DCABES

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

14-17 October 2011
Wuxi, Jiangsu, China

Published by

CPS
Conference Publishing Services

Proceedings

2011 tenth International Symposium on
Distributed Computing and Applications to Business, Engineering & Science

DCABES 2011

Proceedings

2011 tenth International Symposium on
Distributed Computing and Applications to Business, Engineering & Science

DCABES 2011

14–17 October 2011
Wuxi, Jiangsu, China

CPS Conference Publishing Services

Los Alamitos, California
Washington • Tokyo

IEEE computer society

IEEE Computer Society Order Number E4415
ISBN-13: 978-0-7695-4415-1
BMS Part # CFP1120K-CDR

Additional copies may be ordered from:

IEEE Computer Society
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314
Tel: + 1 800 272 6657
Fax: + 1 714 821 4641
http://computer.org/cspress
csbooks@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: + 1 732 981 0060
Fax: + 1 732 981 9667
http://shop.ieee.org/store/
customer-service@ieee.org

IEEE Computer Society
Asia/Pacific Office
Watanabe Bldg., 1-4-2
Minami-Aoyama
Minato-ku, Tokyo 107-0062
JAPAN
Tel: + 81 3 3408 3118
Fax: + 81 3 3408 3553
tokyo.ofc@computer.org

Individual paper REPRINTS may be ordered at: <reprints@computer.org>

Editorial production by Juan E. Guerrero
Label art production by Mark Bartosik

IEEE Computer Society
***Conference Publishing Services* (CPS)**
http://www.computer.org/cps

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

DCABES 2011

Table of Contents

Distributed/Parallel Algorithms

Distributed/Parallel Applications

E-Business

Grid Computing and Cloud Computing

Image Processing

Internet of Things and Applications

Network Information Security

Numerical Methods for PDEs

Software Tools and Environments for Distributed or Parallel Platform

Swarm and Intelligence and Applications

Technology of Computer Application

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 2011 International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES) will be held at Jiangnan University in Wuxi, China. Besides Wuxi, the series of meeting have also been held in Hangzhou, the Three Gorges Dam Project region (Yichang), Greenwich (UK), Wuhan, and Hong Kong.

The conference topics include not only its traditional topics 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 DCABES2011 conference had received a great number of papers submitted cover a wide range of topics, such as Parallel/Distributed Computing, Image Processing, Network Technology and Information Security, E-Commerce and E-Business, Information Processing, Internet of Things, Swarm Intelligence, and so forth. Papers submitting to the conference come from over 18 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. It is indeed a pleasure to work with them and obtain their suggestions.

I am also grateful to Professor Maurício Vieira Kritz, Professor Jifeng He, Professor Zhiwei Xu, and Professor Yuhui Shi for their contributions of keynote speeches in the conference.

My sincere thanks are given to the China Ministry of Education (MOE), National Nature Science Foundation of China (NSFC), and Jiangnan University.

Finally, I would like to also thank Dr. Wei Fang, Dr. Jun Sun, Dr. Haojie Zhou, Dr. Yanrui Ding, Dr. Zhiguo Chen, Miss Litao Gong, Mr. Jianbin Yang and Mr. Bo Chen for their efforts in conference organizing activities, my postgraduate students, such as Miss Ying Li, Mr. Wenqing Guan, Miss Ji Zhao, Miss Di Zhou, for their time and help. Without their time and efforts this conference could not have been organized smoothly.

Enjoy your stay in Wuxi. Hope to meet you again at the DCABES 2011.

