algorithm	The number of misclassified data points	The number of misclassified data points	The vector of clustering prototype	Error squme sum	*m*	*fc*
The traditional FCM algorithm	16	10.67%	p_1=(50062,34242,14684,02492) p_2=(58946,27460,44154,14273) p_3=(68484,30750,57283,20741)	0.1554	1.25	0.97055
The feature weighted FCM algorithm [13]	7	4.67%	p_1=(50060,34278,14624,02461) p_2=(59378,27450,43438,13315) p_3=(66274,30151,55673,20642)	0.0145	1.25	0.9878
The feature weighted FCM algorithm [13]	7	4.67%	p_1=(50060,34280,14621,02460) p_2=(59361,27441,43149,13289) p_3=(66284,30159,55694,20668)	0.0110	1.25	0.9880
The feature weighted FCM algorithm [13]	6	4%	p_1=(50064,34241,14675,02469) p_2=(59620,27591,43375,13409) p_3=(66537,30234,55980,20900)	0.0232	1.8	0.9069

note: the feature weighted FCM algorithm can improve the clustering effect. In this part, we cite paper [13] in order to evaluate the result. While iterative step is over, the corresponding weighted matrix is: $w = (0, 0.4699, 1.4628, 5.8717)$.

From table1, we find that the new FCM algorithm can decrease the number of misclassified data point, and it can also make the center of every sort closer to the actual. In the experiment, the parameters in the two algorithms are: $c = 3, \varepsilon = 10^{-5}$ .For the new FCM algorithm, $G_{\max} = 1000$ .While the interactive step is over, the weight calculated by the feature weighted FCM algorithm is：

While $m = 1.25$,

$w = (1.1513, 1.7808, 7.1383, 30.1492)$

While $m = 1.8$,

$w = (1.1698, 1.7348, 7.1501, 30.0958)$

In addition, we also know from table 1, in the new algorithm, while m choose different data, the clustering result is close, and the result while letting $m = 1.8$ is better than the corresponding result while letting $m = 1.25$, but fc is changed strongly. All things show that the value of fc is changed strongly not because of the better clustering result, but because of the change of m.

In order to make the feature with larger otherness give more contribution to classification, we can process the normalized separate degree by power, thus the distance of all classes and the separate degree will be larger, accordingly the clustering result will be better.

Suppose the power of the feature weight be m', the relationship of m' and fc is showed in the following figure.

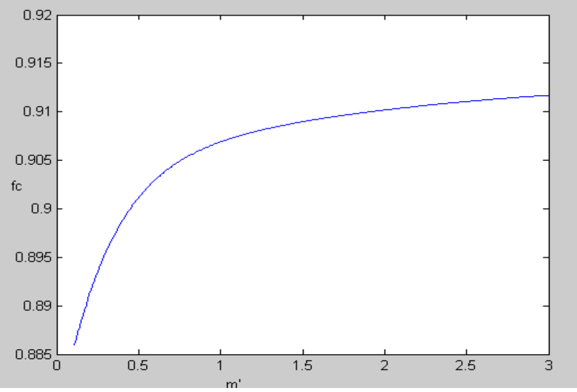


Figure1 the relationship of m' and fc

Figure 1 tells us that the value of fc is also increasing while the value of m' is increasing, but the former increases very fast while $m' < 1.5$, however, it increases slowly while $m' > 1.5$ .so we may let $m' = 1.5$. By computer program we get:

Table 2 The clustering result from setting $m' = 1.5$

m	The number of misclassified data points	The number of misclassified data points	The vector of clustering prototype	Error squme sum	fc
1.8	6	4%	p_1=(50065,34241,14674,02467) p_2=(59652,27591,43415,13413) p_3=(66516,30234,55961,20913)	0.0238	0.9090
1.25	7	4.67%	p_1=(50060,34280,14621,02460) p_2=(59425,27454m43206,13295) p_3=(66242,60254,55672,20681)	0.0113	0.9889

Let us outline the corresponding weights:

While

$m = 1.8$, $w = (1.2633, 1.6516, 15.3711, 75.4540)$

While

$m = 1.25$, $w = (1.2315, 1.7145, 15.3166, 75.3691)$

Table2 shows the clustering result while setting $m = 1.8$ is better than the result while setting $m = 1.25$, and m becomes smaller because of letting fc larger.

Processing the separate degree of intra-cluster with power is discussed in the above part, and we can also process the balance and the separate degree of the original prototype with powers. The relationship of m_1, m_2 and fc is illustrated in Figure 2 and Figure 3.

From the two figures, we have: the partition coefficient of fc becomes larger with the increase of the value of m_1, and it obtains a maximal value while m_2 is set to various values.

The following table shows the optimal value of m_2 while m_1 is set to various values.

Let $m_1 = 3$ and $m_2 = 0.55$, from computer program, we obtain the clustering result showed in table4 :

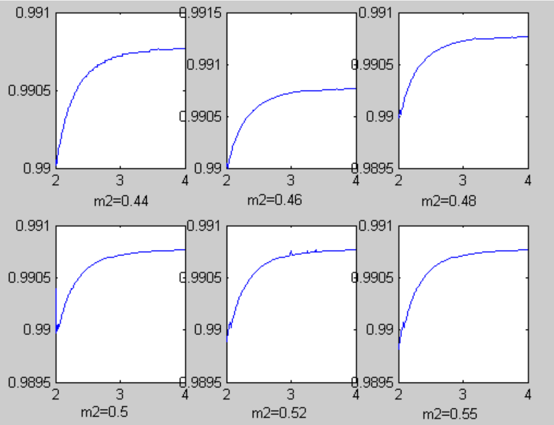


Figure 2 The relationship of m_1 and fc while m_2 is set to a fixes value

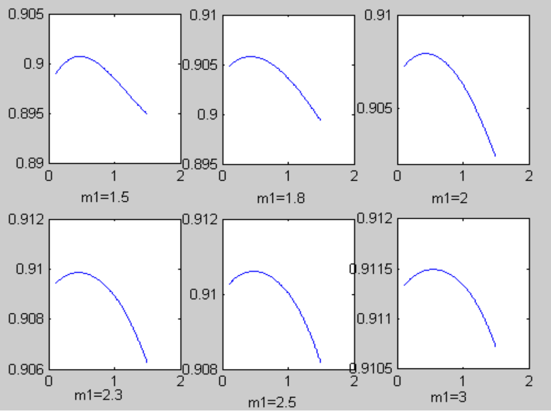


Figure 3 The relationship of m_2 and fc while m_1 is set to a fixes value

Table3 The optimal m_2 while setting m_1 to various values

m_1	1.5	1.8	2	2.3	2.5	3
m_2	0.48	0.44	0.44	0.46	0.48	0.55

Table 4 The clustering result while setting $m_1 = 3$ and $m_2 = 0.55$

m	The number of misclassified data points	The number of misclassified data points	The vector of clustering prototype	Error squme sum	fc
1.8	6	4%	p_1=(50066,34234,14681,02461) p_2=(59758,27600,43571,13438) p_3=(66471,30239,55914,20956)	0.0272	0.9130
1.25	7	4.67%	p_1=(50060,34280,14621,02460) p_2=(59545,27478,43276,13291) p_3=(66124,30130,55598,20678)	0.0112	0.9907

The corresponding weights are:

While

$m = 1.8$, $w = (1.0883, 4.6469, 6.7291, 127.2641)$

While

$m = 1.25$, $w = (1.554, 4.7258, 6.6754, 123.1579)$

From the above, we find: the value of fc is changed larger, but the clustering result is not better than the result showed in the table 2.

Now we use other data to test the result. And the data is subject to: (1) $m_1 = m_2 + 1$, (2) try best to make the value of fc larger. Let $m_1 = 1.5$ and $m_2 = 0.5$, we get the corresponding clustering result by computer program:

Table 5 the clustering result while $m_1 = 1.5$ and $m_2 = 0.5$

m	The number of misclassified data points	The number of misclassified data points	The vector of clustering prototype	Error squme sum	fc
1.8	5	3.33%	p_1=(50063,34239,14678,02472) p_2=(59574,27589,43325,13410) p_3=(66577,30241,56009,20886)	0.0230	0.9012
1.25	6	4%	p_1=(50060,34280,14621,02460) p_2=(59274,27423,43070,13279) p_3=(66344,30164,55722,20644)	0.0108	0.9879

And the corresponding weights are:

While $m = 1.8$,

$$w = (1.0827, 1.8219, 3.3249, 12.0077)$$

While $m = 1.25$,

$$w = (1.0746, 1.8475, 3.3264, 12.0672)$$

Table5 shows: the clustering result while setting $m_1 = 1.5$ and $m_2 = 0.5$ is better than setting $m_1 = 3$ and $m_2 = 0.55$ .Because the clustering result from the latter is not the best, whereas it can choose powers agilely, and its partition coefficient of fc can gain larger values.

4 Conclusion

For the traditional FCM algorithm, the new FCM algorithm is unsupervised and automatic, and it can decrease the number of misclassified data points, furthermore, it also makes the clustering prototype closer to the actual.

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

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

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

[5] 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

[6] Trauwaert E, Kaufman L, et a1. Fuzzy clustering a1gorthms based on the maximum likelihood principle. FSS, 1991, 42: 213~227

[7] Bobrowski L, BezAek JC. C-means clustering with the 11and l norms. IEEE Trans SMC, 1991, 21（3）: 545~554

[8] Guo haixiang, Zhu kejun. Determination of Optimal Classification Number with Genetic Algorithm Based on Matlab .Journal of Chang chun university of technology(natural science edition) 2004,25（1）:12~15

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

[10] Chan K P, Cheung Y S. Clustering of clusters. Pattern Recogition, 1992,25（2）:211~217

[11] Bezdek J C, Pal N R. Some new index of cluster validity. IEEE Trans. SMC, 1998, 28（3）:301~315

[12] Dave R N. Validating fuzzy partitions obtained throngh c-shells clustering. PRL, 1996, 17：613~623

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

Data Mining Based on Genetic Algorithm

Yonghua Qin

1 School of Information technology, Jiangnan University, Wuxi, Jiangsu, 214122

2 Wuxi Telecom, Jiangsu, 214000

Email: yhqin@wst.net.cn

Abstract

Genetic algorithms are considered as a global search approach for solving the complex problems. The procedure of GA is analyzed in detail and the evolutionary operators of GA are introduced. The data mining method is also proposed and the method of GA using in data mining is proposed. At last, the paper presents the GA for data mining of the association rules. Experimental results show that in the association rules data mining, GA could successfully finds the useful knowledge.

Keywords: data mining; genetic algorithm; association rules

1 Introduction

Genetic Algorithms(GAs) are a family of computational models inspired by evolution[1].GAs provide a randomized, parallel, and global search approach to find the optimum solution of problems, especially for optimization problems, based on the mechanics of natural selection and natural genetics. These algorithms encode a potential solution to a specific problem on a simple chromosome like data structure and apply recombination operators to these structures so as to preserve critical information. GAs are often viewed as function optimizers, although the range of problems to which genetic algorithms have been applied is quite broad. GAs have been shown to be an effective tool to use in data mining and pattern recognition[2-4]. GA is usually used as an optimization tool for solving the parameters of the specified problems. There're many application that GA can be used, such as breaking the steganalytic systems[5], image watermark retrieval enhancement[6], and etc.

Generally, data mining [7] or knowledge discovery is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. It has become a popular tool utilized in huge databases to find unsuspected relationships, sometimes called hidden patterns, for predicting future activities [8, 9].Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits.

This paper is organized as follows. In section 2, some methods of data mining is introduced. Section 3 gives the process of GA. The data mining based on GA is viewed in section 5. Finally, conclusions are presented.

2 Data Mining

There are three stages for data mining processing, data pre-processing, data mining tools applied and interpretation and evaluation. For the first stage, heterogeneity resolution, data cleansing and data warehousing are the components. In the second stage, it is needed to extract the patterns from the pre-processed data. At the last stage, the users can get the interest data, such as attributes of interest in databases, goal of discovery, domain knowledge, prior knowledge or belief about the domain. The process can be seen in Figure 1.

In the data mining system, there exists three layers(in figure 2), which are data sources, data mining tools and user layer.

The ways of data mining is the most important in the process of data mining. The methods list below is the most used in the data mining.

- Memory-based reasoning:MBR
- Market basket analysis
- Decision Trees
- Genetic algorithm
- Cluster detection
- Link analysis
- on-line analytic processing:OLAP
- Neural Networks
- Discriminant analysis
- Logistic analysis

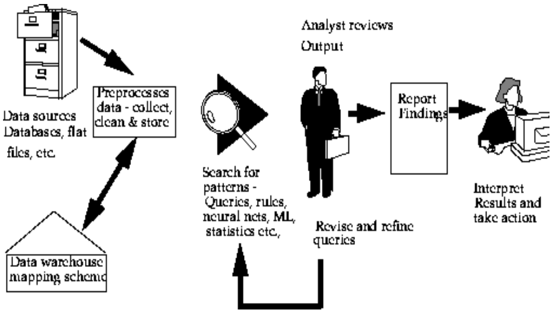

Data sources
Databases, flat files, etc.
Preprocesses data - collect, clean & store
Data warehouse mapping scheme
Search for patterns - Queries, rules, neural nets, ML, statistics etc.,
Analyst reviews Output
Report Findings
Interpret Results and take action
Revise and refine queries

Figure 1 Data mining process

3 Genetic Algorithm[10]

Most of the GAs works as the following steps. Firstly, a population is created with a group of individuals created randomly. Each individual is represented as a chromosome and in the algorithm it is a candidate solution. The individuals in the population will be evaluated according to the fitness function which is provided by the problem. Secondly, the individuals are then selected based on their fitness, the higher the fitness, the higher the chance of being selected. These individuals then "reproduce" to create one or more offspring, after which the offspring are mutated randomly. This procedure continues until a suitable solution has been found or a certain number of generations have passed, depending on the needs of the programmer.

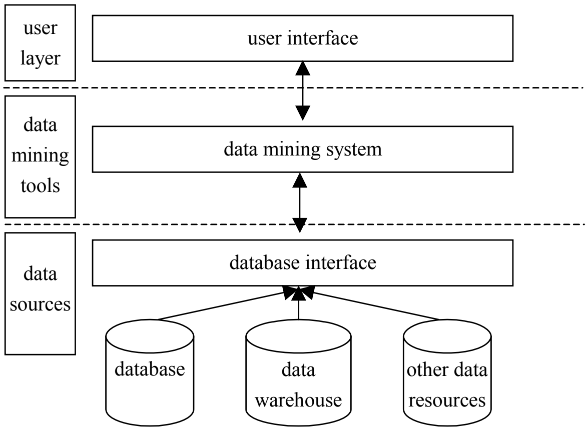


Figure 2 Data mining system

There are many different types of selection, such as roulette wheel selection, tournament selection and stochastic universal sampling. The most common type are roulette wheel selection. In roulette wheel selection, individuals are given a probability of being selected that is directly proportionate to their fitness. Two individuals are then chosen randomly based on these probabilities and produce offspring.

After the selection, the crossover operator occurs. There are many different kinds of crossover, such as single-point crossover, multiple-point crossover and uniform crossover. The most common type is single point crossover. In single point crossover, a random point is chosen at which user swap the remaining alleles from on parent to the other. This particular method is called single point crossover because only one crossover point exists. Crossover occurs according to a set probability. If there's no crossover occurs, the parents are copied directly to the new population.

After selection and crossover, a new population full of individuals is generated. Some are directly copied, and others are produced by crossover. In order to ensure that the individuals are not all exactly the same, mutation operator can be processed according to a little probability.

4 Experiments for GA Used in Data Mining

The first issue of data mining using GA is coding for the real problems, which can use the binary coding or decimal coding. The second issue is to define the fitness function. In the paper, we analysis the rule induction, then the fitness function can be defined as positive example or the counterexample of the rule cover. Generate a set of rules randomly, and judge the given example, calculate the fitness according to the fitness function. In the following steps, crossover operator and mutation operator execute on the set of rules, and the selection operator will be used. After a number of generation, the algorithm will be stopped when it satisfy the condition. The optimal rule can be get by the algorithm. Repeat the above steps and the get the compact rules.

The steps of GA used in data mining

Step 1. Initialize a population, P={P1,P2,…,Pn} and get the support S, confidence C.

Step 2. Calculate the fitness value (f) of each individual according to the fitness function, f=S'/S; then select the satisfied individuals if f(P)>0, otherwise, delete the individual. After the selection, if there're not enough individuals, generate the missing number of individuals randomly.

Step 3. Execute the crossover operator according to a certain probability Pc.

Step 4. Execute the mutation operator according to the given probability Pm.

Step 5. Repeat step 2 to step 4 until a certain condition is satisfied.

The following is an example for data mining based on GA.

According to our company's staff information, a database is found. In order to calculate the fitness, all the string type field is converted to integer type. The construct of the database is (staffID, staffName, staffSex, staffType, staffBirthdate, staffBehave, staffPortfolio). Based on the database, we do the data mining of Association rules. The parameters of GA is: the length of chromosome is 8, population size is 100, crossover rate Pc is 0.8, mutation rate Pm is 0.01. The stop condition is after five iterations, there's not any rules that less than the given value.

<001>→<300> (7% support, 85% confidence), that is <staff:manager>→<portfolio:80-85>

<3>→<100001>(60% support, 99% confidence), that is <staffsex:male>→<behavior:80-85>

<2>→<100>(6% support, 100% confidence), that is <staffsex:female>→<rewards and punishment:rewards>

These results show that the algorithm is effective and confident. Through the same methods, aiming at the different contents of staff information, do the Association rules data mining, we can get some other useful knowledge and then supervise us to manage the staff.

5 Conclusions

In this paper, data mining and genetic algorithm for data mining has been introduced and analyzed. The procedure of data mining is detailed presented. Genetic algorithm is also used in an example for Association rules data mining. The experiment results show that the algorithm is useful and effectively.

References

[1] J. H. Holland, Adaptation in Natural and Artificial Systems, The University of Michigan Press, 1975

[2] De Jong K.A., Spears W.M. and Gordon D.F. 1993. Using genetic algorithms for concept learning. Machine Learning 13, 161-188, 1993

[3] Freitas, A.A. A survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery, See: www.pgia.pucpr.br/~alex/papers. A chapter of: A. Ghosh and S. Tsutsui. (Eds.) Advances in Evolutionary Computation". Springer-Verlag, 2002

[4] Jain, A. K.; Zongker, D. Feature Selection: Evaluation, Application, and Small Sample Performance", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 19, No. 2, February 1997

[5] Y.-T. Wu and F. Y. Shih, "Genetic Algorithm Based Methodology for Breaking the Steganalytic Systems," IEEE Trans. on Systems, Man and Cybernetics, Part B, vol. 36, no. 1, pp. 24-31, Feb. 2006

[6] F. Y. Shih and Y.-T. Wu, "Enhancement of image watermark retrieval based on genetic algorithm," Journal of Visual Communication and Image Representation, vol. 16, pp. 115-133, April 2005

[7] U. Fayyad, "Data mining and knowledge discovery in databases: implications for scientific database," Proc. of the 9th International Conference on Scientific and Statistical Database Management, pp. 2-11, 1997

[8] C. Apte, B. Liu, E. P. D. Pednault, and P. Smyth,"Business applications of data mining," Communications of the ACM, vol. 45, no. 8, pp. 49-53, 2002

[9] X. Chen, X. Zhou, R. B. Scherl and J. Geller, "Using an interest ontology for improved support in rule mining," Proc. of the 5th International Conference Data Warehousing and Knowledge Discovery, pp. 320-329, Prague, Czech Republic, Sep. 2003

[10] http://geneticalgorithms.ai-depot.com/Tutorial/Overview.html

Path Plan of Robot Based on Neural-Fuzzy Control System

Fang Bao[1,2] Yonghui Pan[1,2] Wenbo Xu [2]

1 Jiangyin Polytechnic College. No.168, xicheng road, Jiangyin Jiangsu, China 214405

2 School of Information Technology, Jiangnan University. No. 1800, Lihudadao, Wuxi Jiangsu, China 214122

Email: baofang@mail.jypc.org

Abstract

According to the issue of dynamic path plan of mobile robot in unknown environments from start to the destination with obstacle avoidance, a systemic neural-fuzzy control algorithm is proposed. Fuzzy logic control system is designed to do the input fuzzification, fuzzy reasoning rule base, output defuzzification. The simplified structure of neural network handling the fuzzy control is also designed. Train the network using QPSO. Solve the “dead cycle” problem in U-shaped obstacle through the storage and management strategy of status variable of robot. Experimental results show that under the control of the proposed systemic algorithm, mobile robot can moving toward the target, avoiding all kinds of obstacles, dynamically planning reasonable path, not getting into the dead cycle.

Keywords: fuzzy neural-fuzzy control; dynamic path plan; QPSO; status variable

1 Introduction

Independent navigation of mobile robot in complex dynamic environment is an important issue in robot and artificial intelligent research fields. For real-time automatic navigation, the robot must have the capability of sensing the ambience, ensuring self position, obstacle avoidance, adjusting the direction and speed of itself, thus plan a trajectory from start to the destination. All the above factors are related to the reasoning and controlling upon uncertainty.

Neural-fuzzy system[1] integrate the uncertain expression, logic reasoning capability of fuzzy reasoning system and the adaptive learning, non-linear parallel computing capability of neural network. Based on the input environment variable, by the input fuzzification, fuzzy reasoning, output defuzzification steps, implemen ted by the neural network, neural-fuzzy control system could perform the fuzzy control task, as well as dynamic path plan of robot in unknown situation.

The conventional neural-fuzzy system exist the problem of big network scale, low training speed via using grade-descend training method, and the problem of getting into dead cycle when striding across the U-shaped obstacle is also not be solved well·[2]

According to the issue, this paper proposed a systemic algorithm, under the control of such algorithm, mobile robot could move to the target, plan a real-time path, and not get into the dead cycle.

The following is organized as: sensor system for the path plan is design in section 2, section3 presents the design and realization of neural-fuzzy control system, including fuzzy logic reasoning mechanism, neural network system and network training. Section 4 introduces the U-shaped obstacle avoidance algorithm, the simulation studies and conclusions are provided in section 5.

2 Design of Sensor System

For the purpose of moving towards the target and keep away from all the obstacles, proper sensors must be mounted on the robot to catch the information of environment as well as the relative position of itself.

Firstly, 9 ultrasonic sensors in 3 groups are mounted on the front of robot's head, each group measures the distance to the obstacle in front, right and left, so, both the position of the obstacle and the relative position of itself are achieved. Secondly, an optical localizer is mounted on the top of the robot's head to position the target and guide the direction..

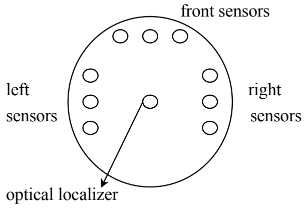


Fig .1 sensor system of the robot

The information collected by above sensors includes the position of the obstacle and the robot, the direction of target. Input such information into the neural-fuzzy system, through the reasoning process, the real-time control information could be outputted. The control information is the moving direction of the robot, mean left, front-left, front, front-right, right and back, thus a dynamic path is achieved step by step.

3 Design of Neural-Fuzzy Control System

3.1 Fuzzy logic reasoning mechanism

Fuzzy logic reasoning mechanism is designed to do the input fuzzification, fuzzy reasoning rule base and output defuzzification.

Fuzzification process transfers the input information caught by sensors to linguistic fuzzy terms. The distance to the obstacle in left, right and front, l_d, r_d and f_d, is transfer to be "far" or "near". Angle between the moving direction and robot-target joint line, θ_d, to be "left", "front" and "right". Membership function used here is the triangle function:

$$p_{ij} = \begin{cases} 1 - \dfrac{2\left|u_i - m_{ij}\right|}{\sigma_{ij}}, if m_{ij} - \dfrac{\sigma_{ij}}{2} < u_{ij} < m_{ij} + \dfrac{\sigma_{ij}}{2} \\ 0, otherwise \end{cases} \quad （1）$$

Where i=1…4, u_i is the input information, {u_1，u_2，u_3，u_4} = { l_d ， r_d ， f_d ， θ_d }, j=1…3, j represents the maximum numbers of fuzzy properties, pij represents the membership degree that the ith input belongs to the jth property, mij is the center of the membership function while σ_{ij} is the width of it, usually be chosen as a constant.

The membership function of l_d, r_d and f_d is shown in the left of Figure 2, while the right part is the membership function of θ_d.

Fuzzy reasoning rule base is designed based on the expert's experience, table.1 shows it:

Table.1 reasoning rule base of the neural-fuzzy control system

	input				output
	ld	rd	fd	θd	O
1	far	far	far	left	l
2	far	far	far	front	f
3	far	far	far	right	r
4	far	far	near	left	l
5	far	far	near	front	r
6	far	far	near	right	r
7	far	near	near	left	l
8	far	near	near	front	l
9	far	near	near	right	l
10	far	near	far	left	fl
11	far	near	far	front	f
12	far	near	far	right	f
13	near	near	near	left	b
14	near	near	near	front	b
15	near	near	near	right	b
16	near	far	near	left	r
17	near	far	near	front	r
18	near	far	near	right	r
19	near	far	far	left	f
20	near	far	far	front	f
21	near	far	far	right	fr
22	near	near	far	left	f
23	near	near	far	front	f
24	near	near	far	right	f

Where the fuzzy reasoning output is the moving direction of the robot, representing in linguistic fuzzy terms as back(b), left(l), front-left(fl), front(f), front-right(fr) and right(r).

Defuzzification process transfers fuzzy output to a crisp signal, the "center of gravity" method is used here, the final output O is:

$$O = \frac{\sum_{k=1}^{24} v_k q_k}{\sum_{k=1}^{24} q_k} \qquad （2）$$

$$q_k = \min\left\{ p_{1k}, p_{2k}, p_{3k}, p_{4k} \right\} \qquad （3）$$

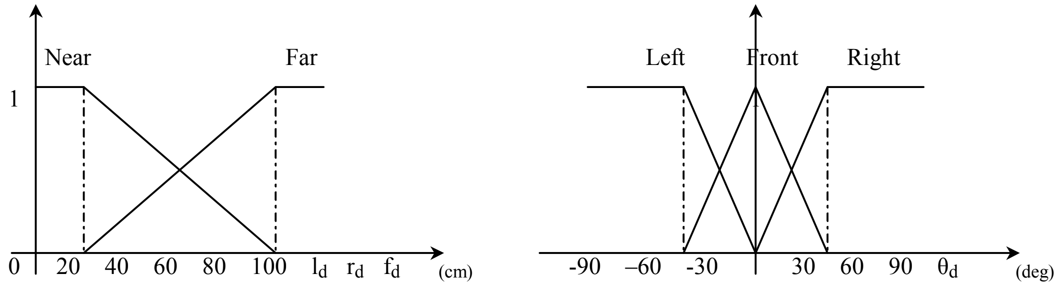


Figure 2　membership function of input information

Where k=1…24 is the number of rules, v_k is the center of the output membership function, q_k is the conjunction inspirit intensity of all input to the kth rule, P_{ij} is define in formula （1）. The output membership function is also a triangle function shown in Figure 3.

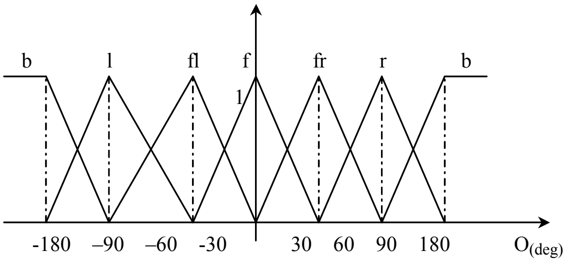


Figure 3　membership function of output signal

The above input fuzzification, fuzzy rule base and output defuzzification process establishes the fuzzy reasoning foundation of the proposed neural-fuzzy control system. In the case neural network be used to implement such fuzzy reasoning, parameters of membership function could be optimized by network training mechanism, and the effective fuzzy control be achieved.

3.2　Neural network system

A 5-layer network including input layer, fuzzification layer, fuzzy reasoning layer, defuzzification layer, output layer is designed to handle the fuzzy reasoning. The structure of neural network is shown in Figure 4.

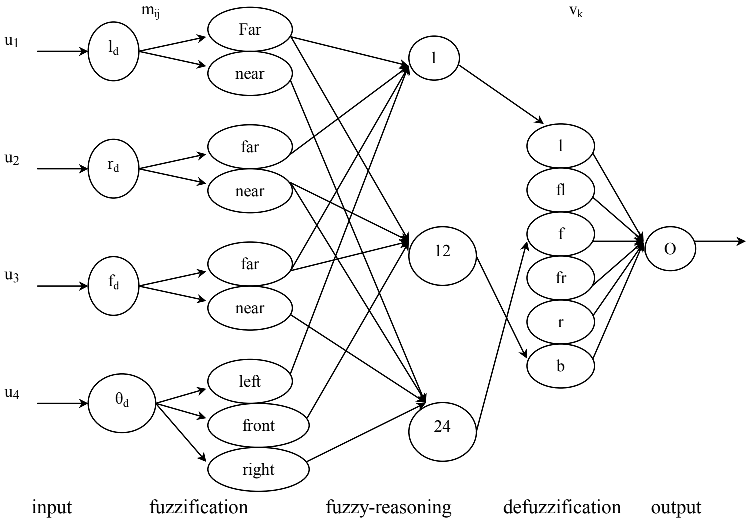


Figure 4　network structure of neural-fuzzy control system

In this network, input and fuzzification layer perform the input fuzzification via input membership function, the weight is mij. Defuzzification and output layer perform the output defuzzification via formula（2）, the weight between them is vk. Fuzzification, fuzzy-reasoning and defuzzification layers represent each rule, each neuron in fuzzy-reasoning layer have 4 inputs and 1 output according to the rule base, the weight between them all could be apprehended as relating with m_{ij} and v_k. So, the weight need to be learned in this network is mij and v_k.