Wenbo Xu, *Jiangnan University, China*
DCABES 2011 Conference Chair

Choi-Hong Lai, *University of Greenwich, United Kingdom*
DCABES 2011 Program Chair

Organizing Committee

Conference Chair
Wen-Bo Xu, *Jiangnan University, China*

Program Committee Chair
Choi-Hong Lai, *School of Computing and Mathematical Sciences, University of Greenwich, United Kingdom*

Steering Committee Co-chairs
Qing-Ping Guo, *Wuhan University of Technology, Wuhan, China*
Choi-Hong LAI, *University of Greenwich, United Kingdom*

Steering Committee
Craig C. Douglas, *University of Wyoming Mathematics Department, Yale University Computer Science Department, USA*
Thomas Tsui, *Chinese University of Hong Kong, Hong Kong, China*
Wen-Bo Xu, *Jiangnan University, Wuxi, China*

Secretariat
Jian-Bin Yang, *Jiangnan University, Wuxi, China*
Li-Tao Gong, *Jiangnan University, Wuxi, China*
Bo Chen, *Jiangnan University, Wuxi, China*
Wen-Qing Guan, *Jiangnan University, Wuxi, China*
Ying Li, *Jiangnan University, Wuxi, China*

Local Organizer Committee Chair
Wen-Bo Xu, *Jiangnan University, Wuxi, China*

Scientific Committee

Xiao-Chuan Cai, *University of Colorado, Boulder, USA*
Jian-Wen Cao, *Research and Development Centre for Parallel Algorithms and Software, Beijing, China*
Jian Chen, *Jiangnan University, Wuxi, China*
Xue-Bing Chi, *Academia Sinica, Beijing, China*
Craig C. Douglas, *University of Wyoming Mathematics Department, Yale University Computer Science Department, USA*
Biao Feng, *Jiangnan University, Wuxi, China*
Wei-Dong Gao, *Jiangnan University, Wuxi, China*
Qing-Ping Guo, *Wuhan University of Technology, Wuhan, China*
Hai-Wu He, *Hohai University, Nanjing, China*
Pui-Tak Ho, *University of Hong Kong, Hong Kong, China*
Zhi-Cheng Ji, *Jiangnan University, Wuxi, China*
Zheng-Yu Jin, *Jiangnan University, Wuxi, China*
Li-Shan Kang,*Wuhan University, China*
David Keyes, *Columbia University, USA*
Souheil Khaddaj, *Kingston University, United Kingdom*
Choi-Hong Lai, *University of Greenwich, United Kingdom*
John Lee, *Hong Kong Polytechnic, Hong Kong, China*
Heather M. Liddell, *Queen Mary, University of London, United Kingdom*
Hai-Xin Lin, *Delft University of Technology, Delft, Netherlands*
Ping Lin, *National University of Singapore, Singapore*
Alfred Loo, *Hong Kong Lingnan University, China*
Frank C.K. Ng, *Chinese University of Hong Kong, China*
Michael Ng, *Baptist University of Hong Kong, China*
Zhen-Qiu Ning, *Jiangnan University, Wuxi, China*
Guo-Dong Shi, *Changzhou University, China*
P.M.A. Sloot, *University of Amsterdam, Netherlands*
Jun Sun, *Academia Sinica, Beijing, China*
Thomas Tsui, *Chinese University of Hong Kong, China*
Hong-Yuan Wang, *Changzhou University, China*
Mei-Qing Wang, *Fuzhou University, Fuzhou, China*
Shi-Tong Wang, *Jiangnan University, Wuxi, China*
Xiao-Jun Wu, *Jiangnan University, China*
Wen-Bo Xu, *Jiangnan University, Wuxi, China*
Yi-Hua Zhu, *Zhejiang University of Technology, Hangzhou, China*
J. Zou, *Chinese University of Hong Kong, China*