It is obviously, the designed network has simple and clear structure, the weights of it are also be simplified.

3.3 Network training based on QPSO

From the above analysis, the weights need to be trained in this network is the following vector:

$$p=\begin{Bmatrix} m11, m12, m21, m22, m31, m32, m41, m42, m43, v1, v2, v3, v4, v5, v6, v7, v8, \\ v9, v10, v11, v2, v13, v14, v15, v16, v17, 18, v19, v20, v21, v22, v23, v24 \end{Bmatrix} \qquad （4）$$

The objective function of network training is the approximate error between actual output and the desired output, let T be the desired output given by expert, the objective function is:

$$E=\frac{1}{2}\left|T-O\right|^2 \qquad （5）$$

Train network with traditional grade-descend method will cause low speed and local convergence, so, the proposed network is trained using Quantum-behaved Particle Swarm Optimization (QPSO) [3].

QPSO is derived from Particle Swarm Optimization (PSO)[4] that proposed by Kennedy and Eberhart in 1995. QPSO initializes a group of random particles (random solution), finds the best value of itself and the best value of the swarm through the fitness function that determined by the essence of the application. Then the loops begin, in each generation, particle's position updates, and the new best value of itself and swarm are computed. Once the maximum generations or the minimum fitness value is attained, the algorithm is finished and the best value of the swarm is the optimized solution of the application. It has been proved that QPSO have better convergent speed and global convergent ability [5].

Once the structure of neural network is ascertained, the weights of all connections in network could construct one particle of QPSO, means such a particle represents a set of network's weights. Made the error function between desired output and actual output to be the fitness function of QPSO, the optimization process of QPSO could found out the particle which has best fitness value. Meanwhile the set of weights that induce the minimum output error is found. So the network could be trained using QPSO[6].

In the proposed network, because the structure has been ascertained, so we can train it with QPSO, made formula（4）be the particle of QPSO, made formula（5）be the fitness function, and the weights of network could be found more quickly.

4 Dead Cycle Avoidance Algorithm in U-Shaped Obstacle

Generic adaptive robot will face the trouble of getting into dead cycle in U-shaped obstacle, analyze the situation in Figure 5 (a):

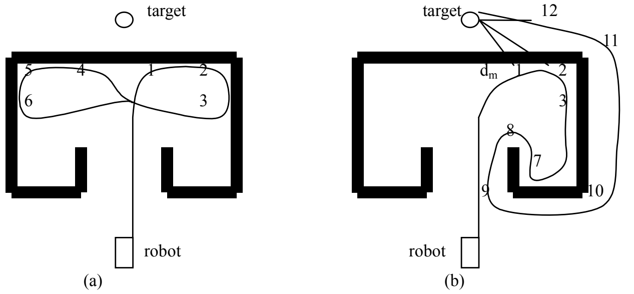


Figure 5 dead cycle in U-shaped obstacle and avoidance

At first, because the target is in front and there is no obstacle be caught, follow rule 2, the robot go straight until reach position 1. Follow rule 5, it turn right, and go straight to position 2 follow rule 19. Follow rule 16, it turn right and go straight until reach position 3. Now, it will turn front-right based on rule 21, and so on. In position 6, it turns front-left based on rule 10, and goes to position 1, the dead cycle is formed.

The key point is, at position 3, if the robot doesn't turn front-right, but goes straight ahead, then the dead cycle could be avoided. So, if the robot thinks that the target still been in the left side but not in the right side, it will go straight based on rule 19.

So, the robot is been designed to remember its own status, and some of the environment information could be specified in certain status. In this way, the dead cycle could be avoided without change the rule base.

The robot's status is designed to be 3 kinds. Generally, the robot is in status 0, its action is strictly follows the true environment information and the rule base. When the situation in table.2 appears, its status turns to 1 or 2, and its action follows the specified environment information and the former rule base.

Table 2 the special status of the robot

obstacle	target	angle between the moving direction and robot-target joint line	status
left	left	largen	1
right	right	largen	2

Specify that in status 1, whatever target been caught by the robot, the target is been set in the left side. As for obstacle, if it is been caught by robot, the obstacle is been set as the true situation, if it is not been caught, the obstacle is been set at the front.

And in status 2, whatever target been caught by the robot, the target is been set in the right side. And for obstacle, if it is been caught by robot, the obstacle is been set as the true situation, if it is not been caught, the obstacle is been set at the front.

So, in Figure 5 (b), after position 1, the robot's situation accords with status 1, its status turns to 1, and its environment information is been specified according to above specification. Then, based on rule 19, the robot goes straight ahead. In position 2, it turns right based on rule 16. In position 3, because the target has already been specified on the left side, it will go straight based on rule 19. In position 7, it turns right based on rule 16. In position 8,9,10 and 11, it will turn left continuously based on rule 4. The robot strides across the U-shaped obstacle successfully.

The last point, how the robot's status turns back to status 0?

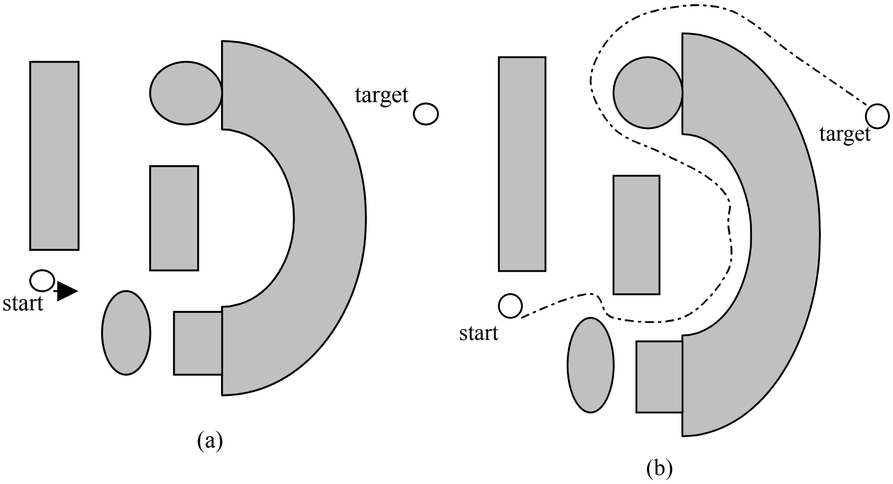
The robot is designed to remember a distance, when its status turns from 0 to 1, the current distance from robot to the target dm is recorded (see in Figure 5 (b)). At each point, the robot compares the current distance from itself to the target with the initial dm, it can be judged that when current distance is greater than dm, the U-shaped obstacle has not been stride over, only when the current distance is smaller than dm in position 12, means the obstacle has been stride over successfully. Now, the robot's status could be turn to 0.

Through the status and distance storage and manage strategy, the proposed systematic algorithm solve the problem of getting into dead cycle in U-shaped obstacle.

5 Simulation and Analysis

To demonstrate the efficiency of the proposed neural-fuzzy control algorithm, a mobile robot is used, the sensor system is shown in Figure 1, its speed is 0.1m/s.

First, the network used in the neural-fuzzy system must be trained. The circle, rectangle, square, U-shaped obstacles and the target have been mounted in the training environment as shown in Figure 6 (a).



(a)

(b)

Figure 6 dynamic path plan of robot

Initial value of parameters in formula （4）is set based on experience:

$$p=\left\{\begin{matrix}90,30,90,30,90,30,-45,0,45,-175,-160,-145,-130,-105,-90,-75,-60,\\ -45,-30,-15,0,0,15,30,45,60,75,90,105,130,145,160,175\end{matrix}\right\} \tag{6}$$

The corresponding width of input membership function is set as constant:

$$\begin{aligned}&\sigma 11,\sigma 12,\sigma 21,\sigma 22,\sigma 31,\sigma 32,\sigma 41,\sigma 42,\sigma 43\\ &=\{160,160,160,160,160,160,90,90,90\}\end{aligned} \tag{7}$$

Train the network using QPSO-based training method, the optimized parameter of the proposed algorithm is:

$$p=\left\{\begin{matrix}96,20,100,26,98,23,-49,0,50,-178,-165,-140,-130,-100,-95,-76,-60,\\ -40,-30,-17,0,0,15,32,43,65,75,90,102,124,140,162,170\end{matrix}\right\} \tag{8}$$

Using this parameter set, under the control of proposed neural-fuzzy control algorithm, with dead cycle avoidance in U-shaped obstacle, the robot could plan the path dynamically, stride over all kinds of obstacle, the path planed is shown in Figure 6 (b).

Change the training environment, the achieved parameter set is changed too. Take certain weight-average parameter set, in general test environment, the proposed control system could achieve the correct path plan.

If the dead cycle avoidance algorithm in U-shaped obstacle is omitted, then, in the same environment, the robot will get into dead cycle absolutely.

The proposed neural-fuzzy control algorithm designs effective fuzzy logic reasoning mechanism and clear network structure, trains the network with QPSO, adopts a special U-shaped obstacle-avoid method, thus solves the problem of low network performance in the conventional neural-fuzzy system.

References

[1] Yupu Yang, Xiaoming Xu, Wenyuan Zhang. Design neural network based fuzzy logic[J]. Fuzzy Sets and System. 2000,114:325

[2] Chengjun Ding, Minglu Zhang, Ping Duan. Application of fuzzy neural network in the information crasis of mobile robot[J]. Control Theory and Application. 2004,21（1）:59-62

[3] Jun Sun,bin Feng,wenbo Xu. Particle Swarm Optimization with Particles Having Quantum Behavior[A]. IEEE int. Conf. on evolutionary computation[C]. Piscataway: IEEE ,2004. 325-331

[4] J.Kennedy, R.Eberhart. Particle Swarm Optimization[A].

IEEE int. Conf. on Neural Network[C]. IEEE,1995. 1942-1948

[5] Chen Wei, Bin Feng, Jun Sun. Simulation study on the Parameters Optimization of RBF-NN based on QPSO Computer Application .2006，26（8）：1928-1931

[6] Fang Bao, Yonghui Pan, Wenbo Xu. A Novel Training Algorithm for BP neural Network[A]. Proceedings of the International Symposium on distributed Computing and Application to Business, Engineering and Science[C]. Hangzhou:Shanghai University Press,2006. 767-770

[7] David Zhang, Sankar K. Pal. A Fuzzy Clustering Neural Network(FCNs) System Design Methodology[J]. IEEE Transaction on Neural Network. 2000, vol.11（5）, pp. 1174-1177

[8] Witold Pedrycz, George Vukovich. Logic-oriented Fuzzy Clustering[J]. Pattern Recognition Letters. 2002, vol. 23, pp.1515-1527

[9] J. Abonyi, F. Szeifert. Supervised Fuzzy for the Identification of Fuzzy Classifiers[J]. Pattern Recoginition Letters. 2003. 24(14). 2195-220

[10] A. Bensaid, L. O. Hall, J. C. Bezdek, L. P. Clarke. Partially Supervised Clustering for Image Segmentation[J]. Pattern Recognition. 1996. 29（5）. 859-871

Fuzzy Entropy of Vague Set and Its Construction Method *

Fengsheng Xu   Jianmin Gong   Haijun Li

Department of Computer Science and Technology, Dezhou University, Dezhou, Shandong, 253023, China

Email: xfs@dzu.edu.cn

Abstract

The drawbacks of the existing fuzzy entropy of vague sets are analyzed, the axiom definition of fuzzy entropy of vague sets is proposed, and construction method of fuzzy entropy of vague sets are given out.

Keywords: Vague sets; Fuzzy sets; membership function; fuzzy entropy

1 Introduction

In the fuzzy set theory introduced by Zadeh in 1965[1], each object $x \in U$ is assigned a single value between 0 and 1, called the grade of membership, where U is a universe of discourse. Formally, a membership function $\mu_F : U \to [0,1]$ is defined as a fuzzy set F, where $\mu_F(x)$, for each $x \in U$, denotes the degree of membership of x in the fuzzy set F. As pointed out by Gau *et al*.in [2], the drawback of using the single membership value in the fuzzy set theory is that the evidence for $x \in U$ and the evidence against $x \in U$ are in fact mingled together. They also pointed out that the single number reveals nothing about its accuracy. To tackle this problem, Gau *et al*.proposed the notion of *vague sets*, which allows using interval-based membership instead of point-based membership as in fuzzy sets. They used a truth-membership function t_V and a false-membership fuction f_V to characterize the lower bound on μ_V. These lower bounds are used to create a subinterval on [0,1], namely, $[t_V(x), 1-f_V(x)]$, to generalize the $\mu_V(x)$ of fuzzy sets,where $t_V(x) \le \mu_V(x) \le 1 - f_V(x)$. Since the theory of vague sets was introduced, many approaches for fuzzy entropy of vague sets have been proposed[3~10], which have discussed this issue from different points of view.But the drawbacks of these methods exist.

The drawbacks of the existing fuzzy entropy of vague sets[5~10] are analyzed, and the axiom definition of fuzzy entropy of vague sets is put forward, and construction methods of fuzzy entropy of vague sets are given out.

2 Vague Sets

Let U be a universe of discourse, with an element of U denoted by x.

Definition 1 A vague set V in U is characteriz ed by a truth-membership function t_V and a false-membership fuction f_V. Here $t_V(x)$ is a lower bound on the grade of membership of x derived from the evidence for x and $f_V(x)$ is a lower bound on the negation of x derived from the evidence against x. $t_V(x)$ and $f_V(x)$ both associate a real number in the interval [0,1] with each element in U, where $t_V(x) + f_V(x) \le 1$. Then

$t_V(x)$： $U \to$[0，1] and $f_V(x)$： $U \to$[0，1]

Suppose $U = \{x_1, x_2, \cdots, x_n\}$. A vague set V of the universe of discourse U can be represented by

$$V = \sum_{i=1}^{n} [t_V(x_i), 1 - f_V(x_i)] / x_i \qquad （1）$$

where $t_V(x) \le \mu_V(x) \le 1 - f_V(x)$ and $1 \le i \le n$.

This approach bounds the grade of membership of

* DSTB(NO.20070141)

x to a subinterval $[t_V(x),1-f_V(x)]$ of [0,1].In other words,the exact grade of membership $\mu_V(x)$ of x may be unknown,but is bounded by $t_V(x)\le\mu_V(x)\le 1-f_V(x)$, where $t_V(x)+f_V(x)\le 1$.

The uncertain degree about x is characterized by the difference $|t_V(x)-f_V(x)|$. The unknown degree about x is characterized by the difference m_A, where $m_x=1-t_V(x)-f_V(x)$. If it is small,the knowledge about x is relatively precise;if it is large,we know correspondingly little.if $t_V(x)$ is equal to $(1-f_V(x))$, the knowledge about x is exact, and the vague set theory reverts back to fuzzy set thory. if $t_V(x)$ and $(1-f_V(x))$ are both equal to 1 or 0, depending on whether x belong to V or not, the konwledge about x is very exact and the thory reverts back to ordinary sets.

Definition 2 The complement of a vague set A in U, denoted $\overline{A}$, is defined by $t_{\overline{A}}=f_A$ and $1-f_{\overline{A}}=1-t_A$.

Definition 3 A vague set A in U is contained in another vague set B in U, written as $A\subseteq B$, if and only if $t_A\le t_B$ and $1-f_A\le 1-t_B$.

Definition 4 Two vague sets A and B in U are equal, written as $A=B$, if and only if $t_A=t_B$ and $1-f_A=1-f_B$.

Definition 5 The union of two vague sets A and B in U is a vague set C, written as $A\cup B$, whose truth-membership and false-membership functions are related to those of A and B by $t_C=\max(t_A,t_B)$ and $1-f_C=\max(1-f_A,1-f_B)=1-\min(f_A,f_B)$.

Definition 6 The intersection of two vague sets A and B in U is a vague set C, written as $A\cap B$,whose truth-membership and false-membership functions are related to those of A and B by $t_C=\min(t_A,t_B)$ and $1-f_C=\min(1-f_A,1-f_B)=1-\max(f_A,f_B)$.

3 The Drawbacks of Existing Fuzzy Entropy of Vague Sets

The definitions of fuzzy entropy of vague sets in[5~10] are as follow:

$$VE_1(A)=\frac{1}{n}\sum_{i=1}^{n}\frac{1-|t_{x_i}-f_{x_i}|+m_{x_i}}{1+|t_{x_i}-f_{x_i}|+m_{x_i}} \qquad (2)$$

$$VE_2(A)=\frac{1}{n}\sum_{i=1}^{n}\frac{2t_{x_i}f_{x_i}+m_{x_i}^2}{t_{x_i}^2+f_{x_i}^2+m_{x_i}^2} \qquad (3)$$

$$VE_3(A)=1-\frac{1}{n}\sum_{i=1}^{n}|t_{x_i}^2-f_{x_i}^2| \qquad (4)$$

$$VE_4(A)=\frac{1}{n\ln 2}\sum_{i=1}^{n}(-\frac{t_{x_i}+1-f_{x_i}}{2}\ln\frac{t_{x_i}+1-f_{x_i}}{2}$$
$$-(1-\frac{t_{x_i}+1-f_{x_i}}{2})\ln(1-\frac{t_{x_i}+1-f_{x_i}}{2})) \qquad (5)$$

$$VE_5(A)=\frac{1}{n}\sum_{i=1}^{n}\frac{m_{x_i}+1-|t_{x_i}-f_{x_i}|(1+t_{x_i}+f_{x_i})/2}{m_{x_i}+1+|t_{x_i}-f_{x_i}|(1+t_{x_i}+f_{x_i})/2} \qquad (6)$$

$$VE_6(A)=\frac{1}{n}\sum_{i=1}^{n}\frac{m_{x_i}+1-|t_{x_i}{}^2-f_{x_i}{}^2|}{m_{x_i}+1+|t_{x_i}{}^2-f_{x_i}{}^2|} \qquad (7)$$

The drawback of above fuzzy entropy of vague sets are explained by following example.

Example 1 let $A=[0.2,0.8]/x_1$ and $B=[0.5,0.5]/x_1$, according to Eq.（1）~ Eq.（6）, we can obtain fuzzy entropy of vague sets A and B are 1.

Above result is explained in the voting model, two persona are for and two persons are against and six persons remains neutra in A, but five persons are for and five persons are against and no one remains neutra in B. Obviously, each people has his own opinion in B, either for or against, but some remain neutra in A, fuzzy property of A is bigger than fuzzy property of B, based on intuition of humanbeing. However, it is wrong that fuzzy property of A is equal to fuzzy property of B in the above result.

Let $x=[t_x,1-f_x]$, if $t_x=f_x$, according to Eq.（2）~ Eq.（7）, its fuzzy property is 1. Obviously, these methods don't think that fuzzy entropy of vague sets is affected by unknown degree. Based on intuition of humanbeing, if $t_x=f_x$, then fuzzy entropy of vague sets will increase with the rise of unknown of vague sets.

4 The Axiom Definition of Fuzzy Entropy of Vague Sets and Its Construction Method

In order to overcome the drawbacks of the existing fuzzy entropy of vague sets, the axiom definition of fuzzy entropy of vague sets is as follow:

Definition 7 VE: $VS(U)\to$[0, 1] is called fuzzy entropy of vague sets, if it is satisfied with the following conditions:

（1）$VE(x)=0$, if and only if $x=[0,0]$ or $x=[1,1]$

（2）$VE(x)=1$, if and only if $x=[0,1]$

（3）$VE(x)=VE(\bar{x})$

（4）Let $x=[t_x,1-f_x]$, $y=[t_y,1-f_y]$, and $|t_x-f_x|=|t_y-f_y|\neq 0$, if $\pi_x>\pi_y$, then $VE(x)>VE(y)$.

（5）Let $x=[t_x,1-f_x]$, $y=[t_y,f_y]$, and $\pi_x=\pi_y$, if $|t_x-f_x|>|t_y-f_y|>0$, then $VE(x)<VE(y)$.

The above definition not only consider the influence of unknow degree to fuzzy entropy of vague sets, but also consider the influence of uncertain degree to fuzzy entropy of vague sets, and fuzzy entropy of vague sets is increasing fuction of π_x and decreasing fuction of $|t_x-f_x|$, and its fuzzy entropy is Maximum when it is completely unknow.

We construct two konds of fuzzy entropy of vague value and vague sets in the following：

$$VE_7(x)=\frac{1-|t_x-f_x|+m_x}{2}$$

$$VE_7(A)=\sum_{i=1}^{n}\frac{1-|t_{x_i}-f_{x_i}|+m_{x_i}}{2} \qquad （8）$$

$$VE_8(x)=a(1-|t_x-f_x|)+bm_x$$

$$VE_8(A)=\sum_{i=1}^{n}(a(1-|t_{x_i}-f_{x_i}|)+bm_{x_i}) \qquad （9）$$

where $a+b=1$, $a,b>0$.

Eq.（8）is special case of Eq.（9）, the influence of unknown degree to fuzzy entropy of vague sets is the same as the influence of uncertain e degree to fuzzy entropy of vague sets in Eq.（8）. The influence of unknown degree to fuzzy entropy of vague sets is b, the influence of uncertain e degree to fuzzy entropy of vague sets is a in Eq.（8）, b and a can be determined by objective reality, its flexibility is well.

Theorem 1 $VE_7(x)$ and $VE_8(x)$ are fuzzy entropy of vague value x which are satisfied definition 5.

Proof （1）$VE_7(x)=0$, if and only if $1-|t_x-f_x|+m_x=0$, if and only if $2-|t_x-f_x|-(t_x+f_x)=0$, if and only if $x=[0,0]$ or $x=[1,1]$.

$VE_8(x)=0$, if and only if $a(1-|t_x-f_x|)+bm_x=0$, if and only if $1-|t_x-f_x|$=0∧ m_x=0, if and only if $x=$[0, 0] or $x=$[1, 1].

（2）$VE_7(x)=1$, if and only if $1-|t_x-f_x|+m_x=2$, if and only if $|t_x-f_x|=0 \wedge m_x=1$, if and only if $x=[0,1]$.

$VE_8(x)=1$, if and only if $a(1-|t_x-f_x|)+bm_x=1d$, if and only if $a|t_x-f_x|+b(t_x+f_x)=0$, if and only if $|t_x-f_x|=0\wedge t_x+f_x=0$, if and only if $x=[0,1]$.

（3）It can be proved by its symmetry.

（4）Let $x=[t_x,1-f_x]$ and $y=[t_y,1-f_y]$, $|t_x-f_x|=|t_y-f_y|\neq 0$, if $m_x>m_y$, then

$$VE_7(x)-VE_7(y)=\frac{1-|t_x-f_x|+m_x}{2}-\frac{1-|t_y-f_y|+m_y}{2}=\frac{m_x-m_y}{2}>0$$

, hence $VE_7(x)>VE_7(y)$.

$$VE_8(x)-VE_8(y)=a(1-|t_x-f_x|)+bm_x-a(1-|t_y-f_y|)-bm_y$$
$$=b(m_x-m_y)>0$$

, hence $VE_8(x)>VE_8(y)$.

（5）Let $x=[t_x,1-f_x]$ and $y=[t_y,f_y]$, $m_x=m_y$, if $|t_x-f_x|>|t_y-f_y|>0$, then

$$VE_7(x)-VE_7(y)=\frac{1-|t_x-f_x|+m_x}{2}-\frac{1-|t_y-f_y|+m_y}{2}=\frac{|t_y-f_y|-|t_x-f_x|}{2}<0$$

, hence $VE_7(x)<VE_7(y)$

$$VE_8(x)-VE_8(y)=a(1-|t_x-f_x|)+bm_x-a(1-|t_y-f_y|)-bm_y$$
$$=a(|t_y-f_y|-|t_x-f_x|)<0$$

hence $VE_8(x)<VE_8(y)$

Theorem 2 $VE_7(A)$ and $VE_8(A)$ are fuzzy entropy of vague set A which are satisfied definition 5.

Example 2 let $A=[0.2,0.8]$ and $B=[0.5,0.5]$, then $\pi_A<\pi_B$. Based on the method in this paper, we can obtain:

$VE_7(A)=0.8$, $VE_7(B)=0.5$

$VE_8(A)=0.96$, $VE_8(B)=0.9$, where $a=0.9$, $b=0.1$

it can be seen that methods in this paper not only overcomes the drawback of existing fuzzy entropy of vague sets, but also are very reasonable.

5 Conclusion

The drawbacks of existing fuzzy entropy of vague sets in [5~10] are analyzed, and the axiom definition of fuzzy entropy of vague sets is put forward, and construction methods of fuzzy entropy of vague sets are given out. Methods in this paper not only overcomes the drawback of existing fuzzy entropy of vague sets, but also are very reasonable.

Acknowledgment

Financial supports from Dezhou science and technology bureau are highly appreciated. The helpful comments from two anonymous reviewers are also gratefully acknowledged.

References

[1] Zadeh L A, “Fuzzy sets”, Inform and Control, Vol.8, No.3, 1965, pp.338~356

[2] Wen-Lung Gau, Daniel J Buehrer, “Vague sets”, IEEE Transaction on System, Man and Cyberbetics, Vol.23, N0.2, 1993, pp.610~614

[3] Burillo P, Bustince.H, “Entropy on intuitionistic fuzzy sets and on interval_valued fuzzy sets”, .Fuzzy sets and System, No.78, 1996, pp.305~316

[4] Eulalia Szmidt, Janusz Kacprzyk, “Entropy for intuitionistic Fuzzy sets”, .Fuzzy sets and Systems, No.118, 2001, pp.467~477

[5] Liu Yun-Sheng, Huang Guo-Shun, “Entropy for vague sets”, Mini-Micro Systems, Vol.27, No.11, 2006, pp.2115-2119

[6] Vlachos Ioannis K, Sergiadis George D, “Inner product based entropy in the intuitionistic fuzzy setting”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol.14, No.3, 2006, pp.351-366

[7] Huang Guo-sun, Liu Yun-sheng, “Similarity Measure of Vague Sets Based on Entropy”, Journal of Chinese Computer Systems, Vol.29, No.1, 2008, pp.139-143(in chinese)

[8] Huang Guo-sun, Liu Yun-sheng, “On the Fuzzy Entropy of Vague Sets”, Computer Enginerring and Applications, Vol.41, No.33, 2005, pp.48-50(in chinese)

[9] Li Tian-zhi, Liang Jia-rong, Fan ping, Gong Jian-min, “Fuzzy Entropy of Vague Sets”, Application Research of Coputers, Vol.24, No.10, 2007, pp.93-95（in chinese）

[10] Fan ping, Liang Jia-rong, Li Tian-zhi, “New Fuzzy Entropy of Vague Sets”, Computer Enginerring and Applications, Vol.43, No.13, 2007, pp.179-181(in chinese)

Radial Basis Function Neural Network Based on Ant Colony Clustering

Peng Tian[1] Xianfang Wang[1,2] Zhou Wu[1] Feng Pan[1]

1 School of Information & Control Engineering, Jiangnan University, Wuxi, Jiangsu 214122, P.R.China

2 Henan Institute of Science and Technology, Xinxiang, Henan 453003, P.R.China

Email: tianpeng325@yahoo.com.cn; wangxianfang@sina.com; wuzhou2118@yahoo.com.cn

Abstract

By introducing ant colony clustering algorithm into neural network of radial basis function for optimally clustering N objects into K clusters. Based on ant colony algorithm and its feature of parallel search optimum, and employs the global pheromone updating and the heuristic information to construct clustering solution. The rate of clustering is accelerated by increasing the utilization of the pheromone. The heuristic information is applied to improve the efficiency of the algorithm. Uniform crossover operator is used to further improve solutions discovered by ants. The experimental results demonstrate that the very encouraging results in terms of the quality of solution found and the precision to RBF-ACC is improved evidently.

Keywords: Ant Colony Algorithm; Clustering; Radial Basis Function Neural Network; Optimization; Uniform Crossover

1 Introduction

Due to the speed of convergence compares quickly, is the classical forward neural network Radial basis function neural network(RBFNN),has the extensively applied foreground in the engineering [1].It is the single and implicit layer of a kind of two forward layer to network, the transforming from the importation space to the implicit layer space is nonlinear. The transforming function of the implicit cell is RBF[2,3,4]. The function of that model mainly is decided by training samples and then using arbitrarily complicated network structure to the training sample to carry on the training, to get an arbitrarily small mistake rate. The network will lose the ability of expand or hardly attain the perfect results, owning to the great deal of training samples. Before dividing the samples to different kinds, they should be gathered to be analysis. According to the type of sample data to organize the subnet network of RBF, studying the regulations which is not discover. The training time of the network will be cut down, the expansion ability will be raised and the ability of approach [5] will be improved evidently.

Based on consideration of those, the novel radial basis function neural work based on ant colony clustering (RBF-ACC) was proposed[6,7,8]. ACC has strongly ability of searching the global optimization. Not only do it make the clustering more faster, but also do the clustering center to be more representation, applies it in the RBFNN, may promote the network greatly robustness.

2 Ant Colony Clustering

ACC simulates real ant's cooperation process which is completed by many ants, each ant searches the solution independently in the candidate solution space, and then leave behind certain information content on the solution[9,10,11,12]. If the solution has too much pheromone, it will be selected more often. ACC is produced by this influence. ACC is a kind of global optimization heuristic algorithm, what based on is the information element and the track theory seeks for the explanation which to the ant food source the behavior makes. In the original state, the ant choose a path stochastically, during the same time, more pheromone will be accumulated in the short route. Therefore, the subsequent ant will favor in chooses this way which will lead to the pheromone becoming more and more. According to the pheromone of clustering center, the ant put the periphery data together, for the purpose of obtaining data species.

2.1 The basic philosophy of ACC algorithm

Supposes the data sample is contains N data $\{x_1, x_2, \cdots, x_N\}$, Each sample point expression is $x_k = \{x_{k1}, x_{k2}, \cdots, x_{kn}\}$ .The clustering is divided the N objects into K clusters.

Defines the following objective function:

$$\min F(w,m) = \sum_{j=1}^{K} \sum_{i=1}^{N} \sum_{v=1}^{n} w_{ij} \left\| x_{iv} - m_{jv} \right\|^{2} \qquad （1）$$

Satisfied

$$\sum_{j=1}^{K} w_{ij} = 1, i = 1, \ldots, N \qquad （2）$$

$$\sum_{i=1}^{N} w_{ij} \geq 1, j = 1, \ldots, K \qquad （3）$$

In the formula, x_{iv} is object i v th is; w is $N \times K$ dimension matrix, the element

$$w_{ij} = \begin{cases} 1, & if\ x_i \in clusterj \\ 0, & if\ x_i \notin clusterj \end{cases} \qquad （4）$$

m is a $K \times n$ cluster center matrix, m_{jv} expression the mean of all samples' attribute v in class J , The formula is as follows:

$$m_{jv} = \frac{\sum_{i=1}^{N} w_{ij} x_{iv}}{\sum_{i=1}^{N} w_{ij}}, j = 1, \ldots K, v = 1, \ldots n \qquad （5）$$

In ACC, construct the solution by R artificial ant to. Each ant constructs the string $S = \{c_1, c_2, \ldots, c_N\}$ by the length of N, and $\{c_i \mid i = 1, 2, \ldots, N\}$ is the sign of the object of i , $c_i \in 1, 2, \ldots K$. The object of x_i , x_j belongs to the same kind if $c_i = c_j$. Otherwise they are not in the same kind. For example, the clustering problem contains $N = 8$ $K = 3$. Supposing the solution is $S = (2,1,3,2,2,3,2,1)$ which means the first object belongs to the kind of 2nd, the second object belongs to the kind of 1st, and so on.

At the beginning of the algorithm, initializing the pheromone matrix τ with N*K dimension, and make τ_0 as its value. The element τ_{ij} represents that the object i is opposite of pheromone density of class j . Through every turn, based on the pheromone matrix constructs solution by each artificial ant, further more it uses the crossover operator to improve the quality of solution, and then updates pheromone matrix. Therefore, in the direction of unceasingly renews the pheromone matrix, the ant improves the quality of solution more and more, until the iterative times is over. In order to explain the process of algorithm, we put N=8, k=3 in the clustering problem.

2.2 The structure of solution

In ACC, the ant structures solves S by using pseudo-random scale factoring rule （6）, Assigns a kind of marking for S in each element, namely to the ant which is located at the element, enables by the probability $\tau_{ij} \square \left[\eta_{ij}\right]^{\beta}$ choice to achieve biggest kind of S to take this element the kind; based on the probability with $1 - q_0$ and according to （7）to assign one kind as the element i ,In the algorithm, ant's pseudo-random proportion formula is

$$s = \begin{cases} \arg\max\limits_{j \in \bar{K}} \left\{ \tau_{ij} \square \left[\eta_{ij}\right]^{\beta} \right\}, if & q \leq q_0 \\ J, & otherwise \end{cases} \qquad （6）$$

In the formula , q for uniform distribution random number in $[0,1]$; q_0 is a Constant $(0 < q_0 < 1)$; $\eta_{ij} = 1/d_{ij}$ represents the heuristic information value; In the formula d_{ij} expresses the center's distance between object i and the kind of j , the whole heuristic information is a $N \times K$ matrix; the heuristic factor is β , expresses the relative importance of the heuristic information; J is the random variable which is decided by type （7）.

$$p_{ij} = \frac{\tau_{ij}\left[\eta_{ij}\right]^{\beta}}{\sum_{k=1}^{K}\tau_{ik}\left[\eta_{ik}\right]^{\beta}}, j = 1,\ldots,K \qquad （7）$$

Here p_{ij} reflects the distributed probability of the object i belonging to class j.

In rule （6）, the first process is used to develop the existing knowledge, and the second process is made use of exploring the new solution space. In order to explain the process, making $q_0 = 0.8$ and consider the result is $S = (2,1,3,2,2,3,2,1)$. At first, producing a group of uniform distributions stochastically vector(0.693254, 0.791554,0.986554,0.988556,0.245698,0.968854,0.091458,0.348956).Selecting the product $\tau_{ij}\square\left[\eta_{ij}\right]^{\beta}$ for greatest kind of achievement, owning to the correspondence's random numbers 1st, 2nd, 5th, 7th and 8th are smaller than q_0; On the contrary, the element 3, 4 and 6 correspondence's random numbers are bigger than q_0. Thus, this element is set to be species by the probability of p_{ij}.

2.3 Crossover operator

Put the partial search into article algorithm, for the sake of raising the approximate solution efficiency which in the algorithm. Especially when the question territory's inspiration information was not obtained easily, joins the partial search to be more possible to find a better solutiono. At present, the partial search implements to all solutions, except, it also can be implemented by the part of results. This article only implements the partial search to the best 2O% feasible solution. Before the partial search, all solutions are carried on the rising foreword arrangement, according to the goal function value.

The partial search operation has many kinds, is distinguish with traditional spot crossover or two spots overlapping, In order to have the new solution, we operate parameter uniform crossover [13]. Uniform crossover is more generalization. The potential crossover spot are made by each spot. The random number variable has the same length with unit that is produced stochastically. Using p_{ls} as threshold value to judge that which father provide the variable value to sub-individual.

First , the number in [0,1] is stochastic producing, if in relevant position's random number is smaller than the threshold value probability, then the father body 1 provides this element; Otherwise the father body 2 provide the element. It is different from the tradition the gene variation which carries on by the very small probability; this process can produce more solutions stochastically, preventing the population precocious convergence.

2.4 The global pheromone updating

The pheromone renews can dynamic respond that the ant individual the information which produces in the movement. Therefore, the pheromone matrix will be renewal, using the global optimal solution, if the ant completes one search.The information element renewal formula is as follows:

$$\tau_{ij}\left(t+1\right) = \left(1-\rho\right)\tau_{ij}\left(t\right) + \rho\Delta\tau_{ij}^{bs} \qquad （8）$$

$$\Delta\tau_{ij}^{bs} = \begin{cases} 1/F_{bs}, & \text{If in the optimal solution the ith element belongs to kind of j} \\ 0, & otherwise \end{cases}$$

（9）

F_{bs} expresses optimal solution goal function value; $\rho(0 < \rho < 1)$ is the whole pheromone volatilization coefficient, ρ is more bigger, the pheromone volatilizing becoming more quicker.

2.5 RBFNN based on ant colony clustering

A novel radial basis function neural network based on ant colony clustering (RBF-ACC) which has the global optimization cluster results enables the training process to be improved.

3 The Result of Experiment

In order to confirm the validity that is based on RBF-ACC neural network training, carrying on the simulation to RBF as well as RBF-ACC.

We choice the glutanic acid fermentative process, put time, cell concentration, glutanic acid density, residual sugar density and so on as the input. And then make the glutanic acid density, the residual sugar density as the output. The cluster center choice produces stochastically, carries on the simulation to the RBF neural network, the result is as follows:

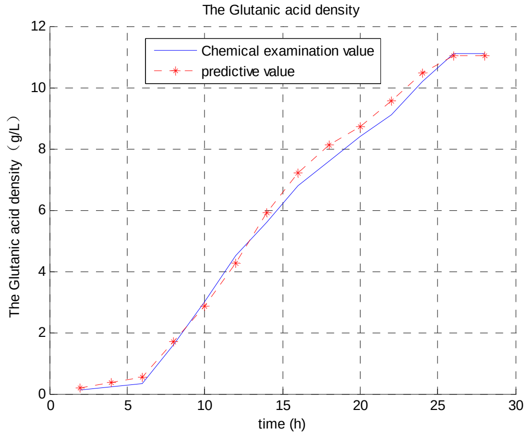


Figure 1 The Glutanic acid density

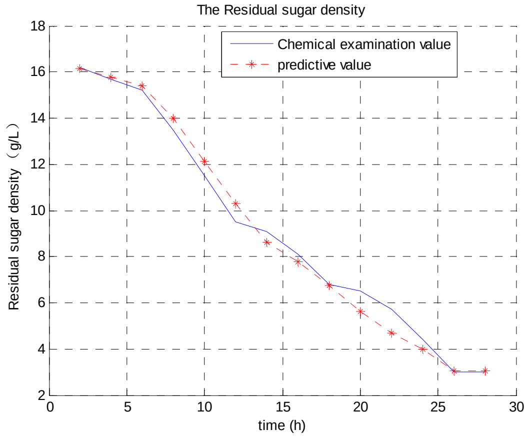


Figure 2 Residual sugar density

The application based on RBF-ACC, put dissolves oxygen, temperature, PH, rotational speed and so on as the input. And then make the glutanic acid density, the residual sugar density as the output. The number of clustering K=3, the ant group counts R=10. The biggest iterative times NC=1000, the global optimization pheromone volatilizing coefficient ρ =0.1, the result is as follows:

Through the simulation comparison and analysis, proved that is the RBF-ACC has the high test collection classification rate of accuracy, the training regulations repayment rate ,besides ,it also has the small training error. So it causes the training result to be more ideal.

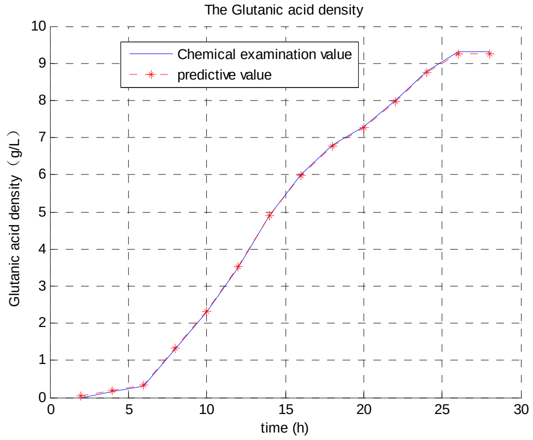


Figure 3 The Glutanic acid density

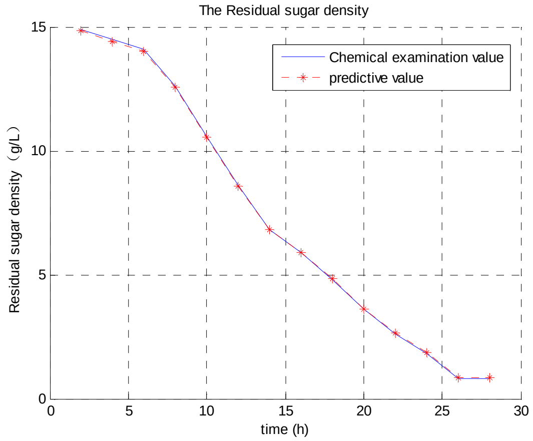


Figure 4 Residual sugar density

4 Conclusion

This article proposed RBF-ACC is a kind of highly effective learning algorithm, this algorithm fully using the characteristic that the ant looking for food and making use of ACC to seek for the central point of the primary function .It provides a useful method to solve the difficulty problem that is the RBF neural network solution center vector. At the same time, it also optimized the training speed, the training precision for RBFNN. In a word, RBFNN will be more popular in application field.

References

[1] Lan Flood, Nabil Katram. “Neural networks in civil engineering”, Principles and understanding[J]. Comp Civ Engrg ASCE, 1994, 8（2）, pp.131-148

[2] Zhong Wei, Yu Jinshou. The mirosoft sensor based on RBFNN for production qualities estimation of hydrogenation cracking fractionation tower[J]. Automation in Petro-Chemical Industry,1999,（5）, pp.19-21

[3] Lei Xu, Adam Krzyzak, Comptitive learning for clustering analysis, RBF net and curve detection. IEEE Trans. on neural networks, 1993,4（4）, pp.636-649

[4] S.chen,Cowan, P.M. Grant. Squares learning algorithm for radial basis function networks. IEEE Trans. on neural networks, 1991,2（2）, pp.302-309

[5] Guoyong, Li. “The intelligent control and matlab realize” [M]. Beijing: Publishing House of Electronics Industry, 2005, pp.33-37

[6] Hathway R J,Bczdek J C. Optimization of clustering criteria by reformulation[J]. IEEE transactions on Fuzzy Systems, 1995, 3（2）, pp.241-245

[7] M Dorigo, V Maniczzo, A Colorni. The ant system: Optimization by a colony of cooperating agents[J]. IEEE Transactions on Systems, Man, and Cybernetics Part B, 1996, 26（1）, pp.29-41

[8] Lumer E,Faieta B. Diversity and adaption in populations of clustering ants[C]. In: Proc of the Third International Conference on Simulation of Adaptive Behavior: From Animals 3, Cambrige, MA: MIT Press, 1994, pp.501-508

[9] Colomi A, Dorigo M, Maniezzo V. “An investigation of some properties of an ant algorithm”[A]. Proceedings of the Parallel Problem Solving from Nature Conference[C]. Brussels, Belgium: Elsevier Publishing, 1992, pp.509-520

[10] Wu Q H, Zhang J H, Xu X H. “An ant colony algorithm with mutation features”[J]．Journal of Computer Research and Development, 1999, 36（10）, pp.1240-1245

[11] Shelokar P S. Jayaraman V K, Kulkarni B D. “An ant colony approach for clustering”[J]. Analytica Chimica Acta, 2004, 509,pp.187-95

[12] B Wu,Y Zheng, S Liu et al.SIM: A Document Clustering Algorithm Based on Swarm Intelligence[C]. In: Proc of the IEEE World Congress on Computational Intelligence, Hawaiian, 2002, pp.477-482

[13] Spears W M. De Jong K A. “On the virtues of parameterized uniform crossover”[A]. Proceedings of the Fourth International Conference on Genetic Algorithms.1991,pp.230-236

Realizable Fuzzy Petri Net of Feedback*

Lingdong Kong[1, 2] Hong Zhang[3] Lifang Kong[1]

1 School of Environment and Spatial Informatics, China University of Mining and Technology
Xuzhou, Jiangsu 221008, China

2 Department of Software Engineering, Yancheng Institute of Technology
Yancheng, Jiangsu 224003, China

3 School of Computers Science and Technology, China University of Mining and Technology
Xuzhou, Jiangsu 221008, China
Email: njkld@163.com

Abstract

Reliability was believed as the first basic standpoint of the system model of Petri net. In this paper, the state components and transition components of Fuzzy Petri net were expanded locally based on analysis of Petri Net fuzzy features, it not only expresses fuzzy knowledge but also was realized with facility using program to reason. Combining with characteristics of knowledge expressions in gas outburst prediction, a reasoning mechanism of fuzzy Petri net was given and a fuzzy Petri net of feedback that can be realized was put forward, at the same time, it was given a detailed description about algorithms analysis.

Keywords: Fuzzy Knowledge Expression, Fuzzy Petri Net, Reliability, Feedback, Algorithms Analysis

1 Introduction

Fuzzy Petri Net (FPN) is the combination of Petri Net and knowledge representation; it was first used to describe the fuzzy production rules. FPN supports the necessary causal relationship of many - to - one and many - to - many, has proven itself as a practical and important inference and learning tool [1]. Combined with the author’s experience accumulation in Software Engineering teaching and software development and referred to the data flow diagram (DFD) in Software Engineering and the state transition diagram advantages that make improvements, at the same time combined with the potential knowledge characteristics of uncertainty and fuzzy in the study of gas outburst prediction, a realized feedback Fuzzy Petri Net was given.

2 Fuzzy Petri Net Knowledge Analysis

Different literatures give different definitions of Fuzzy Petri Net; different definitions are often based on different applications as background. This paper mainly refers to Chongyi Yuan[1], Zhibin Jiang[2], Xingui He[3], Zhehui Wu[4], and other [5-9]related expositions to have a comparative analysis on the fuzzy consisting of a Petri network element. Combined the ambiguity of knowledge in the gas outburst prediction and uncertainty background, gave the definition of fuzzy Petri net.

2.1 Possibility of fuzzy knowledge expression of Petri Net

Petri Net originated in the description of the signal transmission, so it adapts to describe the system which has the characteristics of resource flow. The model of the system has two elements: the state element and the change element. The change of systems is the qualitative

* Graduate Innovation Program of Jiangsu Province of China under Grant No.CX07B_139Z; the Natural Science Foundation of Jiangsu Province of China under Grant No.BK2005021.

description of natural rules. It only gives whether there is a relation between the resources and transitions and its modes (consume or produce), and describes the relations among transitions on the dependence of nature. The systematic behavior of the Petri Net was expressed as resources mobile.

It is natural that our description of knowledge and information system can be used to express the impact factors and variables as the state, and the dependent relationship of cause and effect can be described by transitions. More general we can understand that things may changes in the status of different factors, the impact of such factors may be isolated, and the more likely the combined, then the impact of each of the factors on the results all has its own degrees. In considering the integrated multi-dimensional factors, the weighted combination judgments methods are accustomed to be used.

Petri net is reflected in the inter-dependence on such things, the whole world is composed of such tangible and intangible networks [2]. All things are changing, affecting and relying on. This also gives us the thinking way to use Petri Net to express knowledge and reasons. Thus, it can be certain that it is not only feasible but is the natural habit of the people's thinking that Petri Net for knowledge.

2.2 Petri net fuzzy analysis

Fuzzy Petri net is the fuzzy combination of ordinary Petri Net with fuzzy knowledge expression. The ordinary Petri net is mainly composed of three ingredients: state, change and connecting arc. Therefore, Fuzzy is conducted on the basis of these three ingredients, and also lays particular stress on them. At the same time, the degree of fuzzy is also different. The following was selected from literature [2-4] on the three components of the fuzzy configuration of Fuzzy Petri Net with integrated analysis.

1) State (position) node is corresponding to the credibility of fuzzy knowledge: in the literature [3] the makers of nodes can be any real number which is in the ordinary Petri Net a natural extension, and marked by the number of markers circle. And in the literature [2]-[4] the markers were vaguely restricted between [0, 1].

2) Transition (transfer) node corresponds with the threshold of the fuzzy knowledge state transitions: in literature [3-4] there is a contact transfer threshold limit and restriction in between [0, 1]. In literature [3] the transition node was expressed in vertical line with threshold. In literature [4] it was expressed by small boxes and the threshold was written in them. In literature [2] the threshold concept that given in advance was used but not clearly reflected in the net. But the confidence factor was marked directly on changes. The author thinks it very clear that the expression of a single affecting factors. But the performance of multidimensional factors in the expression of multiple output or outcome of is inadequate; the author of this paper mainly lays emphasis on multi-dimensional non-linear factors.

3) Connecting arc (including input and output connections) corresponds with the confidence factor of the fuzzy knowledge transitions connect (or related): here it is necessary to stress the concept of input and output connections. There are two connections in Petri Net: input connection refers to the connection from the position node to the transfer node; output connection refers to the connection from the transfer node to the position node. A Petri net is the chart of using the interconnection of this two-node and two-connection. In literature [3-4] the input connections and output connections in the fuzzy Petri Net all have connected intensity; they were showed by the side (arrow) that indicates the connected intensity. But in literature [2] there is a clear different, omitting connected intensity, directly show the confidence factors on the transfers which was expressed using vertical lines.

4) Activation (ignition, start) and Reasoning knowledge of Fuzzy Petri Net should be relative: This is the essence of fuzzy Petri Net. In the ordinary Petri Net, only when all the contacts of a transfer node at least has one identification it can be ignited, the result of ignition is to reduce a marker from its every input node and increase a marker in its all output nodes. By such a step by step operation, some transfer points ignited

constantly and the nodes number of some poison contacts are constantly changing. So Petri net is very suitable to describe concurrent or parallel behavior and the operational various systems. With the expression of fuzzy knowledge it is also suitable for a variety of factors (or state) impact of a combination association changes.

In the literature [3-4] because of the using of the incident intensity concept, the corresponding input connected intensity are all expressed as a binary function, that is, the Reliability and the Confidence Factor of transitions are weighted. Only when the minimum of the various input intensity of this transfer contact bigger than or equal to the threshold of this transfer contact, the system model can be ignited (activated). In literature [2] although there is no clear explanation of incident intensity the ignition mechanism is basically the same, that is, the reliability of the state and the confidence factor of the transition are weighted, but only the expression of different ways.

With comprehensive comparison they all have different expansions on the three main components of Petri Net. The fuzzy state and the transition identification are restricted between [0,1], then it is more convenient to operate and correspond in the fuzzy inference procedure and implement the tri-value principles in the process of the computer implementation, at the same time the symbols are different; the transfer transition of state are all based on the weighted Reliability and Confidence Factor and also contrast with threshold to achieve, and facilitate the realization of the inference based on the production rule; in the knowledge representation and inference procedure the two important and fuzzy index plays a decisive role in the dynamic behavior (inference) of Fuzzy Petri Net: the incident function of input intensity and the ignition threshold of each transfer node.

2.3 Characteristics of knowledge expressions in gas outburst prediction

Gas Outburst precursor analysis: Audio omen: commonly known as coal banger, usually has low thunder acoustic (sound of artillery release) in deep coal bed, crack (shorts), splitting, brouhaha, rustle and so on. Silent omen: coal becomes softening, shine darkles, drop bits and small blocks flake away, coal has a slight tremble, the plank pressure increases, methane emission increases or suddenly high or low, coal surface temperature and air temperature decrease[10].

Here the author only makes it as an example to illustrate the fuzzy characteristics of the gas outburst prediction knowledge expressions in order to the description examples of the following algorithm.

The main two comprehensive targets of judging the coal outburst are D and K, the confirmation of them is given by the Institute of Fushun, its calculating formula is the fatalness index of coal outburst and its calculating formula as shown in Table 1.

Table 1 Target of Outburst and Calculating Formula

D	K	Outburst dangerous	Calculating formula
<2.5	-	Not dangerous	$D=\left(\frac{0.0075H}{f}-3\right)(P-7.5)$ $K=\frac{\Delta P}{f}$
⩾2.5	and<15	Not dangerous	
⩾2.5	and⩾15	dangerous	

From the Table.1, we can see that the description of the outburst dangerous is obviously fuzzy. On one hand under different targets the degree description of the not dangerous is fuzzy, on the other hand coal bed outburst fatalness comprehensive index and the coal's outstanding fatalness comprehensive index are all fuzzy within a certain scope. With a further research we can indicate the vertical deep from the earth surface, coal's consistence modulus and gas pressure etc. through fuzzy knowledge directly.

Further various influence factors or targets all have an inevitable incident relationship, such as the existence of the outburst danger often in the condition of the coal bed outburst danger comprehensive index D⩾2.5 and coal outburst danger comprehensive index K⩾15.In the data of daily observation process the degree of the danger can be determined preliminarily, and the corresponding measures can be adopted. Moreover, with the continuous expansion of the depth the changes of various kinds of index information are different, and that the information flow is formatted naturally. These are advantages of the Petri Net description, it will be further expatiated in the following knowledge representation and reasoning.

3 Fuzzy Knowledge Expressions under Software Realization

Reliability is the first basic viewpoint of the system model of Petri Net, and knowledge expressions always the hot research topics of knowledge processing, but it is accompanied with the acquisition and application of knowledge, the straightforward and simple indication knowledge description of Petri Net is one of its main strong points. However, the discussion based on the acquisition and inference of Fuzzy Petri knowledge are very few, and the author has already begun to explore the realization of the workflow based on the object-oriented Petri net [6].

Combined the characteristics of Fuzzy Petri Net with the author's experiences in the Software Engineering teaching and software development and referred to the data flow diagram (DFD) in Software Engineering and the state transition diagram advantages that make improvements.

As previously noted, at the same time with our study of the knowledge characteristics of the gas outburst prediction and the already started work of fuzzy Petri Net workflow mining development and realization, the author gives the following expansion and expression of Fuzzy Petri Net creatively.

The starting point is the author's puzzle on Petri Net for over the years: good technical, theoretical, tight mathematical proof, but application and the realization were very rare in reports, with a view to achieve the realizability groping under the guidance of Petri net theory. For the convenience of the knowledge indication in the following text, the author makes a unified definition of the main expression: Reliability-R, Reliability-R, and Threshold-TH.

3.1 Expansion of component composition

1) Expansion of the state components: In Petri theory it is indicated by a circle with a Token, it has very intuitionistic features and the initial indication and description of knowledge are also straightforward. The expression of one state is very clear, but in the state of the ever-increasing number of state, the mark of this state began dazzling, and this is also the reasons that the stratification emerges and the object-oriented simplify in the process of Petri Net modeling, which can reduce the complexity of modeling. We further refer to the software life-cycle thinking in software engineering and think over the mark of the state point from the point of view of analysis, design, implementation and safeguarding, then just a circle that there are major limitations. In the design, we hope to achieve a smooth transition from analysis to design, while software testing and maintenance can be carried out according to demand (described) to verify conveniently. So with UML and the data flow diagram do the following expansion, in Figure 1 (a) is indicated by a routine marked Petri net, (b) makes the mark of the library transfer into a circle and also divides the circle into two parts, the above half of which is used to indicate the name of the library and the below half of which is used to indicate the mark. In this (c) uses the Statement (S) to replace the name of the library with the further combination with the characteristics of the Fuzzy Petri Net knowledge, the marks was substituted by the Reliability (R) of Statement Factors (S) and the scope of the R all in the circle. On one hand this overcomes the shortcoming of the relatively disordered situation with more state factor data, on the other hand the indication of fuzzy knowledge was settled beautifully and at the same time it laid the foundation for Petri net modeling program.

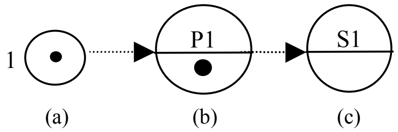


Figure 1 Expansion of the state components

2) Expansion of the transition components: Taking the above fuzzy Petri net analysis, the transition elements need marked threshold. Ordinary Petri Net as long as all the input at least has one Token then

transition can take place. Generally the threshold do not require labeling, but the introduction of Fuzzy Petri Net, threshold as an important analysis index will appear frequently, and therefore it needs improvement on how reasonable and clear to indicate the name of the transition, threshold and confidence factor. As shown in Figure 2, (a) is an ordinary Petri net Transition elements, (b) is indicated by the same recommended rectangular box. But the box is divided into two parts, the first half of it is the transition logo and the second part refers to the threshold of the transition, which was in order to adapt to the knowledge expression and learning. It will be illustrated in the following detailed examples. In Figure 2 (c) gives a transition or the threshold of rules.

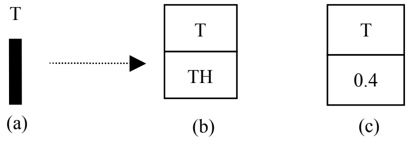


Figure 2 Expansion of the transition components

3) Amendment of the connecting arc weight: on the basis of the expansion of state and transition components and combined with the expression of fuzzy knowledge and also on the basis of the ordinary labeling Petri net weight Confidence Factor (CF) was used to indicate weight and the connected intensity of arc. In Figure 3 on the premise that the credibility of the state factor S1 is R1, its Confidence Factor is CF1. According to the direction of the arrowhead we can see that the Confidence Factor is CF1 under the situation that it is bigger than the threshold TH1, the Reliability of the rules of conversion is R2.

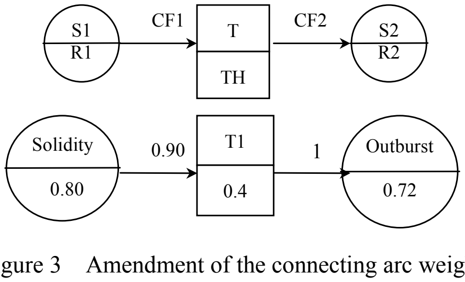


Figure 3 Amendment of the connecting arc weight

3.2 Algorithms Analysis

3.2.1 Related definitions

On the basis of the expansion of the relevant components the definition of the Fuzzy Petri Net and the reasoning algorithm analysis are given.

Definition 1 Fuzzy Petri Net has a six-tuple

FPN = (S, T, F, R, CF, TH) Which

1) (S, T, and F) compose a basic Petri net.

2) R: S-S [0, 1] is as the correlation function of S. The R (Si) refers to the token of the corresponding state Si, the Reliability of this state.

3) CF: F-F [0, 1] is as the correlation function of F,

CF (s, t) is used to indicate the confidence function from the input state to transition CF (t, s) is used to indicate the confidence function from transition to input.

4) Correlation function of TH: T-[0, 1] is used to indicate the threshold of the before and after transition.

5) Inference (transition) have, if R (Si) • CF (s, t) ≥ TH, then the inference can happen. Special status for a variety of factors reasoning using the weighted reasoning method, $\sum_{i=1}^{n} R(S_i) \cdot CF(S_i, T) \cdot W_i \geq TH(t)$

W_i states that every state is on the weight transition.

In the above definition of a fuzzy inference, its rules can be expressed as some Reliability of the state S_i, after the conversion of the definite Confidence Factor (CF), to fulfill the knowledge inference under the guidance of a certain threshold.

Definition 2 inference with feedback conditions

In the knowledge reasoning verification process, the necessary feedback is obligatory for the correcting of the deviation.

1) The amendment of the threshold, adjust the size of the threshold appropriately in the process of reasoning can filter some of the transitions.

2) The amendment of the confidence function, it can be adjusted in the reasoning process; certain deviation because of some conditions can be corrected.

3) The amendment under the influence of multiple factors weights.

4) The above three amendments to be comprehensive.

Through the inference with feedback conditions and with the certification process of knowledge the certain derivation which was caused by subjective reasons and experiences can be adjusted automatically to improve the accuracy of certification and inference and dig out the potential inference rules.