Local Organizer Committee

Zhi-Cheng Ji, *Jiangnan University, Wuxi, China*
Guo-Dong Shi, *Changzhou University, China*
Fei Liu, *Jiangnan University, Wuxi, China*
Xi-Sheng Wu, *Jiangnan University, Wuxi, China*
Yuan Liu, *Jiangnan University, Wuxi, China*
Li-Xia Liu, *Jiangnan University, Wuxi, China*
Shi-Tong Wang, *Jiangnan University, Wuxi, China*
Xiao-Jun Wu, *Jiangnan University, Wuxi, China*
Hong-Yuan Wang, *Changzhou University, China*
Xiao-Feng Gu, *Jiangnan University, Wuxi, China*
Feng Pan, *Jiangnan University, Wuxi, China*
Jiu-Zhen Liang, *Jiangnan University, Wuxi, China*
Xiu-Hong Chen, *Jiangnan University, Wuxi, China*
Xi-Huang Zhang, *Jiangnan University, Wuxi, China*
Hong-Wei Ge, *Jiangnan University, Wuxi, China*
Feng Ding, *Jiangnan University, Wuxi, China*
Zhen-Qiu Ning, *Jiangnan University, Wuxi, China*
Yi-Hua Zhu, *Zhejiang University of Technology, Hangzhou, China*
Jun Sun, *Jiangnan University, Wuxi, China*
Wei Fang, *Jiangnan University, Wuxi, China*
Hao-Jie Zhou, *Jiangnan University, Wuxi, China*
Yan-Rui Ding, *Jiangnan University, Wuxi, China*
Zhi-Guo Chen, *Jiangnan University, Wuxi, China*
Yue-Yang Li, *Jiangnan University, Wuxi, China*

Program Committee

Craig C. Douglas, *University of Wyoming Mathematics Department, Yale University Computer Science Department, USA*
Peter Jimack, *Faculty of Engineering, University of Leeds, United Kingdom*
Chris Jesshope, *Hull University, United Kingdom*
Jian-Wen Cao, *Laboratory of Parallel Computing ,Institute of Software Chinese Academy of Sciences, China*
Xue-Bing Chi, *Suppercomputing Center, Computer Network Information Center, Chinese Academy of Sciences, China*
Wen-Bo Xu, *School of Internet of Things Engineering, Jiangnan University, China*
Alfred Loo, *Department of Computing and Decision Science, Lingnan University, Hong Kong*
Man Leung Wong, *Department of Computing and Decision Sciences, Lingnan University, Hong Kong*
Faouzi Alaya Cheikh, *Department of Computer Science and Media Technology, Gjovik University College, Norway*
Nikos Christakis, *Department of Applied Mathematics, University of Crete, Heraklion, Greece*
Hai-Xin Lin, *Faculty of Information Technology and Systems, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Netherlands*
Souheil Khaddaj, *Faculty of Computing, Information Systems and Mathematics, Kingston University, United Kingdom*
Mei-Qing Wang, *College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China*
Albert Zomaya, *School of Information Technologies University of Sydney, Australia*
Dan Liu, *China Criminal Police University, China*
Fei Liu, *School of Internet of Things Engineering, Jiangnan University, China*
Shi-Tong Wang, *School of Digital Media, Jiangnan University, China*
Xiao-Jun Wu, *School of Internet of Things Engineering, Jiangnan University, China*
Feng Ding, *School of Internet of Things Engineering, Jiangnan University, China*
Qing-Ping Guo, *School of Computer Science & Technology, Wuhan University of Technology, China*
Zhen-Qiu Ning, *Jiangnan University, Wuxi, China*
Yi-Hua Zhu, *Zhejiang Universitry of Technology, Hangzhou, China*