3.2.2 Description of Algorithm

Input: the Reliability of the reasoning.

Output: the Reliability of the reasoning outcome.

1) Determine the state

```
For any one state S
{
  If ($^{\bullet}S_i = \phi$)
  Then S refers to the initial state, $R(S_i)$ as the initial
      Reliability, more generally, can be understood
      as monitoring, collection, observation or the
      value by experiences;
  If ($^{\bullet}S_i \neq \phi$ && $\cdot S_i \neq \phi$)
  Then S as the intermediate state, $R(S_i)$ as a
      middle inference result of some rules, it can be
      used as the precondition for another inference
      rules;
  If ( $^{\bullet}S_i \neq \phi$ )
  Then S as the result state, $R(S_i)$ as the Reliability
      of the reasoning outcome;
}
```

2) Implementation of reasoning

```
Procedure Reference ( $R(S_i)$ )
{
  $CF : CF(S_i,T)$ ; $TH$ : $TH(t)$
  While
      ( $R(S_i) \cdot CF(S_i,T) \geq TH(t)$ or
      $\sum_{i=1}^{n} R(S_i) \cdot CF(S_i,T) \cdot W_i \geq TH(t)$ )
  Do $R(S_{i+1}) = R(S_i) \cdot CF(S_i,T)$;
  Return $R(S_{i+1})$;
}
```

3) Feedback conditions certification

```
Procedure feedback ( $R(S_i)$ , $CF$ , $TH$ , $W_i$ )
{
  Set the possible range of $R(S_i)$ , $CF$ , $TH$ , $W_i$ ;
Do {
      If (threshold set is too small it will result in
          many unproduced reasoning)
        Then leads to a Step increase $TH$ ;
      If (threshold set is too big it will result in many
          reasoning do not appear)
        Then leads to a Step decrease $TH$ ;
      If (the confidence factors)
        Then adjust $CF$ (using step-by-step
          approximated adjustment);
      If (the distribution of the weight of affecting
          factors is unreasonable)
        Then adjust $W_i$ ;
   } While (accord with the certification conditions
and the correctional results basally) { }
```

4) Interpretation of results

Interpretation of the results in the process of Specific reasoning is the conclusions to be presented in the use of the results but also in the process of combining the actual situation further analysis.

4 Conclusions

The strict theory of Petri Net and mathematics certification have been fully recognized, with fuzzy reasoning technology Fuzzy Petri Net has a lot of advantages in the expression of fuzzy knowledge. In consideration to the feasibility of Fuzzy Petri Net it is particularly important for us to further knowledge reasoning and mining. This paper focuses on the characteristics of knowledge of Fuzzy Petri Net, combine with the situation of research field, and expand to put forward a realized feedback Fuzzy Petri net.

References

[1] Chongyi Yuan, Petri net theory and application, Beijing: Publishing House of Electronics Industry, 2005

[2] Zhibin Jiang, Petri Net and their manufacturing systems modeling and control, Beijing: China Machine Press, 2004

[3] Xingui He, Fuzzy theories and fuzzy techniques in knowledge processing, Beijing: National Defence industry press, 1999

[4] Zhehui Wu, Introduction to Petri Net, Beijing: China Mac-

hine Press, 2006

[5] Amit Konar, Uday K. Chakraborty, Reasoning and unsupervised learning in a fuzzy cognitive map, Information Sciences, Vol.210,2005, pp.419~441

[6] X. Li, F. Lara-Rosano, Adaptive fuzzy Petri nets for dynamic knowledge representation and inference, Expert Systems with Applications, Vol.19, 2000, pp.235-241

[7] Zouhua Ding, Fuzzy Timed Petri Net Definitions, Properties, and Applications, Mathematical and Computer Modelling, Vol.41, 2005, pp.345-360

[8] Sheng-Ke Yu, Knowledge representation and reasoning using fuzzy Pr/T net-systems, Fuzzy Sets and Systems, Vol.75, 1995, pp.33-45

[9] Witold Pedrycz, Generalized fuzzy Petri nets as pattern classifiers, Pattern Recognition Letters, Vol.20, 1999, pp. 1489-1498

[10] Yu Bufan, Gas prevention technology, Beijing: China economy Press, 1987

Particle Swarm Optimization with Adaptive Mutation Operator

Yujuan Chen

School of Information technology, Jiangnan University, Wuxi, Jiangsu, China
Email: yu-juan-chen@163.com

Abstract

PSO algorithm is one of the useful intelligent algorithms for solving constrained and unconstrained global optimization problems. But a major problem for PSO algorithm is premature convergence in solving multimodal problems. By analyzed the reason of premature convergence, an improved particle swarm optimization (PSO) algorithm based on adaptive mutation operator is developed. The performance of the proposed algorithm is tested by solving two benchmark functions. Experimental results show that the proposed algorithm have better performance.

Keywords: particle swarm optimization, premature convergence, mutation

1 Introduction

Particle swarm optimization is a stochastic, population-based evolutionary computer algorithm for problem solving. Particle swarm optimization (PSO) is a population based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995[1,2], inspired by social behavior of bird flocking or fish schooling. It is a kind of swarm intelligence that is based on social-psychological principles and provides insights into social behavior, as well as contributing to engineering applications. PSO shares many similarities with evolutionary computation techniques such as Genetic Algorithms (GA). The system is initialized with a population of random solutions and searches for optima by updating generations. However, unlike GA, PSO has no evolution operators such as crossover and mutation. In PSO, the potential solutions, called particles, fly through the problem space by following the current optimum particles. The PSO algorithm is easy to implement and exhibits good performance in solving hard global optimization problems and engineering applications, and compares favorably to other meta-heuristics, such as Genetic Algorithms (GAs).

Since PSO was introduced, many variations of the PSO algorithm have been developed in the literature to realize faster convergence and better convergence quality, which are the most important two topics in the research of stochastic optimization. PSO has also been used in a wide range of applications, such as multi-objective optimization, classification, pattern recognition, job shop scheduling, real time real time robot path planning, image segmentation [3], weights modification of neural network[4], biometric security systems[5].In many problems, PSO find the optimum value more quickly than traditional evolutionary algorithms, hut easily plunging into the local minimum in the later phase of convergence, especially for the problem space which is non-convex set.

In this paper, we modified the standard PSO algorithm by adding an adaptive mutation operator. The modified PSO will be tested on several famous benchmark functions and the results demonstrate that it is a feasible way to avoid the premature problem in optimizing complex multimodal functions.

The remainder of the paper is organized as follows: Section 2 introduce the concept of PSO algorithm. Section 3 proposes the PSO algorithm with an adaptive mutation operator. Experimental results are reported in Section 4. Finally, the conclusions are given.

2 PSO Algorithm

PSO is a novel natural inspired evolutionary computation technique. The main concept of the algorithm lies in sharing information among the collaborating individuals. PSO is initialized with a group of random particles and then searches for optima by updating generations. Each potential solution is assigned a randomized velocity, which fly through the problem space. Each particle adjust its flying according to its own flying experience and its companion's flying experience. In every iteration, each particle is updated by following two best values, which are called *pbest* and gbest. The pbest is the best solution it has achieved so far. The *gbest* is the best value obtained so far by any particle in the population.

In the paradigms of canonical PSO, supposing the target problem has D dimensions and M particles live in the swarm. The velocity of a particle can be represented by a D dimensional vector $v=(v_1,v_2,\ldots,v_D)$. The particle's velocity is restricted by a maximum velocity v_{max}. The position of a particle can be represented by another *D* dimensional vector $x=(x_1,x_2,\ldots,x_D)$. The pbest is denoted as $P=(p_1, p_2, \ldots, p_D)$. Then the PSO algorithm is manipulated according to the following two equations:

$$v_i^{k+1} = wv_i^k + c_1 r_1 (p_i^k - x_i^k) + c_2 r_2 (p_g^k - x_i^k) \tag{1}$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \tag{2}$$

Where w is the inertia weight, c_1 and c_2 are two positive constants called cognitive and social parameters respectively, which are usually called learning factors, r_1 and r_2 are random numbers uniformly distributed in (0,1). i=1,2,…,M and k =1,2,… is the current iteration.

The inertia weight w is calculated by the following equation.

$$w^k = (w_{start} - w_{end}) \cdot (K_{\max} - k) / K_{\max} + w_{end} \tag{3}$$

Where Kmax is the maximum iteration, wstart is the initial value of w, wend is the value when Kmax is achieved. The inertia weight can balance the search ability of global search and local search of the algorithm. The algorithm with larger inertia weight has good global search ability and vice versa.

The pseudo-code of the procedure is as follows

Step 1. Initialize the population

Step 2. Calculate the fitness value of each particle

Step 3. If the fitness value is better than the best fitness value in history, then set the current value as *pbest*

Step 4. Update the global best value(*gbest*) by comparing the *pbest*

Step 5. Calculate the particle velocity according equation (1)

Step 6. Update the particle position according equaton (2)

Step 7. Repeat from step 2 to step 6 until maximum iterations or minimum error criteria is not reached

3 PSO with the Adaptive Mutation Operator

PSO algorithm converges rapidly in the first frontal iterations and then slows down or stops. This behavior has been attributed to the loss of diversity in the population [6]. In the searching procedure of PSO, the population will fly to the *gbest* position. But if the best location is a local minimum, the other particles will all trap into the position as the global best individual attracts all members of the swarm,. Therefore the algorithm will fall into the local best, which is called premature convergence. Looking at the positions of the particles when the swarm had stagnated, it was clear that the points were very tightly clustered and that the velocities were almost zero. The points were often not that far from the global optimum but the update equations, due to the almost zero velocity, were unable to generate new solutions which might lead the swarm out of this state. This behavior can also lead to the whole swarm being trapped in a local optimum from which it becomes impossible to escape. Hence, a natural idea is that when the particles cluster in a small space for a number of iterations or the global best particle can't update for a while, give a mutation to a particle randomly and generate a new position. The new position can help the particles jump out of the local minimum and the population can begin new search for the best solutions. This mechanism potentially provides a means both of escaping local optima. In our proposed algorithm, the criterion of population cluster is judged by equation (4) and the new particle is generated randomly in the defined search space. And the mutation happens according to the cluster degree and the mutation rate is also decided by the cluster degree.

$$div(population) = \frac{f_{\max} - f}{f_{\max} - \overline{f}} \tag{4}$$

The mutation rate is

$$Pm = \begin{cases} 0.1 & div > 1 \\ 0.01 & div \le 1 \end{cases} \tag{5}$$

We call the proposed algorithm Adaptive mutated PSO (AMPSO).

4 Experimental Settings and Results

To test the performance of AMPSO algorithm and compare with PSO algorithm, we use two benchmark functions as the problem. The two benchmark functions are uni-modal and multimodal respectively and can hardly get favorable results. The figures of two functions are shown in Figure 1[7] with dimension 2.

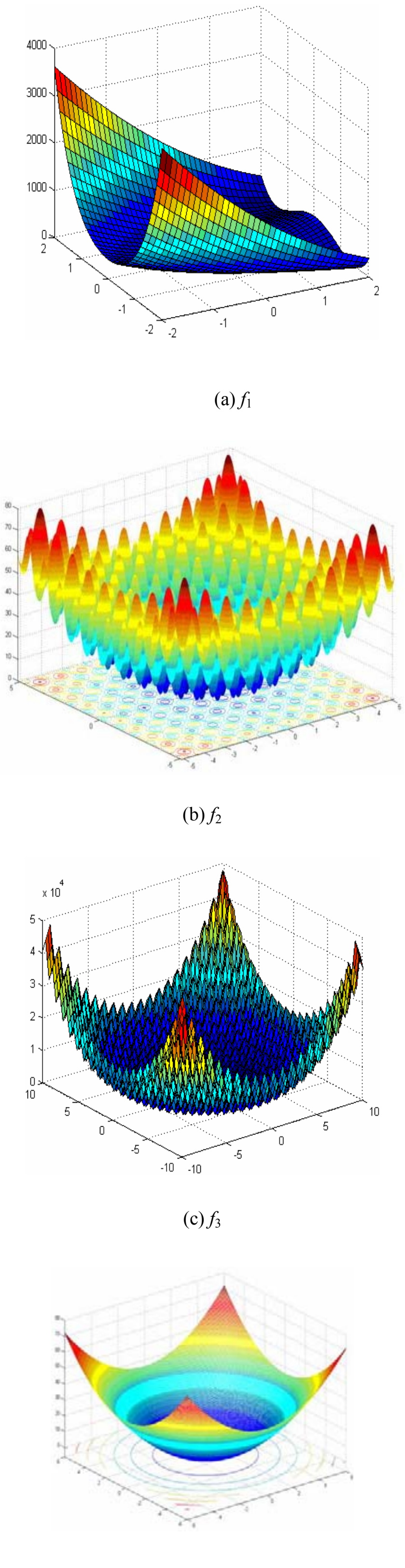
$$f_1 = \sum_{i=1}^{n} (100 \cdot (x_{i+1} - x_i^2)^2 + (x_i - 1)^2) \tag{6}$$

$$f_2(X) = \sum_{i=1}^{n} (x_i^2 - 10 \cdot \cos(2\pi x_i) - 10) \tag{7}$$

$$f_3(X) = \frac{1}{4000} \sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1 \tag{8}$$

$$f_4(X) = \sum_{i=1}^{n} x_1^2 \tag{9}$$

The maximum position of f_1 to f_4 is 30, 5.12, 600, and 100. Two sets of parameter are experimented. The maximum iteration is 1000 and the dimension of each function is 10 and the population size is 20. Another set is, the maximum iteration is 3000 and the dimension of each function is 30 and the population is 30. The maximum value and minimum value of inertia weight is 0.9 and 0.4 respectively. The experiments run 100 times randomly for each function.



(a) f_1

(b) f_2

(c) f_3

(d) f_4

Figure 1 Function graph

Table 1 average results of 100 runs with 10 dimensions

Function	PSO	AMPSO
f_1	62.34	48.25
f_2	9.83	8.65
f_3	0.097	0.088
f_4	0	0

Table 2 average results of 100 runs with 30 dimensions

Function	PSO	AMPSO
f_1	35.21	27.83
f_2	63.63	52.58
f_3	0.075	0.071
f_4	0	0

5 Conclusions

In this paper, the reason of premature convergence in PSO algorithm is analyzed. Then the mutated PSO algorithm is proposed based on the cluster degree of the population and compares its performance with the standard PSO by test the two benchmark functions. Experiments show that the proposed PSO has better performance than the standard PSO.

References

[1] J.Kennedy and R.C.Eberhart, “Particle Swarm Optimization”, Proceedings of IEEE International Conference on Neural Networks, vol.IV, Penh. Australia, pp.1942-1948, IEEE Service Center, Piscataway, NJ.1995

[2] SHI Y, EBERHART R C. A modified particle swarm optimizer.Proceedings of the IEEE International Conference on Evolutionary Computation. 1998, Anchorage, Alaska, pp.69-73

[3] Y.Shi and R.C.Eberhart, “Particle Swarm Optimization: developments, applications and resources”, Congress on Evolutionary Computation 2001, Seoul, Korea, pp. 81-86. IEEE Press, Piscataway, NJ, 2001

[4] Eberhart R. C. and Hu X., “Human tremor analysis using particle swarm optimization”, Proceedings of the IEEE Congress on evolutionary computation (CEC 1999). Washington D.C., pp. 1927-1930.1999

[5] Veeramachaneni K., Osadciw L. A., and Varshney P.K., “Adaptive multimodal biometric fusion algorithm using particle swarm”, Proceedings of SPIE Vol. 5099,pp. 211-221,2003

[6] Riget, J., Vesterstroem, J.S.: A diversity-guided particle swarm optimizer - the ARPSO. Technical Report 2002-02, Department of Computer Science, University of Aarhus

[7] http://www-optima.amp.i.kyoto-u.ac.jp

Research on Multi-valued Multi-input Multi-output Logic Functions Optimization Algorithm*

Qiu Jianlin Li Feng Chen Jianping Gu Xiang He Peng

School of Computer Science and Technology, Nantong University, Nantong, Jiangsu 226019, P.R. China
E-mail: qiu.jl@ntu.edu.cn

Abstract

In this paper, we make an approach to the logic optimization algorithm including two-valued logic optimization algorithm and multi-valued logic optimization algorithm, then present the algorithm to calculate essential prime cube and special relative redundancy cube and construct two-valued logic optimization. We discuss the algorithm converting two-valued logic into multi-valued logic on the basis of building an assignment graph and present multi-valued logic optimization algorithm. we analyze and study logic function mini-covering based on studying the relevant theory of the multi-valued logic optimization which including the identification of judging macrocosm, the identification of essential prime implicants and relative redundancy and absolute redundancy implicants, and so on. We design and realize a software system on logic optimization in programming language C. It overpasses testing of Benchmark and right validate and it shows that the function of the software system on logic optimization is perfect and the optimization efficiency is so high viaw testing.

Key words: Logic optimization, Algorithm, Two-valued logic, Multi-valued logic, Mini- covering

1 Introduction

The two-valued logic optimization of sum of products (SOP) is concerned in the last forty years. Gimpep presented the idea of covering cubes--choosing the sum of the products in 1965[1]. Karp showed that a cover is the problem of NP in 1972[2].Time consumption should be considered as a perfect algorithm. So optimization algorithms are given considering time consumption. The common algorithm--ESPRESSO[3] based on the two-lever logic optimization algorithm needs much more time to calculate the complement set of the function. It will need more time-consumption to solve a larger problem. Bryant limited the indicated order of the variables in Binary Decision Diagram (BDD) and put emphasis on the uniqueness[4]. But it remains a problem to identify the orders of the variables in multi-valued logic function.

The two-valued logic optimization algorithm can convert two-valued logic into multi-valued logic and reduce the number of the products by building an assignment graph[5]. It leaves aside the consumption of memory space and time while calculating the complement set. Select the best minterm as a base minterm in order to calculate the best covering by rearranging cubs[5]. The multi-valued logic optimization algorithm can make up for the problem such as non-uniqueness, great consumption and so on in traditional Karnangh Maps, Binary Decision Diagram(BDD), etc.

On the basis of analyzing and mastering the algorithm of Espresso, Vanilla, Beister, etc[3] [6-8][10], a software system based on multi-input multi-output logic functions optimization algorithm is designed. It allows completeness enumerate inputs and non-completeness enumerate inputs. For two-level SOP multi-input & multi-output logic function, max-input variables is 128,max-output variables is 256,sum of max-input-output variables is 300, max-input products is 20000 according to the true fact of hardware of computer. It overpasses testing of Benchmark and right validate.

* This work was supported in part by the Science Foundation of JiangSu Grants BK2001130 and the Science Foundation of JiangSu Education Department Grants (03KJB520103, 05KJD520166, 06KJB520090) and the Science Foundation of Nantong of JiangSu Grants (AL2007033, K2006008, C3024).

2 Two-Valued Logic Optimization

Supposing that the function F is a multi-input and multi-output logic function with n input variables and m output variables, and then it can be described as follows: $F(x_1, x_2, \ldots x_n): \ B^n \to \{0,1,*\}^m$

The variable that appears in the true form is coded with a 1, and the variable that appears in the complement form is coded with a 0, and the variable that doesn't appear in a product term is coded with a * .For example, the function $F(x_1, x_2, \ldots x_n) = x_1 \ \bar{x}_2 + x_1 x_3 \bar{x}_4$ can be denoted by $F(x_1, x_2, \ldots x_n) = \ \{10**, 1*10\}$.

2.1 Identification of the Essential Prime

We can compute the adjacent and intersecting cube of a prime cube AIC $(\mathrm{P_i})$ [9]according to the prime cube $\mathrm{P_i}$ .Firstly it must be adjacent to $\mathrm{P_i}$ and $(\mathrm{P_i}) \subset (\mathrm{C_{ON}} \cup C_{\mathrm{DC}})$. Secondly, it must intersect $\mathrm{P_i}$ in don't-care. Thirdly, it must intersect $\mathrm{P_i}$ in the onset $\mathrm{C_{ON}}$ but not include the prime cube $\mathrm{P_i}$.

The main steps of generating essential prime cube E [9] can be described as follows:

Expand the prime cube $\mathrm{P_i}$, and then compute the AIC $(\mathrm{P_i})$ .If $\mathrm{P_i}$ subsumes AIC $(\mathrm{P_i})$,then $\mathrm{P_i}$ is a redundant cube, and delete it from the $\mathrm{C_{ON}}$, or else put $\mathrm{P_i}$ into don't-care and E ,and delete the product which is subsumed by $\mathrm{P_i}$ from the don't-care.

Note:(1)If the set which is expanded according to $\mathrm{P_i}$ subsumes a set of $\mathrm{C_{ON}}$, delete the set from $\mathrm{C_{ON}}$.

(2)If $\mathrm{P_i} \#\ \mathrm{AIC}(\mathrm{P_i}) \neq \varnothing$, $\mathrm{P_i}$ is an essential prime implicant. It should be put into the set E and the don't-care C_{DC}. Then delete it from the set $\mathrm{C_{ON}}$.

2.2 Identification of the Relative Redundancy

The principles are just as follows:

Choose the relative redundancy P with the least number to cover all of the mini-terms which are not covered by the set E. Combine the special calculated relative redundancy with the calculated set E, and then we can get the optimized result of the logic function.

The main steps of generating P [9]can be described as follows:

1. Find the set which intersects the don't-care C_{ON} and put it into the set C_{ON}.

2. Rearranging cubs——delete the redundancy from C_{ON} to shrink the covering scale of C_{ON}.

3. Select the max set with the least degree of adjacency (DA) if any of the products does not intersect the don't-care C_{DC} or the don't care is empty, and put it into the don't care C_{DC} and the set P, then delete it from C_{ON}.

4. If the number of the sets which are contained by the set $\mathrm{P_i'} = \mathrm{P_i} \#(C_{\mathrm{ON}} \cup C_{\mathrm{DC}} \# \mathrm{P_i})$ is not more than one for $\mathrm{P_i}$ of $\mathrm{C_{ON}}$, we should rearrange cubs. Put it into C_{DC} and P after expanding and rearranging $\mathrm{P_i}$, then delete it from $\mathrm{C_{ON}}$. If the number of the sets which are contained by the set $\mathrm{P_i'} = \mathrm{P_i} \#(\mathrm{C_{ON}} \cup C_{\mathrm{DC}} \# \mathrm{P_i})$ is more than one, put the set with the largest degree of adjacency (DA) into C_{DC} and P, then delete it from $\mathrm{C_{ON}}$ .Then go to the step 3. The algorithm should be continued until C_{DC} is empty.

Note: (1) If the subset of the set in $\mathrm{C_{ON}}$ is not covered by the other CoH $\mathrm{C_{ON}}$ and $\mathrm{C_{DC}}$, expand it and rearrange P in order to cover the other sets of $\mathrm{C_{ON}}$.

(2) If the degrees of adjacency (DA) of several cubes among the largest cube are the same, we should put the sets which have a non-empty intersection with the other sets in $\mathrm{C_{ON}}$ into $\mathrm{C_{DC}}$ and P, and then delete from $\mathrm{C_{ON}}$.

3 Multi-Valued Logic Optimization

The function with multi-valued multi-input logic function $F(x_1, x_2, \ldots x_n)$ can be defined as follows:

$$P_1 \times P_2 \times P_3 \times \ldots \times P_n \to B$$

Let x_i is input variable and B={0,1,*} (The symbol * stands for a don't-care product. It can be 0 or 1). $\mathrm{P}_i = \{0,1,\ldots,\mathrm{P}_i - 1\}(\mathrm{P}_i \geq 2)$ and $\mathrm{P_i}$ is the prospective value of the variable.

Postulating that X is an input variable and the value $X \subset \mathrm{P}\{0,1,2,\ldots,\mathrm{P_i}{-}1\}$ and S is the subset of P, X^s can be defined as:

$$X^S = \begin{cases} 1 & X \in S \\ 0 & X \notin S \end{cases}$$

For example,

$$F（X_1, X_2, X_3）= X_1^{\{0,1\}} X_2^{\{1,2\}} X_3^{\{1\}} \vee X_1^{\{1\}} X_2^{\{0,1,2\}} X_3^{\{1\}} \vee X_1^{\{1,2\}} X_2^{\{0,3\}} X_3^{\{3,0\}}$$

3.1 Converting Two-valued Variable into Multi-valued Variable

Definition 1: Postulating that S is a subset of input variable and $|S|=2$ that means S contains two input variables, delete the input variable which is contained by S for each product and the number of the remained products can be denominate as R_S.

The feature of the assignment graph:

(1) There are n notes and each note stands for an input variable.

(2) $R_{(x_i,x_j)}$ stands for the weight of the edge between the notes i and j.

We can convert two-valued multi-inputs variable into multi-valued multi-inputs variable and cut down the number of the products by building the assignment graph. But we can not get the best assignment combination directly by doing that, so the assignment graph should be improved.

Definition 2: The times that the variable x_i appears in the un-complemented form is the un-complemented weight of x_i. The times that the variable x_i appears in the complemented form is the complemented weight of x_i.

The feature of the enhanced assignment graph:

(1)There are n notes and every note stands for an input variable.

(2) Every note has its own corresponding un-complemented weight.

(3) $R_{(x_i,x_j)}$ stands for the weight of the edge between the notes i and j.

We can convert two-valued variable into multi-valued variable and cut down the number of the products by building the enhanced assignment graph. The process can be described just as follows:

STEP1:Calculate the un-complemented weight of every note.

STEP2:Calculate the weight $R_{(x_i,x_j)}$ between the note i and the note j.

STEP3:If(i,j) satisfies with the condition—the note i has been chosen and $R_{(x_i,x_j)}$ is the smallest (the note j has not been chosen) then (i,j)can be selected.

If $R_{(x_i,x_j)}$ has only one value to choose (all of the weight crossing the note i are the same) the choose the note j which is nearest to the un-complemented weight of the note i. Do that again and again until all of the notes are chosen.

3.2 Multi-Valued Logic Optimization

The process of multi-valued logic optimization makes up with the steps of expanding products and choosing the special prime implicants. The step of expanding products can be realized by covering the prime implicants in order to get sub function. The step of choosing the special prime implicants can be realized by choosing the special prime implicants from the sub function in order to get non-redundancies covering. The number of the sub functions affects the optimized efficiency to a great extent. We can get a covering much more close to the mini covering if the set of the sub function is larger. Select the best minterm as a base minterm in order to calculate the best covering from the standards cubes by rearranging cubs. It produces sub functions in the process in order to find the essential prime implicants earlier. It can reduce the time calculating the minterm with the smallest number of adjacents by rearranging the given cubs. The algorithm is based on building table of indices and rearranges according to the number of adjacents. It can be drawn just as follows:

STEP1: Check X_1^i (1<i<m, m stands for the number of values).

STEP2: Calculate the number of distinct values for each X_1^i (i∈[0, n-1], n stands for the number of states),then update the table of index by the weight of the distinct values(the times of the appearance).

STEP3: Calculate the number of distinct values for each x^{i}_{j} (i∈[0, n-1], 1<j<n-1,n stands for the number of states),then update the table of index by the weight of the distinct values(the times of the appearance) until the table of index does not change any more. When we finish the process of updating the table of index, the process of rearranging for the given cubs is completed.

4 Technology of Logic Lunction Mini Convering

4.1 The Matrix Form of Logic Function

Definition 3: Supposing that the function F is a multi-input multi-output logic function with n input variables and m output variables and *P* is a given product, and $p^k \in P$, the expression of p^k by vector can be described as follows: $V(p^k)=[p_1^k p_2^k \ldots p_n^k p_{n+1}^k p_{n+2}^k \cdots p_{n+m}^k]$. $p_1^k p_2^k \ldots p_n^k$ is input and $p_{n+1}^k p_{n+2}^k \cdots p_{n+m}^k$ is output. For the input p_i^k ($1 \le i \le n$) of product p^k, 0 expresses reverse variable, 1 expresses original variable and 2 expresses absent variable. For the output p_i^k ($n+1 \le i \le n+m$)of product p^k, if $p^k \in f_t$, p_{n+t}^k is 4, otherwise p_{n+t}^k is 3.

4.2 The Identification of Judging Macrocosm

Definition 4: Supposing that the function F is a multi-input multi-output logic function with n input variables and m output variables and P_i is a product of the product set $P=\{P_1,P_2,\ldots,P_i,\ldots, P_s\}$ ($P_i \in P$,1≤i≤s,s is a product of the product set P), $V(P_i)=[p_1^i p_2^i \cdots p_n^i p_{n+1}^i p_{n+2}^i \cdots p_{n+m}^i]$; for the given product $Q=[q_1,q_2,\ldots,q_n, q_{n+1},\ldots,q_{n+m}]$,the product set P for cofactor of the product Q is the product set P_i ($1 \le i \le s$) for cofactor of the product Q, it is defined as:

$$(p_j^i)_Q=\begin{cases} \phi & (p_j^i=0 \cap q_j=1)\cup(p_j^i=1\cap q_j=0) \\ 2 & (p_j^i=q_j=1)\cup(p_j^i=q_j=0)\cup(p_j^i=2) \\ 4 & q_j=3 \\ p_j^i & (q_j=2)\cup(q_j=4) \end{cases}$$

Theorem 1: The product set G covers the product P while G for cofactor of the product P is macrocosm: $G_P=U$.

The main process of identification of judging macrocosm (Vanilla algorithm) is:

(1) If G contains a row with each element is 2, then G is a macrocosm. If G contains a column with each element is 0 or 1, then G is not a macrocosm.

(2) If not, choose a splitting variable x_i, computer G_{x_i} and $G_{\overline{x_i}}$, then judge whether each splitting satisfies the condition of the process (1) or not.