Reviewers

Albert Zomaya, *School of Information Technologies University of Sydney*
Alfred Loo, *Hong Kong Lingnan University*
Chaofeng Li, *Jiangnan University, Wuxi, China*
Cheng-Yuan Li, *Jiangnan University, Wuxi, China*
Chi Xuebing, *Academia Sinica*
Choi-Hong Lai, *University of Greenwich*
Chris Jesshope, *Hull University*
Craig C. Douglas, *University of Wyoming Mathematics Department, Yale University Computer Science Department, USA*
Dan Liu, *Criminal Police University*
Di Zhou, *Jiangnan University, Wuxi, China*
Ding Feng, *Jiangnan University, Wuxi, China*
Fang Wei, *Jiangnan University, Wuxi, China*
Faouzi Alaya Cheikh, *Gjovik University College*
Fei Liu, *Jiangnan University, Wuxi, China*
Guo Qingping, *Wuhan University of Technology*
Guodong Shi, *Changzhou University*
Hao Gao, *Jiangnan University, Wuxi, China*
Hong-Wei Ge, *Jiangnan University, Wuxi, China*
Ji Zhao, *Jiangnan University, Wuxi, China*
Juan Mei, *Jiangnan University, Wuxi, China*
Jun Sun, *Jiangnan University, Wuxi, China*
Lin Haixin, *Delft University of Technology*
Maolong Xi, *Jiangnan University, Wuxi, China*
Man Leung Wong, *Lingnan University, Hong Kong*
Meiqing Wang, *Fuzhou University*
Mingliang Li, *Jiangnan University, Wuxi, China*
Nicholas Christakis, *University of Crete*
Ning Zhenqiu, *Jiangnan University, Wuxi, China*
Peter Jimack, *University of Leeds, United Kingdom*
Shitong Wang, *Jiangnan University, Wuxi, China*
Souheil Khaddaj, *Kingston University*
Wang Hongyuan, *Changzhou University*
Wei Chen, *Jiangnan University, Wuxi, China*
Jianwen Cao, *Research and Development Centre for Parallel Algorithms and Software*
Xiaojun Wu, *Jiangnan University, Wuxi, China*
Yi Fu, *Jiangnan University, Wuxi, China*
Yi-Hua Zhu, *Zhejiang University of Technology, Hangzhou, China*
Yueyang Li, *Jiangnan University, Wuxi, China*
Zhiguo Chen, *Jiangnan University, Wuxi, China*
Wen-Bo Xu, *Jiangnan University, Wuxi, China*

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Parallel Domain Decomposition Methods for Ray-Tracing on multi-cores and multi-processors

Cédric Venet and Frédéric Magoulès
Applied Mathematics and Systems Laboratory
Ecole Centrale Paris
Châtenay-Malabry, France
Email: frederic.magoules@hotmail.com

***Abstract*—In this paper, an original parallel domain decomposition method for ray-tracing is proposed to solve numerical acoustic problems on multi-cores and multi-processors computers. A hybrid method between the ray-tracing and the beam-tracing method is first introduced. Then, a new parallel method based on domain decomposition principles is proposed. This method allows to handle large scale open domains for parallel computing purpose, better than other existing methods. Parallel numerical experiments, carried out on a real world problem—namely the acoustic pollution analysis within a large city—illustrate the performance of this new domain decomposition method.**

***Keywords*-Domain decomposition methods; Parallel and distributed computing; Acoustics; Ray-tracing methods; Beam-tracing methods;**

I. Introduction

Environmental noise has become a major problem in lot of countries, since it causes not only stress but also various diseases like hypertension or heart ischaemia. Laws at European levels are promulgated to reduce or restrict such environmental noise; for example by fixing a maximum noise level around densely populated areas. In order to come up with effective solutions to reduce noise, one must be able to assess the effect of various architectural configurations. To do this, fast noise simulation is an invaluable tool. However sound propagates far away in open domains so simulations must be done in a large scale model, and well designed methods are a key point to obtain fast simulations. In order to study and to reduce this acoustic pollution, one has to use either numerical or geometrical methods.

Numerical methods (also called Wave-based methods), like the Boundary Element Method [1], the Infinite Element Method [2], [3], the Finite Element Method [4], [5], and the coupling methods [6] approximate the mathematical equations of the acoustic problems. These methods are very accurate and physically correct, but they need a lot of computational power and memory. Moreover these methods have some difficulties to handle extremely large domains (relative to the wavelength) and open domains, such as exterior acoustic problems.

Geometrical methods are well suited for large domains and open domains. These methods assume that the sound propagates in straight lines, which means that such methods provide accurate results only for high frequencies (relative to the size of the problem). These methods do not usually take into consideration the diffraction and diffusion phenomena. Geometrical methods deal with large domains without any difficulty and are quite fast compared to the numerical methods. Well known geometrical methods are the image-source method [7], [8], the ray-tracing method [9] and the beam-tracing method [10], [11].

In this paper, geometrical methods are considered to solve acoustic pollution problems within a large city. A modified beam-tracing method is first introduced, more precisely a hybrid beam-tracing and ray-tracing version with some similarities to frustum-tracing [12]. Domain decomposition methods [13], widely used for numerical methods in acoustics [14], are here extended to the beam-tracing method and allows us to solve large scale acoustic problem. Parallel numerical experiments performed in the shinjuku area (district of Tokyo, Japan) illustrate the efficiency, performance and robustness of this new parallel domain decomposition method.

II. Overview of Geometrical Acoustics Methods

Geometric acoustic modeling software are commonly used to design three-dimensional architectural environments. Sound travels from source to receiver via a multitude of paths containing reflections, transmissions, and diffractions. Geometrical acoustic methods consist to compute these multiple paths of the sound, and then to collect the acoutic pressure at the microphones. In the following paragraphs some of these methods are shortly reviewed.

The image-source method [7], [8] creates virtual sources for each reflection of a source (and of other virtual sources) by a surface of the model. Virtual microphones are placed in the model at the points where the noise level must be evaluated. Then, for each microphone, the image-source method computes the contribution of the sources by checking that the path between the source and the microphone is not blocked by an obstacle.

The ray-tracing (or ray-shooting) method [9] divides the energy of each source between a huge number of elementary

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.65

particles (similar to their physical counterpart the photon). Those particles propagate in straight lines and gradually loose energy due to the damping of the air. When they encounter a surface, the particles are reflected and loose some energy too. When a particle energy goes below a given threshold, it is deleted. Each time a particle crosses a microphone volume—unlike other geometrical methods, these microphones are spheres and not simple points—its energy contribution is added to the microphone.

The beam-tracing method, instead of splitting the energy between particles, split the energy between beams. The energy is distributed to a beam according to the power of the source in the directions of the beam. Starting from its source, each beam propagates in the air, loosing some energy due to damping. When a beam intersects a surface, it is reflected and looses some energy too. Once the energy of a beam reaches a given threshold, it is deleted. Each time a microphone is inside a beam, the contribution of the beam energy is added to its noise level. This method is harder to implement than the ray-tracing method, mainly because intersections between beams and surfaces are more complex to detect and to handle than between rays and surfaces[1].

On one hand, Ray-tracing methods have been extensively studied in the context of computer generated imagery. A lot of work was done on acceleration structures to speedup the ray/model nearest intersection detection and in particular on the efficient generation of hierarchical partitioning of good quality [15], [16]. The use of vectorial instructions (like streaming SIMD extensions (SSE)) [16], [17] and of graphics processing unit (GPU) [18], [19] to accelerate ray-tracing has also been widely investigated. These optimized methods are very efficient when the rays are heavily correlated, i.e. when they follow similar trajectories, which is usually the case in computer generated imagery. However, this is less relevant in acoustic analysis. On the other hand, beam-tracing is rarely used for computer graphic [20], allowing great quality soft shadows and anti-aliasing. However the additional complexity and slowdown are often not worth it, so ray-tracing is largely preferred. Ray-tracing of large scale models got also some attention [17], [21], [22]. However these results are not directly applicable to acoustics due to the assumptions made.

III. Hybrid ray-beam tracing method

In this paper our approach is based on the geometrical method named in the follwoing the Hybrid ray/beam-tracing method. It is an approximate beam-tracing method which is similar to a ray-tracing method. The detection of intersections between the beams in the model reduces to the detection of a ray/model intersection test. In other words this means that instead of testing the whole volume of the beam against the model, only the guiding ray of the beam is tested. This lead to the second approximation: in this method, the beams are never splits on reflection. This introduces some errors, but this is a trade-off between precision and speed. Using such approximation allows to use very efficient ray shooting methods. Since a beam can be treated very fast, small beams can be used to reduce the approximation error. So in practice, each beam is very small and the errors are not that large. Moreover, a beam is very small near the source, where it has the most energy, and becomes larger far from it, where its energy is mostly depleted. For these reasons, the committed errors are less problematic than it appears. Additionally, it would be easy to split a beam into sub-beams once it reaches a given size in order to limit the potential errors.

[1]If the whole beam does not hit one single surface, it needs to be split into sub-beams, one for each surface that the parent beam intersects. So each beam requires more computation to handle than a simple ray. However, less beams than rays (particles) are needed to obtain the same quality of the results.