(3) If a splitting is not a macrocosm, the algorithm should be stopped and then jump to the conclusion that G is not a macrocosm. Only when each splitting is a macrocosm, G is a macrocosm. If it does not satisfy the two cases of the process (1), then repeat the process unless the case of the process (1) appears.

4.3 The Identification of E、R、P

If a given function F's cover is prime and it does not contain any repetitive prime product, then we can identify the essential prime implicants by testing each g_i∈G is covered by the set G-{g_i}∪D. That means identifying whether(G-{g_i}∪D)g_i is a macrocosm according to the Theorem 1. If it is a macrocosm, we can draw a conclusion that g_i is not an essential product, otherwise g_i is an essential product. The set of the products can be denominated as E.

The identification of redundant products is also easy. Discarding the essential products from G and checking each g_i∈G one by one, if (E∪D) g_i is a macrocosm, we can draw the conclusion that g_i is an absolute redundancy and it should be deleted. We can get the relative redundancy implicant P by discarding the essential prime implicant E and the absolute redundancy implicant R from G.

4.4 Selection of P*

Identifying the E、R、P is easier than selecting the mini subset P* from the set P . The set $P^* \cup E$ covers the function f .

We can quote an assistant function $\xi : B^k \to B$,

$k=|P|$. While $y=[y_1y_2...y_k]\in B^k$, the function ξ can be defined as:

$$\xi(y)=\begin{cases}1 & E\cup\{g^i\in P|y_i=1\}\text{ cov}ers\quad f\\ 0 & other\quad case\end{cases}$$

Once the function $\xi(y)$ is formed, we can get the cover of the function *f* by combining with the set E. Solution of mini covering can be converted into compute a vector y with the mini number of input variables while $\xi(y)=1$. It is also equivalent to compute the maximal prime implicant of the function ξ. If a cover *Q* of the function ξ can be formed according to the given cover *G*, the problem can be solved according to the algorithm of generating prime implicants. But it is easier to form the cover $\overline{Q}$ of $\overline{\xi}$ than to form the cover *Q* of the function ξ.

Theorem 2: For every $g\in P$, *S* is the mini subset of P and the set $(P-S)\cup E\cup D$ does not cover g, $\varphi(g)$ is the set of *S*. We can define a vector $C(S)$ for each $S\in\varphi(g)$ just as follows:

$$C(S)_i=\begin{cases}0 & g^i\in S\\ 2 & g^i\notin S\end{cases}$$

Then $\overline{Q}=\bigcup\limits_{g\in p}\bigcup\limits_{s\in\varphi(g)}C(S)$ is a cover of $\overline{\xi}$.

Supposing that $\overline{B}$ is the transfer-meaning matrix of $\overline{\xi}$, $\overline{B_{ij}}=\begin{cases}1 & \overline{\xi_{ij}}=0\\ 0 & \overline{\xi_{ij}}=2\end{cases}$

Computing the maximal prime implicants of the function $\overline{\xi}$ is equivalent to compute the mini column covering of $\overline{B}$. Every row of $\overline{B}$ means a kind of $S\in\varphi(g)$, discarding the case $g\in P$. Delete the prime implicants that the symbol 1 stands for by column in every row. If one case appears, it means that the row does not have the ability of cover. Once a column cover of column is chosen, any case without covering does not exist. We can get a cover with the smallest number of implicants after choosing a column cover.

The key to the problem is how to locate each row of $\overline{B}$. We can compute $\varphi(g)$ for every $g\in P$. According to the definition of $\varphi(g)$, the set $(P-S)\cup E\cup D$, does not cover g only when S is the mini subset of P ($S\in\varphi(g)$). According to the theorem 2 that means $((P-S)\cup E\cup D)_g\neq U$. Then the train of our thought can be converted into judging whether the set $(P\cup E\cup D)_g$ is a macrocosm. We can get δ by the way of deleting the rows which can not maintain a macrocosm in the process of judging macrocosm.

In order to make the calculus and writing brief, we can entitle $\alpha=(P)_g$ and $\beta=(E\cup D)_g$. According to the property of the cofactor:

$(P\cup E\cup D)_g=(P)_g\cup(E\cup D)_g\quad=\alpha\cup\beta$ and $g\in P$, then we can get $\alpha\cup\beta=U$.

If there is a row with each element being 2 in the set β and the set $E\cup D$, the case without covering g can be excluded aside in the process of judging macrocosm by the algorithm of Vanilla, so $S=\varnothing$. If β is not a macrocosm and each element of each row of α is 2, the products represented by the rows can make up S. If we delete the rows at the same time, $\alpha\cup\beta\neq U$, that means it does not cover g. We can get S and form $\varphi(g)$ during the process of judging macrocosm.

Example: Calculate the mini covering set of the multi-valued multi-input multi-output logic function F with its matrix form is given just as follows:

$$\mathrm{M}(P)=\begin{bmatrix}002 & 443\\ 120 & 344\\ 101 & 444\\ 102 & 344\\ 012 & 344\end{bmatrix}$$

Answer: We can get $\mathrm{M}(D)=\varnothing$ according to the problem.

$$\mathrm{M}(G)=\begin{bmatrix}002 & 443\\ 120 & 344\\ 101 & 444\\ 102 & 344\\ 012 & 344\end{bmatrix}$$

Then we can get as follows according to the identification of E、R、P:

$$\mathrm{M}(G_E)=\begin{bmatrix}002 & 443\\ 120 & 344\\ 012 & 344\end{bmatrix},\mathrm{M}(G_\mathrm{R})=\phi,$$

$$\mathrm{M}(G_{\mathrm{P}})=\begin{bmatrix}101 & 444\\ 102 & 344\end{bmatrix}$$

Calculate the transfer-meaning matrix $\overline{\mathrm{B}}=\begin{bmatrix}1 & 0\\ 1 & 1\end{bmatrix}$ by the relative redundancy implicant set G_{P}. Then choose the first column (j=1) according to the mini covering algorithm. And choose the first row of the relative redundancy implicant set G_{P} as the mini column covering P*= $G_{\mathrm{P}}\left[101 \quad 444\right]$. So the set $\mathrm{P}^{*}\cup\mathrm{G}_{\mathrm{E}}$ is the mini covering:

$$\mathrm{M}(P)=\begin{bmatrix}002 & 443\\ 120 & 344\\ 012 & 344\\ 101 & 444\end{bmatrix}$$

5 Results of experiment and discussion

We have designed a software of logic optimization of multi-input multi-output logic function based on mini covering in language C. It can be divided as follows by function: read the source data files, discarding C_{DC} calculate the cofactor, identify the macrocosm, choose the mini row covering, identify the essential prime implicant E and absolute redundancy implicant R and relative redundancy implicant P, select the implicant with the mini covering from the relative redundancy implicants and display the optimized results. The software system overpasses the testing based on computer of Pentium 1.8 GHZ CPU and 512 MBRAM. And max-input variables is 128, max-output variables is 256, sum of max-input-output variables is 300, max-input products is 20000. It overpasses 105 tests about PLA89、PLA91、PLA93 of Benchmark and we present a part of tested results in the Table 1. The software is proved to be correct and efficient by analyzing the results of experimentation.

(1) From the Table 1, we can see ex4.pla with the max-input variables 128, the Cps.pla with the max-output variables 109 and apex5.pla with sum of max-input-output variables 117+88.

(2) The example which is proved to be a macrocosm is the first example and the example proved to be vacant is the fourth example in the Table 1 with ten examples. And there are seven examples with completeness enumerate functions (not including the don’t-care products) and three examples with non-completeness enumerate functions (including the don’t-care products) in the Table 1. In all of examples except only one example proved to be a macrocosm, the rest nine examples including eight absolute redundancy implicants and a non-absolute redundancy implicants and five relative redundancy implicants and four non-relative redundancy implicants.

(3) The optimized number of the products is 1053 based on the expanding products optimization while the number of the products is 1459 in the source file of seq.pla in the test. And it can identify 572 absolute redundancy implicants. In all of the 1053 products, 572 products can be thrown into the discard by the identification of the absolute redundancy implicants. The ratio of the absolute redundancy implicants to the expanded products is 0.542 with the best efficiency one.

6 Conclusion

Logic optimization can be realized by converting two-valued into multi-valued or converting multi-valued into two-valued. In this paper, we mainly introduce the logic optimization algorithm that converts multi-valued into two-valued and then calls the two-valued logic optimization algorithm. The result optimized by each algorithm is just approximation of the optimization, Maybe there are many different optimized results. We check the validity of logic optimization algorithm by covering compatibility. The optimization algorithms in this paper are improved to be better in the consumption of memory space and time than some general common logic optimization algorithms. Since the optimized efficiency is affected by the order of the chosen column, how to decrease the effect remains to be improved.

References

[1] J.F.Gimpel. A Method of Producing a Boolean Function Having an Arbitrarily Prescribed Prime Implicant Table[J]. IEEE Trans. Electron. Comput. pp.485-488, June 1965

[2] R,M.Karp. Reducibility among Combinatorial Problems in Complexity of Computer Computations[M]. R.E.Miller and J.W.Thatcher, Eds. New York: Plenum, 1972, pp.85-103

[3] R.K.Brayton, G.D.Hachtel, C.T.McMullen and A.L.Sangiovanni-

Vincentelli, Logic Minimization Algorithms for VLSI Synthesis[M]. Boston: Kluwer Academic Publishers, 1984

[4] Randal E, Bryant. Graph-Based Algorithms for Boolean Function Manipulation[J]. IEEE Trans. On Computers, 1986, C35(8): 677- 691

[5] Hafiz Md.Hasan Babu, Moinul Islam Zaber, Md.Rafiqul Islam and Md.Mazder Rahman. On the Minimization of Multiple-Valued Input Binary-Valued Output Functions[C]. Proceedings of the 34th International Symposium on Multiple-Valued Logic, May 19-22, 2004, 321-326

[6] D.M.Giovanni.Synthesis and Optimization of Digital Circuits [M]. San Francisco, USA, McGraw-Hill, Inc, 1994

[7] R.B.Cutler, S.Muroga.Derivation Of minimal sums for completely specified functions[J]. IEEE Trans. On Computers, 1987, C-36(3):277-292

[8] J.R.Slagle.A new algorithm for generating prime implicants [J]. IEEE Trans. On Computers, 1975, C-24(10):924-930

[9] P.Srinivasa Rao and James Jacob. A Fast Two-level Logic Minimizer[J]. Proceeding of the 11th International Conference on VLSI Design, Jan 4-7, 1998, 528-533

[10] Till Mossakowski, Michael Drouineaud, Karsten Sohr. A temporal-logic extension of role-based access control covering dynamic separation of duties[C].10th International Symposiun on Temporal Representation and Reasoning and Fourth International Conference on Temporal Logic, Cairns, Queensland, Australia, July 8-10, 2003, 83-90

Qiu Jianlin, MS, associate professor. He was born in Nantong, Jiangsu, China in 1965. He graduated from Department of CS , Hohai University China and joined the faculty of School of CS, Nantong University in 1985. During 2003-2004, he was a visiting professor at Dept. of Mathematical and Computer Sciences, Colorado School of Mines(CSM), USA. He is a Senior Member of CCF and a member of Information Stored Special Committee of CCF. His research interests include logic synthesis and optimization and computer-aided VLSI design and network security.

Li Feng, a student of MS in Department of CS , Nantong University . She was born in Xuzhou, Jiangsu, China in 1984. Her research interests include logic synthesis and optimization and computer-aided VLSI design.

Chen Jianping, MS, professor. He was born in Nantong, Jiangsu, China in 1960. He graduated from Department of CS , Nanjin University of science and technology of China in 1982 and joined the faculty of School of CS, Nantong University in 1985. His research interests include algorithm optimization design and network security.

Gu Xiang, PHD, associate professor. He was born in Nantong, Jiangsu, China in 1973. He graduated from Department of CS, University of Science and Technology of China in 2004. His research interests include network security and network protocol.

He Peng, MS, lecturer. He was born in Nantong, Jiangsu, China in 1980. He graduated from Department of CS , SuZhou University in 2005.His research interests include network technology and network security.

name	input variab-les	output variab-le s	original files		expanded files		Identify				optimized		optimized efficiency	
			products	letters	products	letters	don't care	E	R	P	products	letters	optimized/ original products	optimized/ original letters
5xp1.pla	7	10	75	296	1	macrocosm	75				macrocosm			
Alu4.pla	14	8	1025	7848	954	7280	0	463	132	359	587	4533	0.573	0.578
apex5.pla	117	88	1227	7106	1149	6322	0	1088	61	0	1088	6089	0.887	0.857
bw.pla	5	28	65	240	3	5	87	0	3	0	vacant			
clip.pla	9	5	167	888	153	778	0	137	16	0	137	695	0.820	0.783
Cps.pla	24	109	654	7156	425	4856	0	186	208	31	192	2225	0.294	0.311
duke2.pla	22	29	87	759	26	148	87	11	5	10	13	70	0.149	0.092
ex4.pla	128	28	620	4404	410	2454	0	410	0	0	410	2454	0.661	0.557
misex1.pla	8	7	32	122	18	70	0	12	4	2	13	53	0.406	0.434
Seq.pla	41	35	1459	17823	1053	12636	0	470	572	11	473	5964	0.324	0.335
Average													0.514	0.493

A Review of Artificial Immune Network Models

Danxia Li[1] Li Liu[1] Choihong Lai[2]

1 School of Information,Southern Yangtze University, Wuxi, Jiangsu, China
Email: li_danxiamn@163.com

2 School of Computing&Mathematical Sciences, Greenwich University Old Royal Naval College
Park Row, London SE10 9LS, UK
Email: c.h.lai@gre.ac.uk

Abstract:
Nowadays, the Artificial Immune Network (AIN), a main part of the Artificial Immune System (AIS), has become one of the most active directions of AIS research areas. It has been widely used in the data analysis, clustering, data mining, function optimization and some other application fields. This paper mainly introduced a genealogical tree of artificial immune networks. Then artificial immune networks for data analysis and optimization were described in details. Finally, comments and prospects of future researches were also discussed.

Keywords: AIS, Artificial Immune Network Model, aiNet, opt-aiNet, dopt-aiNet

1 Introduction

The Artificial Immune System is a new kind of artificial intelligence technique appeared in 1990s. It imitates the immune system of biology (especially human being), having the ability of learning memory, distributed processing and so on. It can provide some new methods of information processing and have been used in such fields as control, clustering, computation, etc. Due to add the interaction mechanism between immune cells on the base of the traditional AIS model, it makes the behaviour of cells more complex and more efficient on information processing. So such kind of network is usually considered to be a Unique Network. The immune network theory was proposed by Jerne [1], as a way to explain the memory and learning capabilities exhibited by the immune system. It assumed that the lymphocyte cells are not isolated and they are connected through the interaction between antibodies. Therefore, the identification of antibody is not completed by a single set of identification but the system reorganization. Jerne believed that one cell can produce only one kind of antibody and he concluded that: (1) all receptors with similar antibody showed by one lymphocyte cell should be equivalent or at least have equivalent immune global light chain and heavy chain; (2) all antibodies produced by single cell and their descendants should have the same unique model[2]. Based on Jerne's work, some models of immune network were developed using differential equations to predict the antibody concentration during and after an immune response. The first models were proposed by Jerne, Farmer etal, and Varela etal, see this paper[3] for a summary.

2 Artificial Immune Network Model

N.K.Jeme proposed the unique network model and wrote down the differential equation of this model on the base of the understanding of the uniqueness of antibody Molecular in1972. In 1974, Jerne released a landmark article and then the artificial immune network became the focus. The immune system can turn out the immune memory through the interaction between network B cells. The unique feature and anti-unique feature showed when the antigen combined with antibody is the Material basis of forming mutual stimulation and mutual restriction. Thanks to the network cell interaction providing a network adjusting function, it insures the structure of network memory stable even if there is no antigen. Later, many people have improved this model, such as Farmer, Perelson, Bersini, Varela and so on. And there appeared many different artificial immune network models. This paper gives a supplement for the genealogical tree and describes its evolution according to different application scopes, On the base of the genealogical tree [4]of artificial immune network model drew by Juan, All models being in the first layer are Jerne and Farmer, Varela. The Couthino model is the root of the tree. And the above models are the foundation of the lower layers models.

2.1 Artificial Immune Network Models for data analysis.

In 1996 Hunt and Cooke [5] proposed an artificial immune network, which was applied to pattern recognition tasks in DNA sequences. That model considered the immune system as a network of B-cells that are related to other B-cells by its affinity and its enmity. Such relationships are based on Hamming distance following the Farmer's work. Any B-cell undergoes cloning and mutation, if the current non-self antigen makes it stimulated enough, in other words, if the B-cell stimulation level is greater than a threshold. The cloning process produces a number of exact copies of the B-cells depending on the stimulation level. The mutation process is based on a random selection between three kinds of techniques. At each end of iteration a subpopulation of the less stimulated B-cells are removed from the network and the same percentage of new cells is generated and incorporated in.

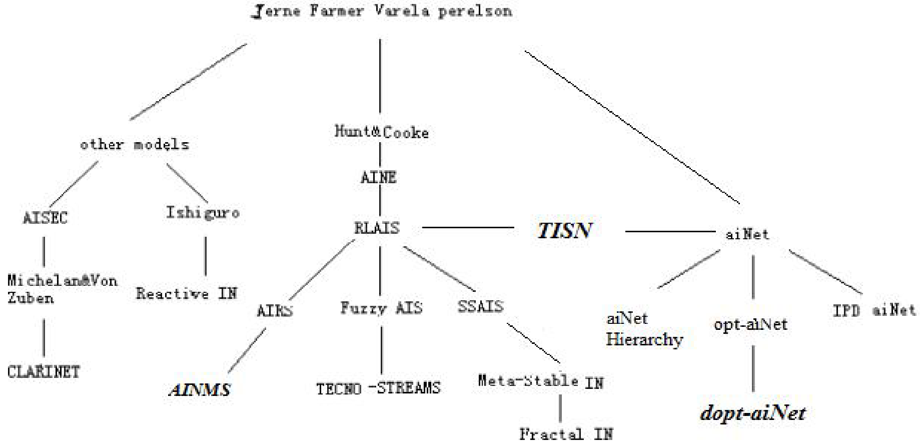


Figure 1 the Artificial Immune Network Genealogical Tree

AINE (Artificial Immune Network):

Timmis[6]proposed AINE network model used to data analyses in 2000 on the base of improving Hunt and Cooke model. The AINE model is composed of ARB collection and its relative relations (immune network). In this model, ARB shows the data in a competitive way. The number of the data is proportioned to the stimulating level of ARB. When ARB no longer stands for the data, it will be removed from the AINE model. The cloning and mutation is used to introduce the individual diversity of the inner immune network. The computation is finished When ARB network reached to the corresponding stability standard or the settled learning algebra. The three computation step (including the computation of ARB stimulating level, B-cell distribution and ARB cell cloning and variation) of the AINE model are crucial to the model.

RLAIS(Resource Limited Artificial immune system)

In2001 Timmis and Neal [7] modified the AINE model introducing the concept of Artificial Recognition Ball (ARB) according to the theory of immune network, this model simulates B-cell interaction relations for data classification analysis. If the classification characteristics vector is considered as antigen, then the B-cell in this system is the randomly generated vector in the same characteristics space. In RLAIS, these antigens are presented to the B-cell system and those B-cells with highest affinity will clone and undergo variation. The system does not keep so many the same B-cells but keep the ARBS which are representations of each same group. These ARBS will compete with each other to try to accept by system. With the introduction of the new antigens, the new B-cells will generate the existing B-cells or ARBS through cloning and variation. Some ARBS will be removed if they are still lack of stimulating after a period of time exposed to the antigen. Only those that have strongest response to the antigen, ARBS will survive. The final system is a kind of system forming antigen clustering.

AIRS (Artificial Immune System)

Timmis[8] proposed a new classification system on the base of RLAIS in 2003. AIRS retains the ARB concept of RLAIS as a supervised learning system.AIRS puts much attention to distinguish ARBS responding to the different types of antigen (characteristics vector of training collection). Those ARBS who have strong response will add to the memory cell pool after further treated and will be kept as the final classification tool when training finished. The mutation of computation would successfully produce memory ARBS because the mutation does not generate the same antigens with training vector but produce antigens that are very similar with the training vector. And the antigens have a high competitive ability. This is very different with immune system. But it is an important reason for AIRS with data extensive ability. AIRS begins to train memory cell pool from the initialization settled memory cell group. AIRS calculate affinity by Euclidean distance, so it tests classifications through real number vector.

Fuzzy AIS

In 2002, Nasraouietal[9] presented a model, based on RLAIS for performing clustering and web mining. The Fuzzy ARB concept is introduced. A Fuzzy ARB defines a fuzzy set over the domain of discourse consisting of the training data set. Each fuzzy ARB is allowed to have its own scale/radius of influence (similar to NAT). Other difference with RLAIS is that those ARBS whose affinity is less than a certain threshold are merged (crossover operator). Stimulation and suppression interactions by both antigens and ARBS are considered. The cloning and mutation operators are applied over cells remaining after removing (remove ARBS with zero B-cells allocated). The resource allocation process is modified.

TSIN

Because the both models were different in the immune

mechanism so we design different immune operator and evolute the populations through the application of immune operator. At last can get antibody population discribling patterns of distribution of data sets. However, these two models have a common flaw, that is, network topology plays a weak role in the course of the evolution and can not guarantee that final antibody stocks topology is consistent with the distribution of the antigen. This paper[8] presents a new type of tree artificial immune network model (Tree Structured Immune Network, TSIN). TSIN, the network topology definition for the DAG $\mathrm{G} = < V, E >$, $V \subset AB$, E as a direction edge collection. The starting points of edge on behalf of the father-generation antibodies. the end points of edge representative offspring antibodies generated through the clone selection process.

2.2 aiNet model

DeCastro[10] and others proposed a kind of immune network named aiNet. The difference between B cell and antibody was ignored in this network. And achieve the network with statistics reasoning technology. aiNet is a Border-weighted map, and there is no need to connect all. It is also called cellular notes collection. Notes are boundary for the collection. Every connecting boundary has a unit of assembling weight or connecting intensity. Clustering in the network is used as inner reflection, and is responsible for inflecting the clustering from data collection to network clustering used as a network within the image, in charge of mapping data collection in the cluster to the network clustering. A hypothetical aiNet structure was generated by the network. The algorithm as following:

(1) At each iteration, do:

1) For each antigenic pattern $Ag_j, j=1,\ldots M$, $(Ag_j \in Ag)$ do:

① Determine its affinity $\mathrm{f}_{\mathrm{i.j}}, i=1,\ldots N$, to all

$Ab_i.f_{i.j} = 1/D_{i.j}, i=1,\ldots,N : D_{i.j} = \| Ab_i - Ag \|, i=1,\ldots,N$

(1)

② A subset $Ab_{\{n\}}$ composed of the n highest affinity antibodies is selected;

③ The n selected antibodies are going to proliferate (clone) proportionally to their antigenic affinity $f_{i.j}$ generating a set c of clones: the higher the affinity, the larger the clone size for each of the n selected antibodies (see Equation(6)).

④ The set C is submitted to a directed affinity maturation process (guided mutation) generating a mutated set C^*, where each antibody k from C^* will suffer a mutation with a rate α_k inversely proportional to the antigenic affinity $f_{i.j}$ of its parent antibody: the higher the affinity, the smaller the mutation rate:

$$c_k{}^* = c_k + \alpha_k (Ag_j - C_k); \alpha_k \propto 1/f_{ij}; k=1,\ldots,N_c; \quad i=1,\ldots,N. \tag{2}$$

⑤ Determine the affinity $d_{k.j} = 1/D_{k.j}$ among Ag_j and all the elements of C*:

$$D_{k.j} = \left\| C_k{}^* - Ag_j \right\|, \quad K=1,\ldots,N_c. \tag{3}$$

⑥ From C^*, re-select $\zeta\%$ of the antibodies with highest $d_{k.j}$ and put them into a matrix M_j of clonal memory;

⑦ Apoptosis: eliminate all the memory clones from M_j whose affinity $D_{k.j} > \sigma_d$:

⑧ Determine the affinity $S_{i.k}$ among the memory clones:

$$S_{j.k} = \left\| M_{j.i} - M_{j.k} \right\|, \forall i,k \tag{4}$$

Clonal suppression: eliminate those memory clones whose $S_{i.k} < \sigma_s$:

1) Concatenate the total antibody memory matrix with the resultant clonal memory for $Ab_{\{m\}} \leftarrow [Ab_{\{m\}}; M_j{}^*]$;

2) Determine the affinity among all the memory antibodies from $Ab_{\{m\}}$:

$$S_{i.k} = \left\| C_k - Ag_j \right\|, \forall i,k. \tag{5}$$

3) Network suppression: eliminate all the antibodies such that $S_{i.k} < \sigma_s$:

4) Build the total antibody matrix $Ab \leftarrow [Ab_{\{m\}}; Ab_{\{d\}}]$

(2) Test the stopping criterion.

To determine the total clone size N_c generated for each of the M antigens, the following equation was employed:

$$N_c = \sum_{i=1}^{n} round(N - D_{i.j}.N), \tag{6}$$

where N is the total amount of antibodies in Ab, $round(\cdot)$ is the operator that rounds the value in parenthesis towards its closest integer and D_{ij} is the distance between the selected antibody i and the given antigen Ag_j, given by Equation (1). In the above algorithm, Steps ① to ⑦ describe theclonal selection and affinity maturation processes as proposed by de Castro and Von Zuben (2000a) in their computational implementation of the clonal selection principle. Steps ⑧ to 3). simulate the immune network activity.

As can be seen by the aiNet learning algorithm, a clonal immune response is elicited to each presented antigenic pattern. Notice also, the existence of two suppressive steps in this algorithm (⑨ and (3.), that we call clonal suppression and network suppression, respectively. As far as a different clone is generated to each antigenic pattern presented, a clonal suppression is necessary to eliminate intra-clonal self-recognizing antibodies, while a network suppression is required to search for similarities between different sets of clones. After the learning phase, the network antibodies represent internal images of the antigens (or groups of antigens) presented to it.

The network outputs can be taken to be the matrix of memory antibodies' co-ordinates $(Ab_{\{m\}})$ and their matrix of affinity (S). While matrix $Ab_{\{m\}}$ represents the network internal images of the antigens presented to the aiNet, matrix S is responsible for determining which network antibodies are connected to each other, describing the general network structure.

3 Artificial Immune Network for Optimization

Castro and Timmis[12] considered the clustering problem as multi-apex optimization problem, and create the algorithm opt-aiNet which is used to multi-apex function optimization. It seeks answers to multi-apex function's optimization problems by adopting artificial immune network model. It can effectively work out the most partial apex data of the objective function, and also keeps the good characters of automatically adjust the number of groups and real-coded Bellowing, I'd like to express its main thought as well as control ideas.

3.1 Opt-aiNet

Firstly bring in a certain quantity of individuals (actual vector) in the objective function's definition area and workable areas to structure the artificial immune network. And then choose to clone those individual ones of every network to choose optimized partial solution. The exact ways are that firstly clone a certain number of ones with multiplication copy operator, then aberrant every clone body with aberrant operator, and also keep one origin individual in those clone group, so that to pick out the most adoptable clone body. If the clone ones have better adaptation than original one, then displace it. When the network is stable, let these individuals network interact with each other, and restrain those who have lower dependency than preset restrain threshold data through negative way, and preserve the left ones as memory units. At last induce new ones arbitrary. Repeat these steps, until reach the convergence standards. When finished, those memory ones are the partial best answers.

In the algorithm of opt-aiNet, adopt affinity and fitness to be iteration group's appraisal index at one time. But here, the affinity is different from "affinity" in common immunity theory. It uses euclidean distance between two individuals in the group to express to measure individual otherness in the group. This will help to take measures to keep group's diversity. This is especially important for multi-peak function optimize. Fitness indicates individual's objection function value and it is used to judge optimal solution's matching degree of individual and part. This is consistent with "affinity" in clone choose theory. During clone choose, it adopts mutation operator inverse to affinity to keep group's high fitness and algorithmic fast convergence. The convergence condition adopted by opt-aiNet algorithm is the quantity based on memory individual. If doing repression for one time, memory individual's quantity don't change. Then this shows that the Internet goes to be stable. In that way, the rest of memory individual is the solution of problem.

But when solving function of some peak value whose distribution is denser, it has the trend of early convergence. Meanwhile, because this algorithm is random in nature, restricted defect of search precision is caused by random factor's affection. Based on the defect of opt-aiNet algorithm at present, there has been more improved algorithm. The algorithm as following:

In order to present an optimization version of opt-aiNet assume the following terminology:

Network cell: individual of the population. In this case no encoding is performed, each cell is a real-valued vector in an euclidean shape-space;

Fitness: fitness of a cell in relation to an objective function to be optimized (either minimized or maximized). The value of the function when evaluated for the given cell;

Affinity: euclidean distance between two cells;

Clone: offspring cells that are identical copies of their parent cell. The offspring will further suffer a somatic mutation so that they become variations of their parent.

The optimization version of opt-aiNet can be summarized as follows:

(1) Randomly initialize a population of cells (the initial number of cells is not relevant).

(2) While stopping criterion is not met do

① Determine the fitness of each network cell and normalize the vector of fitnesses.

② Generate a number Nc of clones for each network cell.

③ Mutate each clone proportionally to the fitness of its parent cell, but keep the parent cell. The mutation follows Eq. (7).

④ Determine the fitness of all individuals of the population.

⑤ For each clone, select the cell with highest fitness and calculate the average fitness of the selected population.

⑥ If the average error of the population is not significantly different from the previous iteration, then continue. Else, return to step ①.

⑦ Determine the affinity of all cells in the network. Suppress all but the highest fitness of those cells whose affinities are less than the suppression threshold σs and determine the number of network cells, named memory cells, after suppression.