IV. Domain decomposition method for beam-tracing

A. Amdahl's law for beam-tracing

It is well known that one of the limitation of parallel computing consists in the input and output of the code, which are in a standard implementation, sequential. These parts of any code are well understood with the Amdahl's law which enunciates that the global speedup of a program with a local speedup of S on a percentage f of the sequential execution time is equal to $S_{total} = ((1-f) + (f/S))^{-1}$ which gives, for a partially parallel program with an efficiency η, $S_{total}(n) = ((1-f) + (f/(\eta n)))^{-1}$, where n denotes the number of threads. It is clear from this formula that the maximal speedup is bounded by $1/(1-f)$.

While the beam-tracing itself can be parallelized, the input, output and result gathering cannot. To be able to use a lot of processes efficiently, the input, output and result gathering need to be made parallel. One of the main goals of the Domain Decomposition Methods (DDM) presented here is to allow the whole program to be parallel. The idea is to split the model into sub-domains. Then the loading, output and result gathering can be done in parallel for each sub-domain. The splitting of the model is fast and can be done only once per model and reused when needed. Another important advantage of the DDM is to allow to consider very large models and a huge number of microphones. This is particularly important when generating volumetric noise map, since the number of microphones quickly becomes very large. Each process only needs to allocate memory for the sub-domains it is currently working on instead of allocating memory for the whole domain.

B. Microphones partitioning

A basic idea consists of simply splitting the set of microphones into multiple sets, one per sub-domain, each sub-domain containing the complete geometry, as illustrated

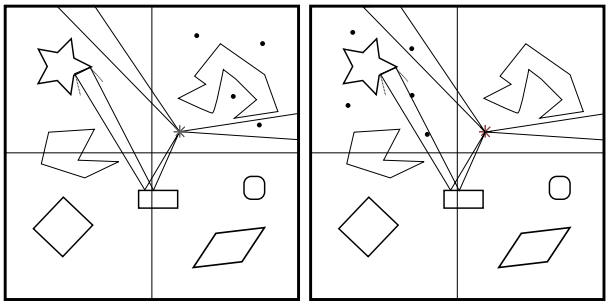


Figure 1. Microphones partitioning method.

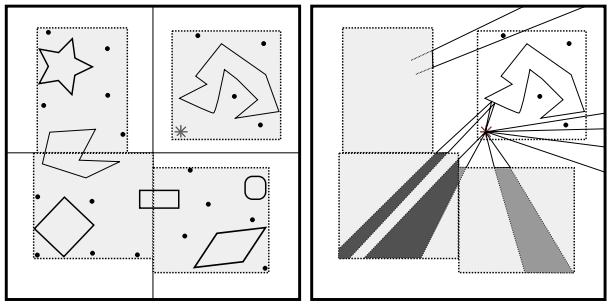


Figure 2. Splitting of the model and microphones (left) and shooting beams (right).

in Figure 1. Each sub-domain is totally independent so the program can be run in parallel. A whole part of the work is duplicated, especially the intersections detection and the loading/preprocessing of the model. If the model is relatively simple and as such most of the computation and memory are used by the microphones, this can be a good alternative. There is usually nothing to implement, only the configuration files need to be changed. Since there is absolutely no communication between the processes, this is a good candidate for distributed systems even those which are loosely connected. In particular, this could be run on a grid and even on a system like SETI@home or Folding@home. If it is used with traditional parallelism, there are also some loadbalancing issues. Indeed the processing time of each sub-domain can vary a lot and there is no easy way to reduce this problem.

C. On demand geometry and microphones loading

Another idea with some similarity to domain decomposition methods is to load sub-domains on demand, i.e. that some hierarchical acceleration structure is precomputed and only the first levels are loaded at the start. If, during the traversal of the structure, a node not yet loaded is reached, the corresponding data are loaded. If it is an interior node, these data are the child nodes informations, if it is a leaf, the data are the corresponding mesh and microphone data. At the end, only the data used have been loaded. This system is pretty complicated to implement and needs a specific precomputed data structure. Also the node loading latency needs to be hidden (for example by changing the current beam to one whose data are ready). The gathering of the final results is also complicated since the data are spread over all processes with potential duplicates needing merging. Moreover, this method is only useful if the beams launched by a process stay in a specific zone of the model. If the beams spread everywhere, the process needs to load the complete model. Since beams in exterior acoustics often go far and are widely dispersed in the model, this method is probably not a good match for exterior problems.

D. Geometry and microphones partitioning

The original method proposed here matches more closely a domain decomposition approach [13]. Here the geometry of the model and the microphones are split into multiple sub-domains, as illustrated in Figure 2. The beam-tracing is done in each sub-domain. When a beam goes out of a sub-domain, it is sent to the next sub-domain it intersects. In practice, the beams are shot in a modified model where all the sub-domains except the current one are replaced by a simplified version of themselves, for example a bounding box. The main difference with a classical domain decomposition method is that there is not a one to one correspondence between the processes and the sub-domains. Indeed, if each process were associated to one and only one sub-domain, the load balancing would be dramatic. For example, consider the case where there is only one source. The sub-domain containing the source would have the most work and the others a lot less. This would reduce the efficiency. This is why a more complex load balancing scheme has to be used. The idea is that each process starts with one sub-domain, but when it has few remaining beams to handle, it start to load one or more sub-domains having a lot of beams remaining. The number of sub-domains which can be loaded simultaneously is limited to contain memory usage. Sub-domains can also be unloaded when few beams remain to be processed. This allows to do most of the results gathering during the processing of other beams, which reduces the final gathering time a lot. Overlapping gathering and processing is efficient since gathering mainly uses the communication system.

The load balancing of the computations is the main problem. The goal is that every process always has something to do. The loading and unloading of sub-domains take a long time and produces no added value hence it must be avoided when possible. These goals are not easy to fulfill since the number of beams to process in a sub-domain is unknown and the time needed to process one varies a lot. The solution used is to process first the exchanged beams and then to equalize the number of remaining source beams in each sub-domain

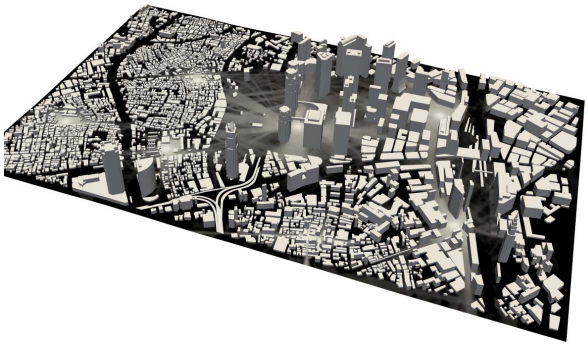


Figure 3. Noise level distribution on the ground in Shinjuku area (district of Tokyo, Japan).

comparatively to the processing power affected to this sub-domain by (un)loading sub-domains when this ratio deviate too much from the average.

V. Applications and numerical experiments

The numerical experiments are conducted for the simulation of urban acoustic pollution. The model is the Shinjuku area (district of Tokyo, Japan). This area consists of an area of 2.5 km by 1.5 km around the point of coordinates $35^o41^{'}23^{''}N$, $139^o1^{'}25^{''}E$. Inside these bounds, there is the busiest train station of the world and a number of skyscrapers (at least ten with a height of 200 meters or more) including the administration center of the government of Tokyo and other commercial and administrative buildings. The rest of the area is composed of a park, smaller buildings and housing. The test consists in the simulation of the propagation of several noise sources with the proposed hybrid ray-tracing method with the DDM-based parallelization. A total number of 10 millions beams for each source is used for the simulation. Figure 3 illustrates the noise level distribution obtained from the simulation. The numerical experiments were run on the NOVA supercomputer (Bull S.A.) composed of 32 nodes; each node is composed of four quadricore processors "Xeon E7310 Tigerton". A total number of 13.7 millions microphones are placed in a regular grid in all the volume of the district, which corresponds to a microphone every 4 meters in each of the three spatial directions. The partitioning used is trivial. Since the domain is the box of size (w, h, d) with the origin $(0, 0, 0)$, it is split into regular boxes along its longest axe, i.e. for n sub-domains, the sub-domain i is the box of size $(\frac{w}{n}, h, d)$ with the origin $(\frac{iw}{n}, 0, 0)$. On each process, a very efficient multithreaded beam-tracer with dynamic worksharing is used. The results are stored in a shared array. Each microphone data are protected by its own spin-lock allowing for safe and fast simultaneous data update.