⑧ Introduce a percentage d% of randomly generated cells and return to step (2).

(3).EndWhile

$$\begin{aligned} c' &= c + \alpha N(0,1) \\ \alpha &= (1/\beta)\exp(-f^{*}) \end{aligned} \tag{7}$$

3.2 Dopt-aiNet

In order to enhance search capacity and quicken algorithmic search speed, Fabricio Olivetti de Franca and so on put forward dopt-aiNet[13] .They perfected the network in many aspect on the base of opt-aiNet and apply it to the optimize problem of dynamic function. The improvement of dopt-aiNet algorithm mainly has following several points:

(1) In order to avoid slowing of algorithmic execution speed due to Internet cell's excessive growth, it divides Internet cell into current group and memory group in the arithmetic of dopt-aiNet. Current group is for keeping newborn cells while memory group is for the cells which are not evolutionary and possible to achieve part extremis by some time.

(2) Adopt the way of golden section to partition value range of mutation step parameter and calculating respectively the results of mutation. After that, choose better parameter value.

(3) Put forward two mutation operations of one-dimensional mutation and gene duplication and operating on each dimension of Internet cell. This is comfortable to fine search.

(4) The way of cell linearity compression replaces original compression way based on euclidean distance. This is easy to set compression valve value.

(5) Set the upper limit in Internet, which can group memory cells.

In this literature, it meanwhile indicates that the mutation step way of golden section can not get whole optimal solution in the problem of multimodal optimization and increase the time of calculation. Besides, because the two new mutation operation is about each dimension's operation of Internet cell, so the computation is very high to problems of high dimension.

4 Conclusion and Future Research

Many research and applied practice show immunity system itself is a information processing system which has good performance. And also prove we can make use of above discussed model and arithmetic to carry out these fine characteristic and solve actual problems. Moreover, the application of Internet model in actual problems is not only limit to the information processing problem such as data analysis, data dig and the multi-peak function optimization, but also applied to computer safety, automatic control system's design, optimizing neural network and many fields. They are the mechanism based on immunity clone, immunity memory and important characteristic, only that the focus of using characteristic and the form of tectonic model and arithmetic are different. But, the artificial immunity internet model's research is still in the start phase. People still design different model and arithmetic aiming at different problems and mostly, they still are probing. In general, the research of artificial immune network models is lack of effective theory guidance and uniform criterion and does not come into being integral system configuration. If we use immunity system's apocalypse in deed, it would have plentiful work to do.

In the actual project, many problems can be abstracted to be the problem of objection function's optimization like complicated system parameter and configuration identification and so on. In the actual application, in order to get more choose or all-round information, It not only require to get the whole optimal solution of objection function, but also require to get more part optimal solution as far as possible. At present, there is some deficiency of opt-aiNet and dopt-aiNet algorithm when they used to solve some function. In the project application problem of the no-linearity, combination, restrict are not perfect. We need to research farther.

References

[1] N. Jerne, "Towards a Network Theory of the Immune System", Ann. Immune. (Inst. Pasteur), pp.373-389, 1974

[2] D. Dasgapta, "Artificial Immune System and their Applications [M]", New York, Springer, 1998

[3] De Castro and J. Timmis. "A New Computational Intelligence Approach", Springer -Verlag, 2002

[4] JuanCG, AngelicaV, FabioA, "A comparative analysis of artificial immune network models[C]", GECCO'05, pp. 361-368, 2005

[5] J.E Hunt, D.E.Cook, "Learning Using an Artificial Immune System Journal of Network and Computer Applications", pp. 189-212, 1996

[6] J.Timmis, M . Neal, J.Hunt, "An Artificial Immune system for Data Analysis", BioSystems, 55: 143–150, 2000

[7] J .Timmis, MNealA, "Resource Limited Artificial Immune System for Data Analysis", Knowledge-Based Systems, 14: 121- 130, 2001

[8] J Timmis, A watkins, "Artificial Immune System: Revisions and Refinements Artificial Immune System (Timmis eds)", Berlin Heidlberg, Springer-Verlag, 2003

[9] O.Nasraouietal, F.Gonza'lez, D.Dasgupta, "The Fuzzy and Application to Clustering and Web Profiling", Inte rnational Conference on Fuzzy Systems, pp. 711-716, Hawaii, HI, May 2002

[10] De Castro, Fernando José Von Zuben, "aiNet: An Artificial Immune Network for Data Analysis", Idea Group Publishing, USA, March, 2001

[11] De CastroIN, TimmisJI, "An Artificial Immune for Multimodal Function Optimization", Proceedings of IEEE Congress on Evolutionary Computation (CEC'02), HaWaii, vol. pp. 699-674, 2002, May

[12] Fabricio Olivetti de França, "An Artificial Immune Network for Multimodal Function Optimization on Dynamic Environments", IEEE, 2006

A Novel Fuzzy Cognitive Map Approach to Differential Diagnosis of Specific Language Impairment

Feng Bin   Wang Zhang

School of Information Technology Jiangnan University Wuxi, Jiangsu P.R China 214122
Email: jsufeng@vip.sina.com

Abstract

A technology for Fuzzy Cognitive Maps learning, which is based on the Quantum-behaved Particle Swarm approach is introduced. The proposed approach is used for the nonzero weight values that lead Fuzzy Cognitive Maps to desired steady states. The working of the approaches are applied to the model for differential diagnosis of specific language impairment (SLI).The methodology is based on fuzzy cognitive maps with Quantum-behaved Particle Swarm algorithm. The development of the model was based on knowledge from the literature and then it was successfully tested on two clinical cases. The results obtained point to its final integration in the future and to its valid contribution as a differential diagnosis model of SLI.

KeyWords: PSO (Particle Swarm Optimization), QPSO (Quantum-Behaved Particle Swarm Optimization), fuzzy cognitive maps, weight matrices, object function.

1 Introduction

The representation of causal relationships and reasoning are important research fields of AI(artificial intelligence).[1]Political scientist Robert Axelrod originally proposed Cognitive Maps in 1976.Cognitive Maps(CM) are a useful model to represent concepts or variables in a given domain and their causal-effect relations[2]. The concepts in a given domain and their causal-effect relationships between these concepts are represented as edges. Kosko enhanced the power of the fuzzy cognitive maps and fuzzy degrees of interrelationships between introduced in 1986.

Speech assessment is a procedure, which should include a complete history of each patient, diagnostic test to examine all of the aspects of speech, language and communication in general, as well as a detailed observation of the patient over a long period of time. However, in many cases there are similar symptoms that correspond to a group of disorders [3]. Thus, the differential diagnosis has to determine which the most probable disorder is and the goal of this study is to offer a model of differential diagnosis in order to facilitate this process [4].

In this paper, an approach for FCMs learning, based on QPSO algorithm, is presented .QPSO is applied to update the weight values of the FCM, so that leads the FCMs to a desired steady state. This paper also presents some basic factors that appear in all three disorders (SLI, dyslexia and autism) with different frequency and severity in most cases. The considered factors are either causative factors or symptoms of the disorders. A detailed and in-depth analysis of the factors is not within the scope of this work. Instead, the development of an advanced system is discussed that is capable of contributing to the differential diagnosis of SLI from other disorders, taking into account the factors that are involved in each disorder. The factors considered in the model, are those that have been found to play an important role in the diagnosis of all three disorders. Some other factors (e.g. memory, auditory processing and orientation in space and time) will be included in the next phase of the study, because these factors need more investigation.

2 Particle Swarm Optimization algorithm and Quantum-Behaved Particle Swarm Optimization algorithm

2.1 particle swarm optimization algorithm

Particle swarm optimization is an evolutionary computation technique developed by Dr Eberthart and Dr Kennedy in 1995, inspired by social behavior of bird flocking or fish schooling .PSO is similar to the other evolutionary algorithm in that the system is initialized with a population of random solutions. However, each potential solution, call particles, flies in the D-dimensional problem space with a velocity which is dynamically adjusted according to the flying experience of its own and its colleagues. The location of the i-th particle is represented as $X_i=(x_{i1},x_{i2},\dots x_{iD})$.The best previous position of the i-th particles is represented as $P_i=(p_{i1,}p_{i2},\dots p_{id})$,

Which is also called pbest, the index of the best particle among all the particles in the population is represented by the symbol g_i, the location P_{gi} is also called gbest. The velocity for the i-th particle is represented as $V_i = (v_{\ i1,}v_{i2,\dots}v_{id})^T$,. Then the swarm is manipulated by equation

$$\begin{aligned} v[i] = w *v[i] + c1 * rand() *(pbest[i]-present[i]) \\ + c2 * rand()*(gbest[i]- present[i]) \end{aligned} \quad (1)$$

$$present\ [i] = persent[i] + v[i] \quad (2)$$

Where w is inertia weight, c1 and c2 are acceleration constants, r1 and r2 are a random function in the range [0,1], Generally, the default values c1 = c2 = 2 have been proposed ,and the parameter w decrease linearly from 0.9 to 0.4.

2.2 quantum-behaved particle swarm optimization algorithm

In PSO algorithm, the i-th particle is depicted by its position vector X_i and velocity vector V_i, which determine the trajectory of particle. The particle moves along a determined trajectory in classical mechanics, but this is not trajectory is meaningless, because X_i and V_i of a particle cannot be determined simultaneously according to uncertainty principle. Therefore, if individual particles in a PSO system have quantum behavior, the PSO algorithm is bound to work in a different fashion .In [5,6], Jun Sun et, introduce a Quantum-behaved PSO(QPSO) algorithm[7][8]. The experiment results indicate that the QPSO works better than standard PSO on several benchmark functions. In QPSO algorithm, only position vector p is needed to depict a particle, and there is only one parameter β .The equation is as follows:

$$mbest = \frac{1}{m}\sum_{i=1}^{M} P_i = \left(\frac{1}{M}\sum_{i=1}^{M} P_{i1}, \frac{1}{M}\sum_{i=1}^{M} P_{i2}, \dots \frac{1}{M}\sum_{i=1}^{M} P_{id}\right) \quad (3)$$

$$P_{id} = \phi * P_{id} + (1-\phi) * P_{gd} \qquad \phi = rand \quad (4)$$

$$x_{id} = P_{id} + \beta * \left|mbest - xid\right| * \ln\left(\frac{1}{u}\right) \qquad \mu = rand \quad (5)$$

φ is a random function in the range[0,1], pi is the best position of particle i ,P_g is the position of the globe point, denoted as mbest, is defined as the mean of Coefficient, M is the population size and u is a random function in the range [0,1],P_i is the best position of all particle among all the particles in the population, the globe point, denoted as mbest, is defined as the mean of pbest positions of all particles. β is called Creativity Coefficient, M is the population size and u is a random function in the range[0,1]. In the process of iteration, ± is decided by the random number, when it is bigger than 0.5, minus sign (-) is proposed, other plus sign(+) is proposed.

3 Fuzzy Cognitive Maps

3.1 fuzzy cognitive maps

Fuzzy Cognitive Maps are soft computing tools. Concepts are pictured different aspects of the system and their behavior, and the dynamics of the system are represented by the interaction of concepts. An FCM models consists of nodes-concepts, $C_i(i=1,2,3\dots N)$, where N is the total number of concepts. Each node-concept represents one primary factor of the system and it is depicted by a value $A_i \in [0,1](i=1,2\dots N)$. The weight, W_{ij}, indicates whether the relation between the two concepts is positive or negative. The direction of causality demonstrates whether the concepts C_i causes the concept C_j or reverse. Thus, the values of weights are in continuum [-1, 1].

The values A_i of a concept C_i is influenced by the values of concepts-nodes connected to it, and are updated according to the equation:

$$A_i(k+1) = f(A_i(k) + \sum_{j=1}^{n} W_{ij} A_j(k)) \qquad (6)$$

Where k stands for the iteration counter; and W_{ji} is the weight of the arc connecting concept C_j to Concept C_j The function f is sigmoid function:

$$f(x) = \frac{1}{1+e^{-\lambda x}} \qquad (7)$$

Where λ>0. In the present study the value of was set to 1.The function can restrict the values A_i of the concepts within [0,1]. The interaction of the FCM results after a few iterations in a steady state, i.e. the values of the concepts are not modified further.

After the determination of FCM's structure, and using the initial concept values A_i and the matrix W_{inital}, which are provided by the experts, then the FCM is let to converge to a steady state through the application of Eq(1).

The heavy dependence on the expert's opinion regarding the FCM's design and the convergence to undesired steady states starting from the expert's recommendations are the two most significant weakness of FCMs .However, we apply the new technique to update the weight matrix of FCM so as to avoid convergence to undesired steady states.

3.2 the new learning approach

The purpose of new approach is to search a proper weight matrix W = $[W_{i,j}]$(i,j = 1,2,…N), and then the FCM is let to converge to a steady state.Set C_i,…,C_N be the concepts of an FCM, and C_{out1} ,…,C_{outm} (1<= m <=N) be the output concepts, while the remaining concepts are considered input or interior concepts. The output concepts keep in strict bounds:

$$A_{out_i}^{\min} \le A_{out_i} \le A_{out_i}^{\max}, i = 1,2...m$$

Thus, the objective function is considered:

$$F(W) = \sum_{i=1}^{m} H(A_{out_i}^{\min} - A_{out_i}) \mid A_{out_i}^{\min} - A_{out_i} \mid + \sum_{i=1}^{m} H(A_{out_i} - A_{out_i}^{\max}) \mid A_{out_i} - A_{out_i}^{\max} \mid \qquad (8)$$

Where H is the famous Heaviside function

$$H(x) = \begin{cases} 0, x < 0, \\ 1, x \ge 0, \end{cases}$$

A_{outi} (i = 1,2,…m) are the steady state values of the output concepts, which are obtained through the application of Eq(6).Obviously, the globe minimization of the objective function F is weight matrices that lead FCM to a desired steady state. The application of QPSO for the minimization of the objective function F starts with an initialization state, where a swarm of M particles is generated randomly, and it is evaluated using F. Then Eq(5),Eq(6) and Eq(7) are used to evolve the swarm. When a weight configuration that globally minimizes F is reached, the algorithm stops.

4 Sli Differential Diagnosis Model

The proposed FCM, depicted in Figure 1, consists of two different types of concepts. The three central concepts (disorder concepts) correspond to the three disorders. The factors presented belong to the second type of concepts, factor concepts which are symptoms and cause factors to the disorder concepts, and they are considered as measurements that can determine the result of the diagnosis. The direction of interconnections between the concepts is shown in Figure 1 by the arrowed arcs. This shows in a simple way which concept influences another concept. However, due to limited space and in order to make the figure simpler, the sign and weights of the connections are not illustrated in Figure 1. These are extracted by assigning the qualitative (linguistic) values: very-very high, very-high, high, etc. to the importance of each diagnostic criterion, respectively. These connections may show a positive or negative dependence between factors and disorders. A positive connection (+) implies that the given factor increases the probability of diagnosis of the connected disorder. Lack of connection between a factor and a disorder suggests that no influence of that factor on the disorder has been found, yet. A negative (-) connection between the factor and the disorder (such as reading ability and autism) implies that the existence of the given factor must lead to reduction of the probability of diagnosing the particular disorder.

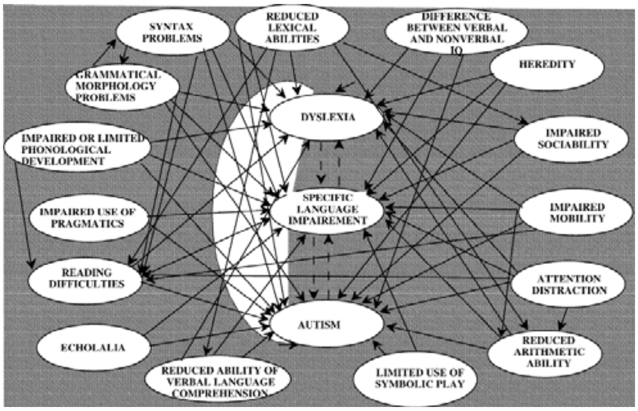

SYNTAX PROBLEMS
REDUCED LEXICAL ABILITIES
DIFFERENCE BETWEEN VERBAL AND NONVERBAL IQ
HEREDITY
GRAMMATICAL MORPHOLOGY PROBLEMS
DYSLEXIA
IMPAIRED OR LIMITED PHONOLOGICAL DEVELOPMENT
IMPAIRED SOCIABILITY
SPECIFIC LANGUAGE IMPAIREMENT
IMPAIRED USE OF PRAGMATICS
IMPAIRED MOBILITY
READING DIFFICULTIES
ATTENTION DISTRACTION
AUTISM
ECHOLALIA
REDUCED ARITHMETIC ABILITY
REDUCED ABILITY OF VERBAL LANGUAGE COMPREHENSION
LIMITED USE OF SYMBOLIC PLAY

Figure 1 FCM differential diagnosis model of SLI from dyslexia and autism.

Apart from describing the direction of causality between two concepts and the sign of causality, the degree of cause and effect between two concepts must be determined, since we do not expect that all factors have the same weight for a given disorder, nor the same weight for each disorder. Each expert describes the degree of influence for each interconnection using a linguistic variable. Thus, each expert of the group of experts suggests a linguistic weight for each interconnection, so a set of linguistic weights for each interconnection is assigned. This set of weights for each interconnection is integrated, using a sum combination method and then the defuzzification method of center of area (CoA) is used and a numerical weight for this interconnection is produced, which belongs to the interval [-1, 1]. In this first phase of the research, published research results have been used as ‘‘experts’’ and these were integrated using the procedure described above. The allowable linguistic variables for this application may belong to the fuzzy sets described below. Each fuzzy set corresponds to a membership function shown in Figure 2. Seven membership functions are suggested to describe the degree of influence, giving the Possibility to the experts to describe in detail the influence of one concept to another:

_ M(very-very low): the fuzzy set for influence around 10%withmembership function μ_{vvl} .

_ M(very low): the fuzzy set for influence around 20% with membership function μ_{vl} .

_ M(low): the fuzzy set for influence around 35% with membership function μ_l ..

_ M(medium): the fuzzy set for influence around 50% with membership function μ_m .

_ M(high): the fuzzy set for influence around 65% with membership function μ_h .

_ M(very high): the fuzzy set for influence around 80% with membership function μ_{vh} .

_ M(very-very high): the fuzzy set for influence around 90% with membership function. μ_{vvh}

The membership functions are not of the same size since it is desirable to have finer distinction between grades in the lower and higher end of the influence scale.

Another new consideration is that in the FCM in which there are nodes that do not accept

Feedback, it is important not to allow the values of those nodes to change. In order for this to be achieved, a check should be made of each node to examine if it accepts inputs from other nodes. If not, then a self-feedback value of the node should be set at 1 and the value of that node after each repetition should remain the same. Therefore, the algorithm is as follows.

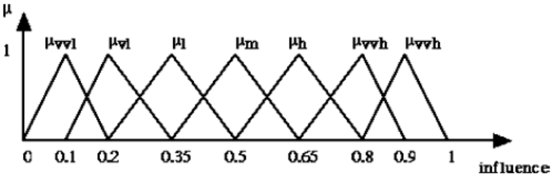


Figure 2 Membership function

Fuzzy cognitive map that models and controls this system is depicted on Figure 1. It consists of 18 concepts that are defined as:

1) Concept1 – the degree of Dyslexia

2) Concept2 – the degree of SLI

3) Concept3 – the degree of Autism

4) Concept4 – the degree of reduced lexical

abilities

5) Concept5 – the degree of Problems in syntax

6) Concept6 – the degree of Problems in grammatical morphology

7) Concept7 – the degree of Impaired or limited phonological development

8) Concept8 – the degree of impaired use of pragmatics

9) Concept9– the degree of Reading difficulties

10) Concept10– the degree of Echolalia

11) Concept11 – the degree of reduced ability of verbal language comprehension

12) Concept12 – the degree of Difference between verbal and non-verbal IQ

13) Concept13 – the degree of Heredity

14) Concept14 – the degree of Impaired sociability

15) Concept15 – the degree of impaired mobility

16) Concept16 – the degree of Attention distraction

17) Concept17 – the degree of reduced arithmetic ability

18) Concept18 – the degree of Limited use of symbolic play

The weight value of the initial weight matrix is:

$$W = \begin{bmatrix}
\text{sli\_f} & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & \text{dys\_f} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & -1 & \text{aut\_f} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.9 & 0.58 & 0.9 & 1 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0.2 & 0 & 0 & 0.2 & 0 & 0 & 0.2 & 0 \\
0.9 & 0.58 & 0.8 & 0 & 1 & 0.2 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.9 & 0.58 & 0.8 & 0 & 0.2 & 1 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.65 & 0.9 & 0.8 & 0 & 0 & 0 & 1 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.58 & 0 & 0.9 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.58 & 0.9 & -0.65 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.2 & 0 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.8 & 0.35 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.9 & 0.9 & 0.58 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0.58 & 0.70 & 0.65 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0.58 & 0.8 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0.8 & 0 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0.2 & 0 \\
0.5 & 0.58 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.2 & 0 \\
0.58 & 0. & 0D & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0.5 & 0 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}$$

Experts determine the direction of the arcs and the weight value among the concepts. The ranges of the weight implied by the fuzzy regions are:

$$w_{41}w_{43}w_{51}w_{61}w_{72}w_{83}w_{92}w_{10,3}w_{11,3}$$

$$w_{12,1}w_{12,2}w_{14,3}w_{16,3}w_{18,3} \in [0.85, 0.95]$$

$$w_{42}w_{52}w_{62}w_{81}w_{91}w_{12,3}w_{13,1}w_{14,1}w_{16,2}w_{17,6} \in [0.55, 0.6]$$

$$w_{53}w_{63}w_{73}w_{15,1}w_{14,2} \in [0.75, 0.83]$$

$$w_{71}w_{93}w_{13,3} \in [0.60, 0.68]$$

$$w_{16,1}w_{18,1} \in [0.47, 0.55]$$

$$w_{11,2} \in [0.32, 0.37]$$

$$w_{13,2} \in [0.68, 0.72]$$

Encode the thirty-nine weights into the vector of the particle, PSO or QPSO algorithm is used to update these values of the FCMs. The bounds implied by the direction of the corresponding arcs of the FCMs are imposed on these weights.

The output concepts for this problem are the concepts Concept1, Concept2, Concept3.The experts have defined the desired regions for these concepts:

$$C_{MINi} \le C_i \le C_{MAXi} (i = 1, 2, 3)$$

The minim and the maximum of the Concept i is decided by the experts.

5 Confirmations Of Results Of The Model For Two Clinical Cases

After the construction of the above differential diagnosis model, two case studies from the literature were investigated (on SLI), in order to confirm its effectiveness. The value of occurrence of each factor in each case study is denoted with similar qualitative degrees, as shown in Table 2. For the cases that the value of a concept factor is 0, it denotes that either there was no information supplied on the given factor or that the given symptom did not exist. The following initial vectors of concepts values are used for each one of the four cases; their values are produced implementing using the defuzzification method of CoA on the linguistic values of Table2.

Table 2 weight of each vector for clinical case

Number	Factor concepts	Case1	Case2
1	Reduced lexical abilities	Very-very high	High
2	Problem in syntax	Very-very high	Very high
3	Problem in grammatical morphology	Very high	Very high
4	Impaired or limited	0	0
5	Reading ability of verbal	0	0
6	Reading difficulties	0	0
7	Echolalia	0	0

(Continued)

Number	Factor concepts	Case1	Case2
8	Reduced ability of verbal language comprehension	0	0
9	Difference between verbal and non-verbal IQ	High	High
10	Heredity	High	0
11	Impaired sociability	Medium	0
12	Impaired mobility	0	0
13	Attention distraction	0	0
14	Reduced arithmetic ability	0	0
15	Limited use of symbolic play	0	0

respectly:

$$A_0 = [0\ \ 0\ \ 0\ \ 0.9\ \ 0.9\ \ 0.8\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0.65\ \ 0.65\ \ 0.5\ \ 0\ \ 0\ \ 0\ \ 0\,]$$

$$A_0 = [0\ \ 0\ \ 0\ \ 0.8\ \ 0.9\ \ 0.9\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0.65\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0]$$

In our experiment, a total of 100 independent experiments have been performed using QPSO algorithm, and the obtained results compared with the results of using PSO algorithm.

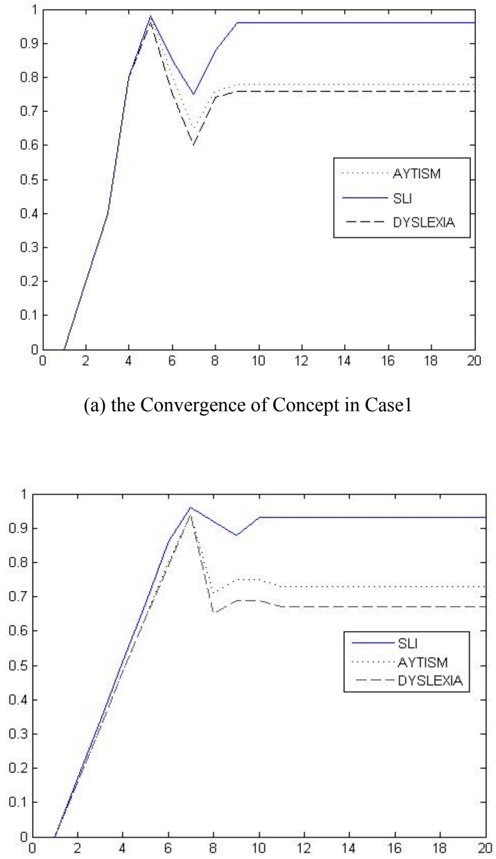
The parameters of PSO algorithm are defined as following: the swarm size is set to 30, the default values c1 = c2 = 2 have been proposed, parameter w decrease linearly from 0.9 to 0.4, the number of iterations required is 40. The parameter of QPSO algorithm is defined as follows:

The swarm size is set to 30, the parameter β decreases from linearly from 1.0 to 0.5, the number of iterations required is 20.the minimize and maximum of the Concept i (i=1, 2,3) as follows:

Table 3 minimum and maximum of Concept determined by experts

	Case1		Case2	
	Min	Max	Min	Max
Concept 1	0.92	0.98	0.93	0.98
Concept 2	0.73	0.80	0.62	0.7
Concept 3	0.78	0.82	0.78	0.80

Figure 3 contains plots of the values of the output nodes, SLI, dyslexia and autism as a function of the number of repetitions for each case. Each node converges to a final value and the node with the maximum value is the most probable diagnosis based on the model. In two cases, even though the information was incomplete, the result given by the model agreed with the published diagnosis. That is in two cases, the correct diagnosis was concluded: SLI, SLI.

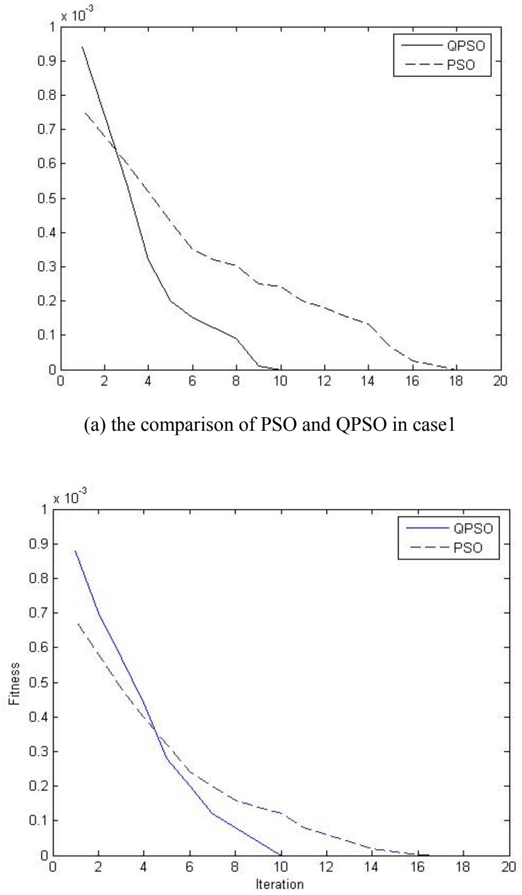


(a) the Convergence of Concept in Case1

(b) the convergence of Concept in Case2

Figure 3 Result of two clinical case

The convergent process of objective function F is illustrated in Figure 4. We can obtained the conclusion from the Figure 4 that the velocity of QPSO algorithm is much faster than that of PSO algorithm. The obtained weight matrixes lies in expert's bounding regions. It is clear that the learning algorithm is capable of providing proper weight matrices for the FCM, and alleviating deficiencies caused by deviation in the expert's suggestions.



(a) the comparison of PSO and QPSO in case1

(b) the comparison of the PSO and QPSO in case2

Figure 4 the comparing of two optimization algorithm

6 Conclusion

QPSO algorithm is the expanding of the standard PSO algorithm in quantum space, and has the virtues of briefness and less parameter. The experiment results indicate that the control problem and QPSO is a promising algorithm.

The ultimate goal of this effort is to develop a sufficient estimation model that can reliably be used complementary to other diagnostic methods to assist the speech pathologist in cases of language and communication disorders that are difficult to discern. Even though this effort is in its initial stage, we hope that when successfully completed it will contribute to the field of differential diagnosis in speech and language pathology.

Reference

[1] Anastasiou D. Dyslexia. Theoria kaiereyna—opsis praktikis [Dyslexia. Theory and research—practical views]. Athens: Atrapos; 1998 [in Greek]

[2] Bishop DVM. What is so special about Asperger's disorder? The need for further exploration of the borderlands of autism. In: Klin A, Volkmar F, Sparrow S, editors. Asperger syndrome. New York: Guilford Press; 2000

[3] Bishop DVM. Handedness, lumsiness and developmental language disorders. Neuropsychologia 1990; 28:681–90

[4] SUN J, FENG B, XU WB. Particle Swarm Optimization with Particles Having Quantum Behavior [A]. Proceedings of 2004 Congress on Evolutionary Computation[C].2004. 325-331

[5] SUN J, XU WB, FENG B. A Global Search Strategy of Quantum-Behaved Particle Swarm Optimization [A]. Proceedings of IEEE conference on Cybernetics and Intelligent Systems[C]. 2004. 111-116

[6] Tsujimura Y, Mafune Y, Gen M. Effects of Symbiotic Evolution in Genetic Algorithms for Job-Shop Scheduling [A]. Proceedings of the 34th Hawaii International Conference on System Sciences[C]. 2001. 1-7

[7] Madureira A, Ramos C, do Carmo Silva S.D. A Coordination Mechanism for Real World Scheduling Problems Using Genetic Algorithms [A]. Proceedings of the 2002 Congress on Evolutionary Computation[C]. 2002. 175-180

[8] Clerc M, Kennedy J. The particle swarm: explosion, stability, and convergence in a multi-dimensional complex space [A]. IEEE Trans on Evolutionary Computation[C]. 2002.6(1):58-7

A Formal Description and Verification of Authentication Protocol

Zhanting Yuan Xu Kang Qiuyu Zhang Shuang Liang

College of Computer and Communication, Lanzhou University of Technology, Lanzhou, Gansu, 730050, China

Email:kxx1983@mail2.lut.cn

Abstract

Identity authentication, as the first step to realize the network security, is the key technology for secure exchange of the online commercial information. Moreover, if there is any leak in the authentication protocol, secret information will leak out consequentially. It is necessary to adopt some formalized method to describe and prove the authentication protocol. In this paper the colored Petri Net is employed to describe the authentication protocol, and simultaneously the combined strategy of *1*-reachable analysis and conversing analysis has been put forward to verify the security of this authentication protocol.

Keywords：Identity Authentication Protocol; Petri Net; Reachable Tree; Reachable Analysis; Security

1 Introduction

With the extensive application of computer networks, e-commerce, e-government, e-banking, e-stocks and other commercial activities have been developing quickly. In any online commercial activity, the identity authentication of users is a very important issue. However, most of identity authentication protocols have been found to be flawed after being used for some time. How to ensure that there is no leak in the authentication protocol is one of key factors for the security of online commercial activities. Most of protocols are described by using diagrams or natural languages. In those descriptions, the dynamic characteristics of the protocols can not be seen intuitively, and the usable resources for users can not be shown, and of course the non-formal description method has no advantage to prove the security. One kind of structure of identity authentication and authentication protocol has been proposed in[1]. In this paper the specific authentication procedure of this protocol is given, which is described formally by using the Petri net and verified by proposed combing strategy of *1*-reachable[2] analysis and conversing analysis.

This authentication method can realize the simple verification of the protocol analysis and solve the problem of explosion of the state space also.

2 Authentication Protocol Design

This authentication protocol[3] is based on the traditional dynamic password authentication mechanism of the request/response authentication methods and the transfer security has been improved in[4]. More than that, the public key certificates and public and private key pair have been added, so two-way authentication of a client and server has been achieved.

After finishing the initialization, the user has to pass the authentication of the authentication sever before visiting the resources of the resource sever. The user can use authentication information in Authentication Token to express his identity to the authentication sever of the system. Four following information will be exchanged between the client and authentication server to complete two-way identity verification.

For convenient description, the following remarks have been introduced:

- A is a client, and B is an authentication server;
- k_x k_x^{-1} are the public key, private key of the main body x respectively;
- $k_x(\ )$ and $k_x^{-1}(\ )$ are encryption data and decryption by using k_x k_x^{-1} respectively;
- c_x is the public key certificate of x;
- $UseID$, Psw are the user name and password, sent to the authentication server by the user, which are waiting to be verified;
- R_X is random number generated by x, R_B is the random number generated by authentication server used as authentication challenges;
- EA is the list of encryption algorithms provided to the server from client side;
- EA_C is the list of encryption algorithms chosen by the sever B;
- $rule()$ is the computing rule by using random method;
- $seed$ is the seed value of the authentication token;
- N_0 is the information of verified results sent to the client by authentication server.

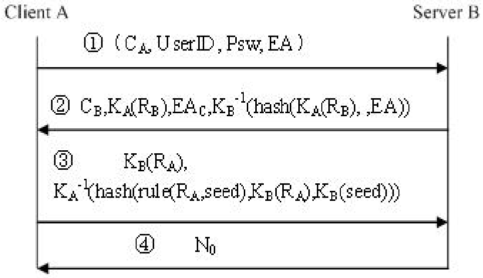

Client A
Server B
① （C_A, UserID, Psw, EA）
② C_B,$K_A(R_B)$,EA_C,K_B^{-1}(hash($K_A(R_B)$, ,EA))
③ $K_B(R_A)$, K_A^{-1}(hash(rule(R_A,seed),$K_B(R_A)$,K_B(seed)))
④ N_0

Figure 1 Identity Authentication System interactive information flow

The authentication procedure of this authentication protocol is specified as follows:

（1）When user logs in to the client resource server and issue resource access requests, the system prompts the user to input a user name and password, and then send (UserID, Psw) as well as the certificate C_A of client A and EA to authentication server;

（2）The authentication server first verifies user name and password;

A) If correct, the authentication server will verify the client's certificate.

1) If passed, B and A will execute the two-way authentication based on the authentication protocol of dynamic password mechanism. B receives the public key C_A from A and then generates a random number R_B, then uses C_A to encrypt R_B and chooses an encryption algorithm to calculate (hash (K_A (R_B), and EA). C_B, K_A (R_B), EA_C, K_B^{-1} (hash (K_A (R_B), EA)) are sent to the client as a data package of authentication for challenge. The random number is kept in the database.

2) If failed, the authentication fails;

B) If not correct, the sever transmits back prompt information to the user to re-enter the login information.

（3）After receiving the data package of C_B,$K_A(R_B)$, EA_C,K_B^{-1}(hash($K_A(R_B)$,EA)), if the head information of the package is judged as authentication information then the client will send the authentication request to the authentication token; after the authentication token receives the certification request, the system prompts the user to input the password of private key protection, then the authentication token will verify B's certificate first;

A) If C_B passes through verification, A will receive the public key of B from C_B, and use it to verify the B's signature. By using A's private key to decrypt R_B then R_A can be derived. After that the randomized rule is called to generate rule (R_A, seed) and the signature program is called to sign authentication information, user name and password, then to form the message $K_B(R_A)$, K_A^{-1}(hash (rule(R_A,seed),$K_B(R_A)$,K_B(seed))) which is used as a response back to authentication server B.

B) If C_B can not pass through verification, that means the authentication server is fake and the authentication of the client to authentication server fails.

（4）When authentication server B receives the $K_B(R_A)$, K_A^{-1}(hash(rule(R_A,seed), $K_B(R_A)$,K_B(seed))) message, according to the user name the sever will check out seed value and the stored random number, then verify the correctness of the signature and send back the verification result to the client A. The certification process is completed.

A) When the user is legal, the control agent module of authentication server will transfer the request launched by the client to the resource sever and build up a transparent proxy between the client and the resource sever, so the client can access resources on this resource server.

B) If the user is illegal, an authentication failure will be caused, so the client request will be declined.

3 Formal Description and Verification Based on Petri-net

3.1 Colored Petri Net

As a mathematical tool, colored Petri net[5] is widely used in communication protocols, operating systems, hardware systems, embedded systems and software designs in[6]. Like basic Petri nets, colored Petri net has both of graphic and language descriptions, between them the conversion can be implemented by using some tool. The Petri net with graphic description is used to describe the identity authentication protocol in this paper. In the colored Petri net with graphic description, the basic concepts are mentioned below:

（1）Place, presented as an oval that describes system state.

（2）Transition, presented as rectangle, describes system activity.

（3）Arc, as arrow to describe the change of the system state when a transition occurs.

（4）Token (token). There is a set of tokens in each place and each token contains a set of data elements of given type.

More details about the colored Petri net[7] can be referred to. Colored Petri net can not only accurately describe the certification process of the identity authentication protocol, but also have a strong capacity of bearing data. The dynamic characteristics of the protocol can be revealed in the model by the news and transitions completely.

3.2 Security protocol verification methods

In most of using of Petri net to verify the security protocol[8], it only verifies the correctness but not the security in. In this paper, based on the sufficient analysis of the characteristics of the protocol which is needed to be verified , a new security protocol authentication method with emphasis on the analysis the security of the proposed. When the protocol becomes very complicated, it is possible that the corresponding state space would explode if the method of reachable tree has been adopted only. This problem can be solved efficiently by using conversing analysis to analyze the security of the protocol. The problem can be solved efficiently also by using the proposed method which combines the conversing analysis and *1*-reachable analysis. Furthermore, this analysis method can realize the simplicity of protocol authentication and is easy-to-use.

Such authentication procedure is divided into the following steps basically;

（1）Build the model of colored Petri net for protocol analysis, and determine the variables which the intruder may change;

（2）Analysis the possible unsafe states of the model;

（3）Transfer the security issues into *1*-reachable issues of a single location in Petri net;

（4）Build the reachable tree and adopt the combined method of conversing analysis and *1*-reachable analysis. Now start from the conditions which ensure successful authentication to analyze the probability with that the intruder can attack successfully, then to state the security of the protocol waiting for analysis.

3.3 Description and verification of Identity Authentication Protocol Petri Net

Step 1:build the Petri net model of this identity authentication protocol, as shown in Figure 2, where

M_1:{ C_A, UserID, Psw, EA }

M_2:{ C_B,$K_A(R_B)$,EA_C,K_B^{-1}(hash($K_A(R_B)$, ,EA)) }

$M_2{}'$:{ $K_A(R_B)$,EA_C,K_B^{-1}(hash($K_A(R_B)$, ,EA)) }

M_3:{$K_B(R_A)$,K_A^{-1}(hash(rule(R_A,seed),$K_B(R_A)$,K_B(seed)))}

N_0: the news of verified results sent to the client by the authentication server.

The specific meaning of each transition is as follows:

t_1: the authentication request sent to the authentication server B by the client A;

t_2: the message M_1 sent to B by A;

t_3: M_1, received by B from A;

t_4: according to user name to verify the legality the of *(UserID, Psw*), if illegal, re-input;

t_5: to generate random number R_B, choose one algorithm in the list of encryption algorithm, and then encrypt C_B, R_B,, and other information to produce news M_2;

t_6: send message M_2;

t_7: execute the random computation with A’s seed and R_B stored in B;

t_8: A receives message M_2;

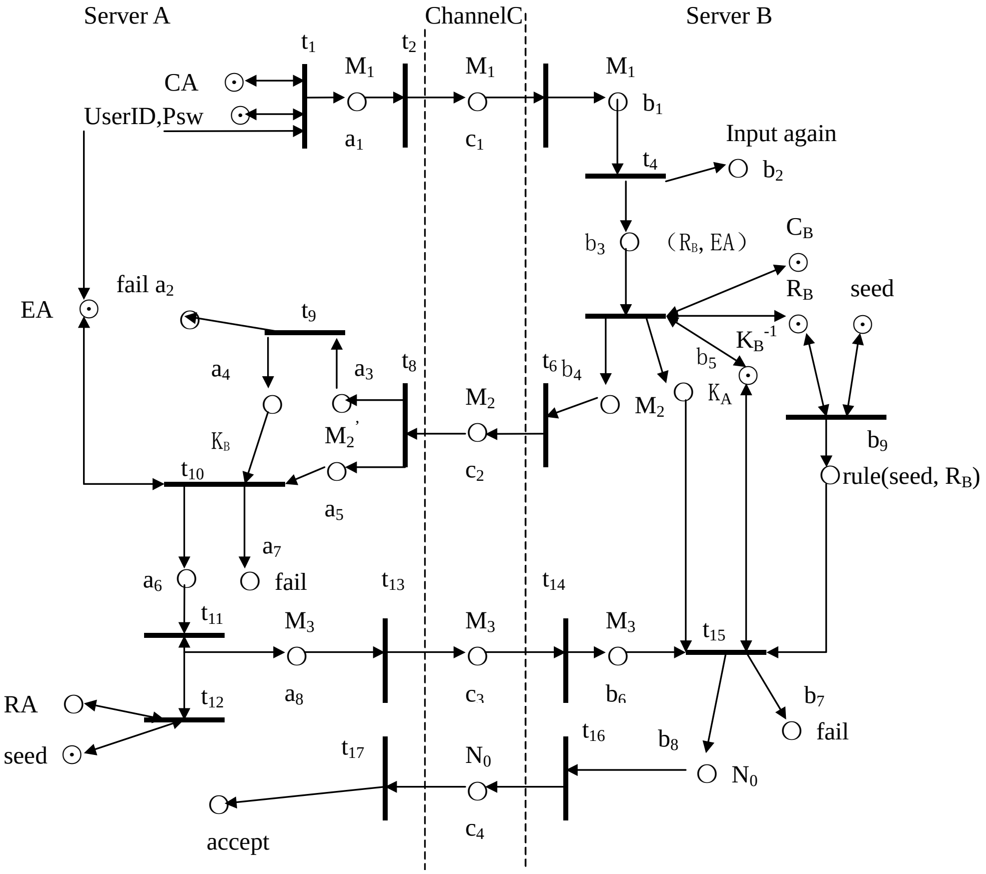


Figure 2 the Petri net description of identity authentication protocol

t_9: verify B’s certificate;

t_{10}: using B’s public key to verify if the hash value and $K_A(R_B)$ are equal. If not, that means M_2 has been faked;

t_{11}: encrypt the results of random computation and decryption result to produce message M_3;

t_{12}: execute the random computation with seed and R_A;

t_{13}: send message M_3;

t_{14}: receive message M_3;

t_{15}:compare result and the hash value. If they are same, the authentication is successful, otherwise fails.

Step 2: the intruder may pretend as A, or as B. The Petri net protocol description is as follow when the intruder pretends as A. as shown in Figure 4.

The transitions T of A and B have been introduced below, the transitions of invader H are introduced here only:

t_3: intercept the authentication request sent to B by A;

t_4: intruder sends faked message to B;

t_6: store the message of intercepted;

t_7: rejigger the message of intercepted and change

C_A to C_H;

t_8: encrypt useful information, and send the encryption information later;

t_{12}: intercept the message sent to A by B;

t_{13}:decrypt intercepted information to imply random number R_B;

t_{14}: send encrypted message to A;

t_{20}: intercept M_3 sent to B by A;

t_{21}: decrypt message by using the intruder’s private key;

t_{22}: encrypt the received information to produce message ~~M_3~~;

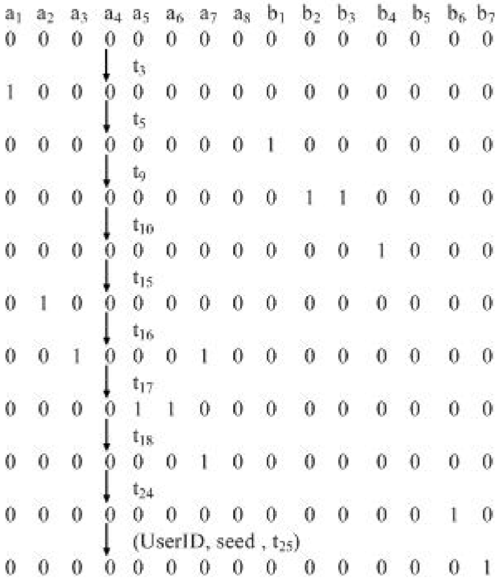
t_{23}: send ~~M_3~~ to B.

Step 3: When the intruder pretends as A, its specific reachable tree analysis is as shown in Figure 3. The analysis is similar if the intruder pretends as B:

The username and seed in the authorization server are matched one-to-one. If the intruder tries to pass authentication by using the client username and its own seed, b_7 receives Token when the transition t_{25} occurs, then the authentication fails. It means that, the procedure of that the intruder pretends as A fails[9].

C_B and K_B^{-1} are stored in the authentication server only which is in the protected sub-network of the system. The authentication server can be attacked with very low probability, or it has low-risk. The intruder can get C_B and K_B^{-1}hardly. For the seed is only stored in the authentication token of the client or authentication server, but not transmitted through the network, it is impossible for the intruder to get the seed in[10]. Therefore, the message can not reach a_8 through t_{25}, the authentication fails. In other words, the intruder can not pass the authentication procedure as pretending as a client or the authentication server.

For the legal client and the authentication server, with C_B, K_B^{-1} and seed, they can receive correct output data sets and so can pass the authentication successfully. If any one of these conditions is absent, then some output data set must be empty. That means the data set has been faked, so the authentication will be refused. That shows that the authentication protocol is safe.



a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	b_1	b_2	b_3	b_4	b_5	b_6	b_7
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
			↓	t_3										
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
			↓	t_5										
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
			↓	t_9										
0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
			↓	t_{10}										
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
			↓	t_{15}										
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
			↓	t_{16}										
0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
			↓	t_{17}										
0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
			↓	t_{18}										
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
			↓	t_{24}										
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
			↓	(UserID, seed , t_{25})										
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 3　specific reachable tree analysis

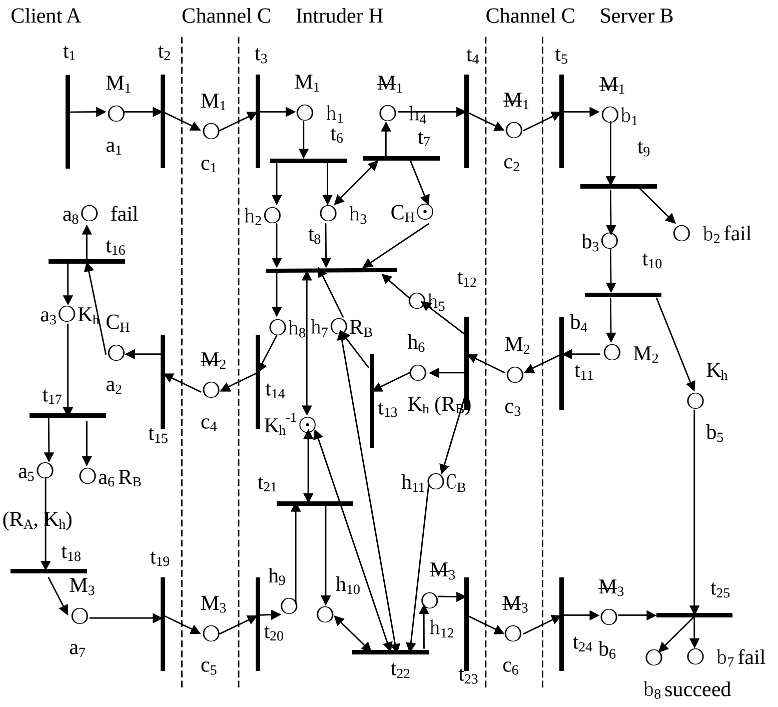


Figure 4　the Petri net description with the intruder attacks

4 Conclusion

Identity Authentication Protocol and its formal research method is an important research area of information security, it is a present well-concerned area also. In this paper, the identity authentication protocol in [1] is specified in detail which is described by Petri net. The conversing and *1*-reachable analysis proves that the identity authentication protocol is safe.

References

[1] Zhang Qiuyu, Liang Shuang. A new dynamic based on PKI authentication system design[J], Computer Application Research, 23(10), 2006, pp.116-118

[2] Chang Junjiang, Wu Zhefai; Petri net mark up the tree [J]; Journal of Software; 1993 06; 24-30

[3] Chin, Hurong. Network security system of identity authentication technology application and development of [J]. China's modern educational equipment, 2007,(01)

[4] Meadow C.Applying Formal Meathods to the Analysis of A Key Management Protocol [J]. Journal of Computer Security, 1,1992

[5] Van der Aalst W M P. The application of Petri nets to Workflow management[J] .The Journal of Circuity,Systems and Computers, l998, 8 (1) :2l-66

[6] B.B.Meh, S.E.Tavares. Modeling and analyzing cryptographic protocols using Petri nets[A]. Advances in Cryptology-AUSCRYPT' 92[C]. Vol 718 of LNCS, Springer Uerlag, 1992.pp.275 - 295

[7] Yuan Chong-Yi. Petri net theory [M]. Beijing: Electronic Industry Publishing House, 1998

[8] Gang-Soo Lee, Jin-Seok Lee. Petri net based models for specification and analysis of cryptographic protocols, Journal of Systems and Software. 37(2), 1997.5, pp.141-159

[9] Peter Stephenson. Modeling a virus or worm attack[J]. Computer Fraud and Security 2002,pp.15-19

[10] Liu Daobin, Guo Li, Bai Shuo. A new security protocol verification methods[J]. Computer Research and Development, 4(10), 2003, pp.1514-1520

Design of Network Testing System Based on Libnet

Wentao Liu

Department of Computer and Information Engineering, Wuhan Polytechnic University, Wuhan, Hubei 430023, China

Email: tcpids@gmail.com

Abstract

This paper presents a method of making network testing system which is used for network performance testing or security system testing and evaluating. The model of this system is designed and implemented and the Libnet is used to build network packet and generate the network traffic in this testing system. The principle of Libnet is discussed and the progress of making the network protocol packets by using Libnet is provided. The method of making amount of protocol packets is presented. An experiment is provided and it shows that it is more portable and flexible.

Keywords: Network Testing System; Packet Generating; Libnet; Network Traffic

1 Introduction

Network testing system can make network performance testing and network evaluation and evaluate network components such as network intrusion detection systems [1], firewalls, routers and switches. It can test the responses time of traffic on IP networks and emulates real application flows across the network to test connectivity and performance. It can tests network throughput and whether a network can support multimedia traffic and testing a network link using the application flows generated by streaming multimedia applications and determines at what rate streaming traffic is received and how much packet loss occurs. It can Tests the connectivity between local computer and another computer and supports a variety of protocols and can test network performance using TCP, UDP, IPX, and SPX networks. In fact, there are many tools which have one or more functions of network testing system such as Tcpdump [2], Ping, Traceroute, Nettimer [3] and so on. Tcpdump is a stable, mature, canonical portable packet collector and it is built by using libpcap. Network researchers frequently use tcpdump in place of bundled packet collectors and some vendors even ship it as bundled packet dumper. It requires reasonable understanding of networking to interpret collected packets. Output format can be easily and portably analyzed using awk, sed, and perl scripts. Traceroute directs a packet to each router along a path without actually knowing the path by setting the IP TTL field from 1 to n until the ultimate destination is reached. When a packet with an expired 0 TTL is received and the hop generates an ICMP with time exceeded response back to the source of sending the packet, and it identify the hop and its round trip delay. The UDP packet is sent to a probably-unused port and when the destination receives the packet it responds with ICMP with port unreachable information. Nettimer is useful for measuring end-to-end network performance and it can simulate or passively collect network traffic and it can also actively probe the network. It doesn’t need the more requirements for the information from the network and the transport protocol can be different kinds. The metrics contains bottleneck bandwidth and link bandwidth. When make more packets in the network in the network testing system, the Libnet can be used to build every kinds of packets for the different functions. Libnet [4] is an API to help with the construction and handling of network packets. It provides a portable framework for low-level network packet writing and handling. Libnet includes packet creation at the IP layer and at the link layer as well as a host of supplementary and complementary functionality. Libnet is very handy with which to write network tools and network test code. Libnet is designed and primarily maintained by Mike D. Schiffman. Many network tools are used with the Libnet, for example, Tcpreplay, Snort, Ettercap and son on.

2 Testing system

Network measurement includes active network measurements and passive network measurements. The active network measurements require sending test protocol packets into the network to determine network topology and end-to-end performance and capability of network paths. The passive network measurements do not send test protocol packets in the network but require capturing of packets in the network traffics and their corresponding timestamps transmitted by applications program running on network-attached devices over various different network links and nodes. There are many systems and tools which can be developed for network testing such as Iperf, Hping, D-ITG and so on. Iperf [5] was developed as a modern alternative for measuring TCP and UDP bandwidth performance. Iperf is a tool to measure maximum TCP bandwidth, allowing the tuning of various parameters and UDP characteristics. Iperf can be used to report bandwidth and delay jitter and datagram loss. Hping [6] is the most popular network testing tools which is designed by Salvatore Sanfilippo. It can be used for firewall testing, advanced port scanning, network testing using different protocols, manual path MTU discovery, advanced traceroute under all the supported protocols, remote OS fingerprinting, remote uptime guessing, TCP/IP stacks auditing and so on. D-ITG [7] (Distributed Internet Traffic Generator) is a platform capable to produce traffic at packet level accurately replicating appropriate random processes for both IDT (Inter Departure Time) and PS (Packet Size) random variables. D-ITG supports both IPv4 and IPv6 traffic generation and it is capable to generate traffic at network, transport, and application layer. The framework of network testing system designed based on Libnet is as fowling in Figure 1.

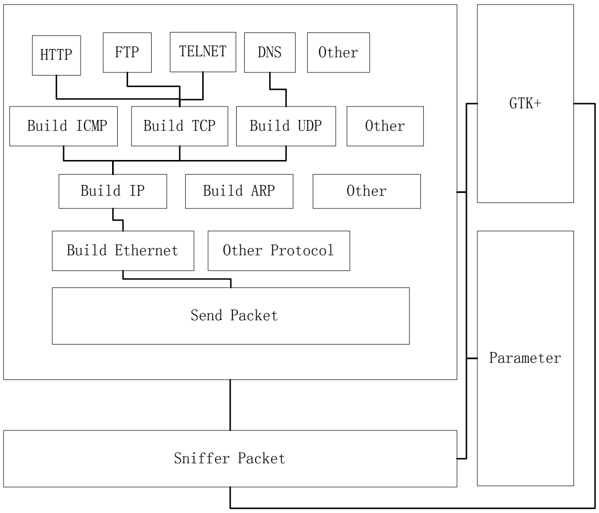


Figure 1 The framework of network testing system based on Libnet

The building packet is the basic part in generating network traffic in network testing system. A distributed multiplatform architecture for traffic generation is designed by S. Avallone, D. Emma, A. Pescapè, and G. Ventre [8] who gave a deep analysis of the generation traffic platform. The performance is important in the packet generating system [9] and in this system it can get high performance. Some commercial traffic generators include IXIA, Omnicor, Spirent and so on. A high performance internet traffic generator is presented by S. Avallone, D. Emma, A. Pescapè and G. Ventre, in 2006. TCPivo [10] is a packet replay tool that is implemented on commodity hardware using widely available open-source software and can be used as a cost-effective means for evaluating the performance of networking devices.

This system can send several kinds of network packets based on TCP/IP protocol and make more network traffic which can be used for many network performance testing and network security system effectiveness. The system can make the analysis of many QoS parameters including throughputs, inter-packet delay, packet loss rate and packet transit delay. The system can record real data carried over an IP network and replay this data using the Libnet packet generating function. For example, in testing intrusion detection system a large amount of normal traffic are built and sent by this system with a great deal number exceeding the IDS's processing capability. Because of too much traffic to process the IDS may drop packets and be unable to detect attacks. The network traffic simulation can be made to evaluate the performance of the network security systems. The system can be used for generating various HTTP workloads and for measuring server performance and other testing function also can be performed by this system.

3 Libnet

The Libnet can make many useful network tools such as ping, traceroute and son on. Libnet support many network protocols such as Ethernet, FDDI, Token Ring, ARP, RARP, MPLS, ICMP, IPv4, IPv6, IGMP, ICMP, TCP, UDP and so on. The Libnet is designed based on Libpcap [11] which is a component of capturing the network packets.

The Libnet includes two methods of building packet: raw socket method and link layer method. The raw socket need not build the link layer header and built the protocol at the IP layer and this method is very easy but it can't get more control of the building packet. The link layer method can build the packet through the link layer and the packet can be built more flexible.

The most important data structure is the Libnet context.

```
struct libnet_context
{
#if ((__WIN32__) && !(__CYGWIN__))
SOCKET fd;
LPADAPTER    lpAdapter;
#else
int fd;
#endif
int injection_type;
libnet_pblock_t *protocol_blocks;
libnet_pblock_t *pblock_end;
u_int32_t n_pblocks;
int link_type;
int link_offset;
int aligner;
char *device;
struct libnet_stats stats;
libnet_ptag_t ptag_state;
char label[LIBNET_LABEL_SIZE];
char err_buf[LIBNET_ERRBUF_SIZE];
u_int32_t total_size;
};
```

The member fd is the file descriptor of packet device. The injection_type may be the raw socket or link. If the injection_type is raw socket, it can use the raw socket to create the network packets and only crate the protocol packet from the network transform layer such as IP protocol. If the injection_type is link, it can create the packet from link layer. The protocol_blocks is protocol headers or data and it is the first protocol block. The pblock_end is last node in list. The n_pblocks is the number of pblocks. The link_type is link-layer type. The link_offset is the link-layer header size. The aligner is used to align packets. The device is the device name. The member stats is statistics and the ptag_state is the state holder for pblock tag. The label is textual label for cq interface. The err_buf is error buffer and the total_size is total size of the packet.

A packet is composed of many protocol blocks and the one block is described by the structure libnet_protocol_block which contains the all parameter of every protocol block.

```
struct libnet_protocol_block.
{
u_int8_t *buf;
u_int32_t b_len;
u_int16_t h_len;
u_int32_t ip_offset;
u_int32_t copied;
u_int8_t type;
u_int8_t flags;
libnet_ptag_t ptag;
struct libnet_protocol_block *next;
struct libnet_protocol_block *prev;
};
```

The member buf is protocol buffer and it contains the real data of protocol such as payload and all fields. The member b_len is the length of buf. The h_len is the header length which is used for checksumming. The ip_offset is offset to IP header for checksum.

The copied is the bytes copied and the type is the type of pblock such as TCP, UDP, ICMP and so on. The flags field is control flag and the ptag is the protocol blog tag. The next is the next pblcok which contains the lower layer protocol data and prev is previous bplock which contains the higher layer protocol data. A series of packets can be sent by using the Libnet context queue and the progress of making many protocol packets is as shown in Figure 2.

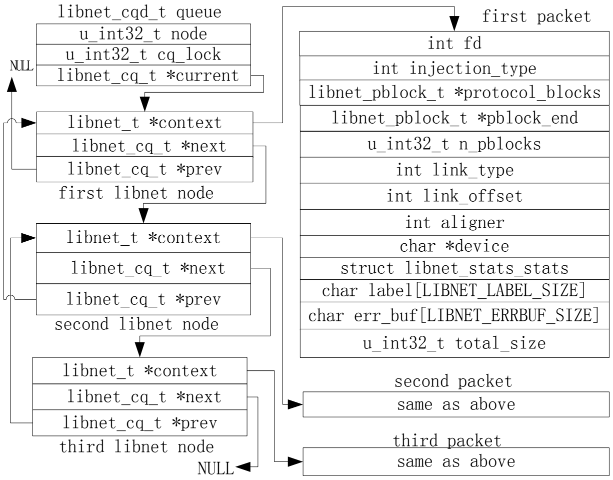


Figure 2 The process of making many packets using the Libnet context queue

4 Experiment

A TCP packet is built by the Libnet in this experiment and the core progress of the creating this packet is as follows. The type 0x20 represent tcp data and the flags 0x00 needs not checksum and flags 0x01 needs a checksum.

```
LibnetHandle = libnet_init(LIBNET_LINK,
DeviceName[DeviceIndex], LibnetError);
LibnetPtag=libnet_build_tcp(SourPort,DestPort,
12, 13, TH_SYN, 235,0,0,
LIBNET_TCP_H+strlen(TcpData),
(unsigned char *)TcpData,
TcpDataLength, LibnetHandle,0);
LibnetPtag = libnet_build_ipv4(
LIBNET_IPV4_H+
LIBNET_TCP_H+TcpDataLength,0,242,0,64,
IPPROTO_TCP,0,SourAddrNumber,
DestAddrNumber,
NULL,0, LibnetHandle,0);
LibnetPtag = libnet_build_ethernet(DestMAC,
SourMAC,ETHERTYPE_IP,NULL,0,
LibnetHandle,0 );
PacketLength = libnet_write(LibnetHandle);
```

The process of building a TCP packet is as shown in Figure 3.

The package Libpcap also provides the technology of making packets and injecting the packets to the network. In Libpcap, the function pcap_sendpacket( ) finish the method of sending the packets. But the user must create several protocol packets manually and must make all fields of protocol.

The Libnet reduces this step and gives user more special protocol functions which can set the field of protocol. In order to create a protocol packets in Libnet, user must use the function libnet_init( ) to get the injection type and network device firstly. When a new protocol packet is created, user can use libnet_build_write( ) to send the packet. The function libnet_destroy( ) is called to free resource finally.

The operating system provides a raw socket mechanism which also can create the network packet but it may be restricted in some operating system such as Windows XP SP2 which can’t send the TCP packet and UDP packet with the fake IP address. The raw socket can’t crate the link layer packets such as Ethernet packets. Libnet can resolve this problem can it can create every protocol packets form link layer.

In order to verify the correctness of the building packet, the Wireshark is used for sniffer the above packet. A comparison to both is made in the experiment.

The context of packet which is got by the Wireshark is as follows.

Ethernet II, Src: 31:52:23:64:5a:b6,
Dst: ac:bc:8f:c3:9c:5c
Destination: ac:bc:8f:c3:9c:5c
Address: ac:bc:8f:c3:9c:5c
Source: 31:52:23:64:5a:b6
Address: 31:52:23:64:5a:b6
Type: IP (0x0800)
Internet Protocol,
Src: 192.168.1.9,
Dst: 192.168.1.65
Version: 4
Header length: 20 bytes
Differentiated Services Field: 0x00
(DSCP 0x00: Default; ECN: 0x00)
Total Length: 52
Identification: 0x00f2 (242)
Flags: 0x00
Fragment offset: 0
Time to live: 64
Protocol: TCP (0x06)
Header checksum: 0xf637
Source: 192.168.1.9 (192.168.1.9)
Destination: 192.168.1.65 (192.168.1.65)
Transmission Control Protocol, Src Port: doom (666),
Dst Port: http (80), Seq: 0, Len: 12
Source port: doom (666)
Destination port: http (80)
Sequence number: 0   (relative sequence number)
Header length: 20 bytes
Flags: 0x02 (SYN)
Window size: 235
Checksum: 0xd849
Hypertext Transfer Protocol
Data (12 bytes)
acbc8fc39c5c315223645ab608004500
003400f200004006f637c0a80109c0a8
0141029a00500000000c0000000d5002
00ebd849000054637020436f6e74656e
742e

In contrast with the original packets, the result is same content as the packet.

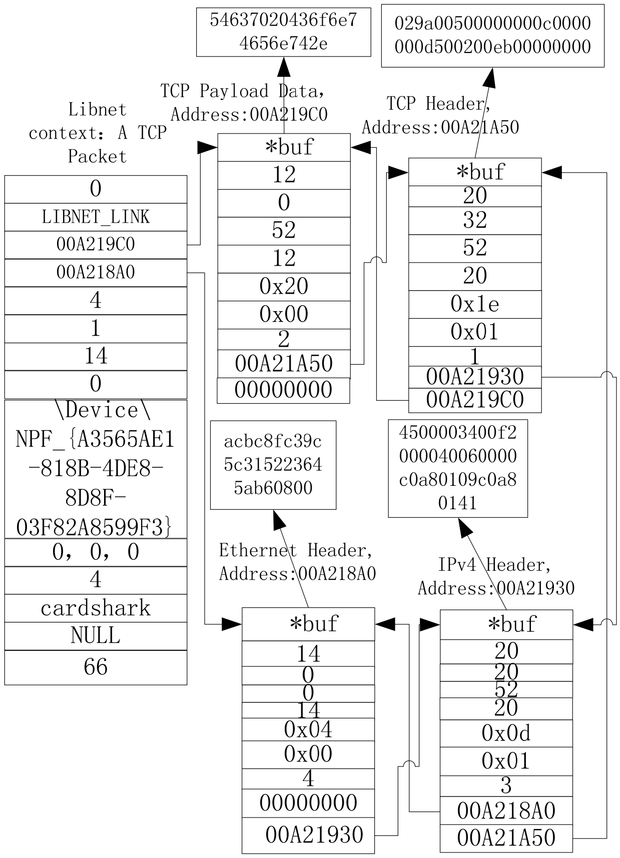


Figure 3　The process of building a TCP packet

5　Conclusion

The network measurement can be performed by the network testing system based on Libnet to get the bandwidth, delay and packet loss of networking. It also can be used for the network security system testing such as firewall, IDS, scanning system and so on. Because the Libnet owns the capability of building more kinds of protocol data packets and sending to network and it can be controlled easily and the all parameters of protocol block can be set discretionarily, this system can get high performance and extendibility.

References

[1] NJ Puketza, K Zhang, M Chung, B Mukerjee, "A Methodology for Testing Intrusion Detection Systems," in

IEEE Transaction on Software Engineering v 22 no 10, Oct 1996

[2] Tcpdump, www.tcpdump.org/tcpdump_man.html

[3] Kevin Lai and Mary Baker. Nettimer: A tool for Measuring Bottleneck Link Bandwidth. In Proceedings of USENIX Symposium on Internet Technologies and Systems, March 2001

[4] Mike D. Schiffman, http://www.packetfactory.net/libnet

[5] Iperf. http://dast.nlanr.net/Projects/Iperf/

[6] Salvatore Sanfilippo. http://www.hping.org/

[7] Alessio Botta, Alberto Dainotti, Antonio Pescapè, "Multi-protocol and multi-platform traffic generation and measurement", INFOCOM 2007 DEMO Session, May 2007

[8] S. Avallone, D. Emma, A. Pescapè, and G. Ventre, “A Distributed Multiplatform Architecture for Traffic Generation” - International Symposium on Performance Evaluation of Computer and Telecommunication Systems, (SPECTS) 2004 July 25-29, 2004

[9] S. Avallone, D. Emma, A. Pescapè, G. Ventre, “High Performance Internet Traffic Generators”, The Journal of Supercomputing, Volume 35, Issue 1, Jan 2006

[10] W. Feng, A. Goel, A. Bezzaz, W. Feng, J. Walpole, "TCPivo: A High-Performance Packet Replay Engine", ACM SIGCOMM 2003

[11] LIBPCAP, http://www.tcpdump.org/pcap3_man.html

An OPC-Friendly Detailed Routing Algorithm

Mande Xie

College of Computer & Information Engineering, Zhejiang gongshang University ,Hangzhou, Zhejiang 310035, China

Email: xiemd@mail.zjgsu.edu.cn

Abstract

The design for manufacturablity brings a huge challenge to semiconductor industry. At present, the advanced IC manufacturers modify the layout file of chip by many resolution enhancement techniques (RETs) to get better lithography images and higher yield. After thoroughly analyzing the present RETs, an improved OPC-friendly detailed routing algorithm is presented. An effective method is introduced to fast build electric amplitude of diffraction (EAD) table and the table can be rapidly refreshed during maze routing. The detailed routing algorithm is easily integrated into a routing system by simplifying a judgement of critical nets.

Keywords: optical proximity correction; maze routing; design for manufacturability

1 Introduction

As the manufacture process of IC rapidly developing, the minimum critical dimension and line space have been smaller than the lithographic wavelength. The lithography technology with shorter wavelength is still too costly and unstable. The 90nm process adopts the 193nm wavelength optical system and major IC fabs have announced that the 65nm process will adopt the 193nm wavelength optical system to leverage the mass capital investment in the 90nm node[1]. In this process, lots of factors such as a light diffraction and photo resistance and development make that mask images differ with printed silicon images This kind of distortion of printed silicon images reduces the yield rate[2]. So before the layout is printed to wafer, lots of resolution enhancement techniques (RETs) are adopted to compensate for the distortion or cancel out the interference from the neighboring light diffraction in the modern IC design flow. Optical proximity correction (OPC) is one of important RETs. However, the OPC process is time-consuming and the results are still limited by the original layout quality. An OPC-unfriendly routing algorithm has the following drawbacks: 1) Some mistakes which can not be corrected by OPC are generated during the routing; 2) A routing path which is expensive in OPC phase is chosen from the several routing paths with the same cost (such as minimum line length) for a net. Lots of lectures[3-6] have point out that physical design should have some changes to incorporate the concept of design for manufacturability. Research on routing with OPC consideration has received much attention in the literature. Huang and Wong[1] presented a mazing routing algorithm with OPC consideration, in which the routing problem of two-pin net firstly is formulated as the shortest path problem with multiple constrains and solved the problem by a Lagrangian relaxation method. When a net routed, the net length constrain and optical proximity error (OPE) constrain will be simultaneously considered. However, this algorithm has three following drawbacks: 1) for two-pin net; 2) Before a net routed, it is difficult to assign the OPE constrains for each net and the literature didn’t present an effective method to solve it. 3) An OPE constraint on a net cannot guarantee the net to meet the requirement for OPC implementation along all the routing grid cells the net passes through. Because over-strong interference situations still have chances to take place locally along a net. Some improvements have be made in the literature [7] and the shortest path problem with minimum OPE and the minimum OPE problem satisfying the net length constraint are proposed. Two kinds of improved maze routing algorithm are proposed to solve two above questions, respectively. But there are the following drawbacks in the literature [7]: 1) A judgement of critical net is itself NP-hard problem, so the method imperceptibly increases the problem complexity; 2) An effective method to fast build electronic amplitude of diffraction (EAD) table is not introduced and the literature does not point out whether EAD table needs renewal and how to carry out renewal. The literature [8] proposed after the initial routing which has gone through timing and congestion closure finished, a fast lithography simulation for a whole layout is performed to generate the edge placement error (EPE) map. Ripup and reroute is only needed to do for the lithography hotspots which some serious errors take place in. Because many times of iteration are need, the efficiency of algorithm is low. The lecture [9] roughly considers OPC in gridless routing, but OPC cost estimation is coarse in the algorithm.

Based on the lecture [7], this paper proposes an improved OPC-friendly maze routing algorithm. The main ideal is similar to that of lecture [7]. The main contribution of this paper includes: 1) The estimation method of OPC cost is improved in maze routing and an effective method is introduced to fast build EAD table and EAD table is dynamically refreshed during routing to reflect the new light intensity distribution; 2) The maze routing algorithm for critical nets and the maze routing algorithm for general nets are merged by simplifying a judgement of a critical net. As a result, the new algorithm is easily integrated into a routing system.

2 Thography system model

Modern optical system is partially coherent image system. The light intensity of aerial image can be computed by Hopking equation[10].

$$I(f,g)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}T(f^{'}+f,g^{'}+g,f^{'},g^{'})\cdot F(f^{'}+f,\ g^{'}+g)\cdot F^{*}(f^{'},g^{'})df^{'}dg^{'} \tag{1}$$

$$T(f^{'},g^{'},f^{''},g^{''})=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}J(f,g)\cdot K(f+f^{'},g+g^{'})\cdot K^{*}(f+f^{''},g+g^{''})dfdg \tag{2}$$

In Eq.（1）, $I(.,.)$ is Fourier transform of light intensity function, $T(.,.,.,.)$ is transmission cross coefficients (TCC) of optical system, and $F(.,.)$ is Fourier transform of mask function. In Eq.（2）, $J(.,.)$ is a mutual intensity function of imaging system, and $K(.,.)$ is a coherent transform function with non-aberrations. Their expressions respectively are:

$$J(f,g)=\begin{cases}\dfrac{\lambda^{2}}{\pi\cdot s^{2}\cdot NA^{2}} & f^{2}+g^{2}<(\dfrac{s\cdot NA}{\lambda})^{2}\\ 0 & f^{2}+g^{2}\geq(\dfrac{s\cdot NA}{\lambda})^{2}\end{cases} \tag{3}$$

$$K(f,g)=\begin{cases}1 & f^{2}+g^{2}<(\dfrac{NA}{\lambda\cdot M})^{2}\\ 0 & f^{2}+g^{2}<(\dfrac{NA}{\lambda\cdot M})^{2}\end{cases} \tag{4}$$

Eq.（1）indicates light intensity distribution is only connected with mask function and TCC.

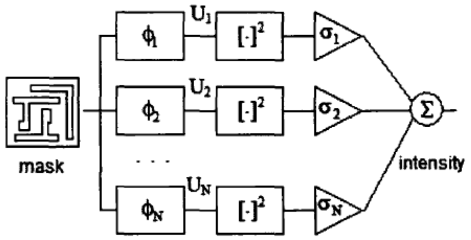


Figure 1 The partial coherent system expressed by the linear coherent system

The mask function is determined by input mask image and TCC function shows the feature of the whole optical system including illumination and imaging feature. Eq.（2）, . Eq. （3）, and . Eq.（4）indicate the TCC of optical imaging system is determined by wavelength λ of light source, numeric aperture NA, amplification coefficient M and coherent coefficient

σ of light source. However, it is too complicate to directly compute light intensity by TCC. In order to make computation of light intensity rapid and effective, the principle that partial coherent system can be approximately expressed by the sum of coherent system can be employed to compute the light intensity of partial coherent system. Figure 1 explains the principle. Obviously, the total light intensity can be expressed by:

$$U_k(x,y) = (\phi_k * g)(x,y) \qquad (5)$$

$$I(x,y) = \sum_{k=1}^{N} \sigma_k \left|U_k\right|^2 (x,y) \qquad (6)$$

In Eq（5）, U_k indicates field intensity function of coherent subsystem k， and it is the convolution of the unit pulse response function ϕ_k of subsystem and mask transform function $g(.,.)$ 。Eq.（6）indicates light intensity $I(x,y)$ is the sum of weighted square of field intensity of all subsystems. If the corresponding part of Eq.（6）is substituted by Eq.（5）, the convolution expression to calculate light intensity is:

$$I(x,y) = \sum_{k=1}^{N} \sigma_k \left|(\phi_k * g)(x,y)\right|^2 \qquad (7)$$

3　Estimation of OPC cost

Because convolution operation is complicate, the method that point light intensity is directly computed by the convolution can not be adopted by routing algorithm, which has itself high time complexity. Based on the linear shift-invariable characteristic of coherent system, the computation of point light intensity can be further simplified. In our method, the linear shift-invariable characteristic is employed for two goals: one is to make the complicated prob lems simplified; the other is to make the different problems normalized. Figure 2 shows the convolution operation of the whole mask and convolution kernel can be converted to the sum of convolution of each image in mask and convolution kernel. As a result, the complication convolution operation is converted to many simple convolution operations. Figure 3 shows any rectangle mask image R in mask can be decomposed into four rectangles. Suppose U_r is the contribution of field of rectangle r, and then:

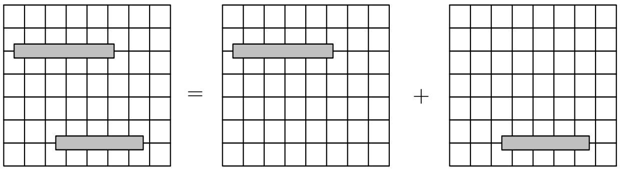


Figure 2　Conversion of convolution of complicated images

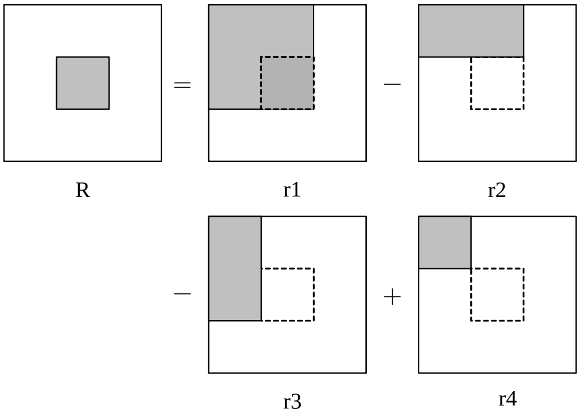


Figure 3　The field evaluation of any rectangle in mask

$$U_R = U_{r1} - U_{r2} - U_{r3} + U_{r4} \qquad (8)$$

Eq.（8）indicates the evaluation of field intensity of any rectangle in the mask can be formulated into the evaluation of the rectangle whose upper-left corner is origin. So field intensities of all rectangles whose upper-left corner is origin can be computed in advance and stored in a table. When the field intensity of any rectangle is computed, it is only needed to look up the table and do simple arithmetic operation. As a result, the complex convolution operation is avoidable. This ideal for simplifying the filed intensity evaluation is also beneficial to the algorithm realization and reuse of code.

When the routing starts, a routing region is first meshed into a routing grid. Since the amplitude of the electric field fluctuates with position obviously and the value might change from positive to negative within a routing grid cell, each routing grid cell is divided into several square sub-cells called optical grid cells. Figure 4 shows a division example. In general, there are lots of images in a mask, so the printed silicon images are influenced by both light interference and light diffraction. The light interference rapidly decays as the distance increasing and when the distance is out of the certain scope, the effort of light interference can be neglected. For this reason, an effective area is introduced and its value is generally several wave lengths. When the point light intensity is computed, an effective area whose center is the object point is defined and the images which lie in the effective area are environment variables of the object point. As the Eq.（9）indicates, the point light intensity is the sum of weighted square of field intensities which are generated by several sub-systems at this point.

$$I(x_0,y_0)=\sum_{k=1}^{N}\sigma_k\left|U_k\right|^2(x_0,y_0) \qquad （9）$$

Given the coordinate of point X is (x_0,y_0), filed intensity is $U(x_0,y_0)$, $\phi(.,.)$ is convolution kernel function and $g(.,.)$ is mask transaction function, and then

$$U(x_0,y_0)=g(x,y)*\phi(x,y)\,|_{x=x_0,y=y_0} \qquad （10）$$

$$g(x,y)=\begin{cases}1 & if\ (x,y)\,lies\ in\ the\ \ polygon\\ 0 & otherwise\end{cases} \qquad （11）$$

If $\phi(.,.)$ and $g(.,.)$ are discretized into the $M\times N$ matrix，and then：

$$U(x_0,y_0)=\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}g(x_0+M/2-i,y_0+N/2-j)\cdot\phi(i,j) \qquad （12）$$

A distributed table of field intensity is built up by the Eq. （12）. Figure 4 is an example, which shows how to compute the point light intensity by the distributed table of field intensity. The point P in this figure is the object point, and the rectangle with dashed line frame indicates its effective area. Each vertex of mask rectangles in the effective area is labeled 1、2、3、4、5、6、7、8, respectively. The step of computing the light intensity I_P is:

1)The field intensity of all vertexes are looked up from the distributed table of field intensity. Obviously, the field intensities of point 1、4、5、8 are positive and the field intensities of point 2、3、6、7 are negative.

2)Compute the arithmetic summary of field intensity of vertex array 1、2、3、4 and 5、6、7、8, and these two arithmetic summaries are labeled as U_{s1} and U_{s2} , respectively.

3)The light intensity I_P of point P is the sum of weighted square of U_{s1} and U_{s2} .

If the object point doesn't lie in the optical grid, the object point must firstly be shifted into the nearest optical grid. Repeat this process and the EAD table can be built up in advanced

Based on the light intensity of optical grid, the light intensity I_{tile} of a routing tile can be defined as the summary of light intensity of all optical grids in the tile.

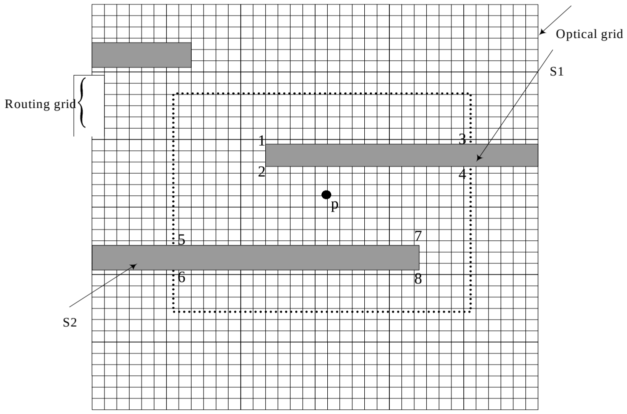


Figure 4　An example of computing point light intensity

Mask images interfere with each other in the lithography process. As a result, aerial images degenerate and distort. To estimate the degeneration degree of aerial images in a routing tile, the function OPE_{tile} is introduced and its definition is:

$$OPE_{tile}(t)=\alpha I_{tile}(t)+\beta\sum_{i\in S(t)}(I'_{tile}(i)-I_{tile}(i)) \qquad （13）$$

Where, $S(t)$ is a set of routed grids in the effective area of routing tile t. If the tile t becomes the routed tile, and then $I'_{tile}(i)$ is the light intensity of $i\in S(t)$. The first item in right of Eq.（13）indicates the light intensity of a tile; the second item indicates the increase interference when the tile t is taken up. The α and β indicate weighted factor. Because the routing of tile t is uncertain before the routing is finished, the second item in Eq.（13）can not be accurately computed. When determining the value of α and β , we must make sure that α is much greater than β to guarantee the first item is dominant and reduce the effect of the second item. In additional, to evaluate the exposure quality of a net, OPE_{net} of a net is defined as the summary of OPE_{tile} of all tiles which lie in the routing path of the routed net. If P is a routing path, then

$$OPE_{net}(P) = \sum_{t \in P}(OPE_{tile}(t)) \qquad （14）$$

When the interference is greater than a threshold, the errors can not be corrected. So the OPE_{net} of a net must be smaller than a given threshold. At the same time, the local I_{tile} must be also smaller than a given threshold, which can be expressed by the following inequality.

$$I_{tile}(t) \le local\_OPE\_bound \qquad （15）$$

Where, $local\_OPE\_bound$ is a parameter given by a user.

4　A mazing routing algorithm with OPC consideration

The literature [5] divides the routing problem into two kinds. One is for a critical net and is defined as:

Problem1: Given $k-1$ routed nets in the region and a set of convolution kernels of lithography system in spatial domain, and find the shortest routing path P with minimum $OPE_{net}(P)$ which is subjected to the following constraints:

Volume constraint;

(b) $I_{tile}(t) \le local\_OPE\_bound \quad \forall t \in P$.

The other is for a non-critical net and is defined as:

Problem2: Given $k-1$ routed nets in the region and a set of convolution kernels of lithography system in spatial domain, and find the routing path P with minimum $OPE_{net}(P)$ which is subjected to the following constraints:

(a)Volume constraint;

(b) $I_{tile}(t) \le local\_OPE\_bound \quad \forall t \in P$;

(c)The length constraints of routing path P .

Both of routing problems are solved by modified mazing routing algorithm which consists of three steps: data structure initialization step, wave propagation step and retrace step. The differences between the algorithm1 for problem 1 and the algorithms for problem2 are:

1) The prior queue employed in algorithm1 is sorted by path length and each popped element is the point with the shortest distance to source point t. However, the prior queue employed in algorithm2 is sorted by OPE_{net} and each popped element is the point with the minimum OPE_{net} to source point t.

2) When there are several routing paths with the same path length in retrace step, the path with minimum OPE_{net} is chosen in algorithm1. However, when there are several routing paths with the same OPE_{net} in retrace step, the path with minimum length is chosen.

An identification of critical net itself is a NP-hard question, so the method employed by literature [5] imperceptibly increa- ses the problem complexity.

```
RouteWithOPC(PQ, s,t, P)
    /*Step 1:propagation phase*/
1   PQ⟵(S,D,E)
2   while (PQ != empty)
3   (g,D,E)⟵ extract the partial solution with minimum E from PQ;
4    if (g=T) then go to Line 11
5    for each adjacent grid cell g'of g do
6        if (g' is not marked as an obstacle) then
7            if (g' or any grid cell g"∈U(g') violates $local\_OPE\_bound$ ) then
8                mark g as an obstacle;
9           else if (D+1 $< Long\_Net\_Bound$)
10              add (g', D+1, $E+OPE_{cell}(g^{'})$) to PQ and prune;
11          else  call algorithm1
     /*Step 2: retrace phase*/
12  use(g, D, E ) to retrace the path
13  refresh the electric amplitude of diffraction table
```

Figure 5　A mazing routing algorithm with OPC consideration

At the same time, it is difficult to integrate the algorithm into the routing system. Based on the fact that the delay of long net is big and it is difficult to route a long net, two kinds of problems are merged by a simplified strategy. Figure 5 shows our algorithm steps.

In Figure 5, D and E indicate the distance to source point and OPE, respectively. $Long\_Net\_Bound$ in line 9 indicates the threshold given by user. A net whose length is greater than this threshold is a long net and a critical net which is routed by calling the algorithm1. If the length of a net is smaller than the threshold, the net is short net and is routed for minimum OPE_{net} . To reflect the latest light intensity distribution, the EAD table employed by the algorithm is not constant input item and is dynamically updated in the routing. At the same time, after a net is routed, the OPC cost of tiles which lie in the routing path is merely updated to accelerate an algorithm. Because this routed net doesn't change the environment of the other tiles except for tiles which lie in the routing path, the strategy for updating OPC cost is valid and efficient.

5 Conclusion

Based on the lecture [5], this paper proposes an improved OPC-friendly maze routing algorithm. After how to build lithography model and EAD table is detailedly described, the method of evaluating OPC cost in mazing routing and the method of dynamically updating EAD table are introduced. The maze routing algorithm for critical nets and the maze routing algorithm for general nets are merged by simplified strategy. As a result, the new algorithm is easily integrated into routing system.

References

[1] L.D. Huang and D. F. Wong, “Optical Proximity Correction (OPC)-Friendly Maze Routing,” Proc.Design Automation Conference, 2004, pp.186-19

[2] A. B. Kahng, Y. C. Pati. “Sub-Wavelength Optical Lithography: Challenges and Impact on Physical Design”. Proceeding of the 1999 international symposium on physical design, California, 1999: 112-117

[3] F. M. Schellenberg. “Resolution Enhancement Technology: The past, the present and extension for the future”, 2004 SPIE Microlithography Symposium

[4] K. McCullen. "Phase correct routing for alternating phase shift masks," in Proc. Design Automation Conf., 2004

[5] P.Gupta and A.B.Kahng, “Manufacturing-Aware Physical Design,” *Proc. International Conference on Computer Aided Design*, 2003, pp. 681-687

[6] L.K.Scheffer, “Physical CAD Changes to Incorporate Design for Lithography and Manufacturability,” *Proc. Asia and South Pacific Design Automation Conference*, 2004, pp. 776-771

[7] Yun-Ru Wu, Ming-Chao Tsai, Ting-Chi Wang. “Maze routing with OPC consideration” Design Automation Conference, 2005. *Proceedings of the ASP-DAC 2005. Asia and South Pacific* Volume 1, 18-21 Jan. 2005 Page(s):198 - 203

[8] Joydeep Mitra, Peng Yu and David Z. Pan, "RADAR: RET Aware Detailed Routing Using Fast Lithography Simulations", *Proc. of DAC 2005*: 369372, pp:369-372

[9] Tai-Chen Chen, Yao-Wen Chang. “Multilevel full-chip gridless routing considering optical proximity correction,” *Design Automation Conference, 2005. Proceedings of the ASP-DAC 2005. Asia and South Pacific* Volume 2, 18-21 Jan. 2005 Page(s):1160 - 1163

[10] Nicolas Bailey Cobb, "Fast Optical and Process Proximity Correction Algorithms for Integrated Circuit Manufacturing", PHD. thesis, UC Berkeley, 1998