In the tests, the speedups are computed using the total execution time, including the loading of the model, the acceleration structure generation and the saving of the results. They were realized with two different MPI backends. The first one uses Ethernet (on top of Infiniband) and the second one uses Infiniband natively. Despite the use of the same hardware, the second one is faster. Figures 4 and 5 show the global speedup of the DDM program for these two means of communication. The efficiency is computed with respect to the sequential version: $\eta = \frac{t_{ref}}{numb\_threads \cdot t}$ where t_{ref} is the execution time of the sequential version and t the time for the parallel version. The speedup is evaluated as: $S = \frac{t_{ref}}{t} = numb\_threads \cdot \eta$. The speedup of the program without domain decomposition show diminishing returns quickly. In fact it is counterproductive to use more than 64 threads, i.e. when using 128 threads the program takes longer to complete than with 64 threads. The use of multiple sub-domains allows a bigger part of the application to be parallelized, and gives better results. However, to have more sub-domains than nodes is not advantageous. The four sub-domains case is a little strange. However, the 8 sub-domains case has the best speedups with 64 and 128 threads (4 and 8 nodes). The speedups may not seem exceptional, but taking into account that it is the whole execution time, they are very good compared to other codes. The differences between the results of the two MPI implementations show that the communication speed influences mostly the one and two sub-domains cases with lots of threads. This is logical

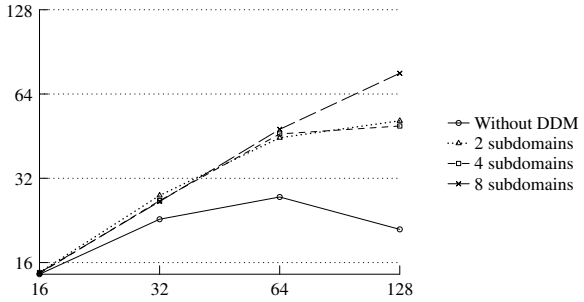


Figure 4. Speedup of the whole DDM program (using Ethernet) with respect to the number of threads and sub-domains.

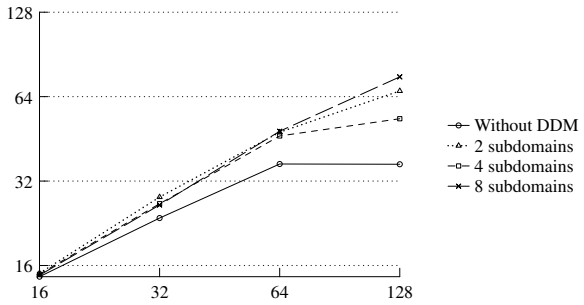


Figure 5. Speedup of the whole DDM program (using Infiniband) with respect to the number of threads and sub-domains.

since the one and two sub-domains cases need the most communication bandwidth to gather the results. Anyway, due to the high performance of the new domain decomposition method proposed, the difference remains small.

VI. Conclusion

In this paper a new domain decomposition method for ray-tracing is proposed. Numerical simulations have been carried out to evaluate the performance and efficiency of this new method for analyzing large scale urban acoustic pollution. This new method allows to use more processors than other methods and scales to extremely large models by parallelizing the whole algorithm, including the gathering process. Multi-cores and multi-processors implementation of this method were run on a NOVA supercomputer (Bull S.A.). Since the quantity of memory needed and the interconnection bandwidth requirement are reduced by this method, it also could be used on inexpensive systems, such as cluster of workstations with an Ethernet switch.

References

[1] J. Zhang, W. Zhao, and W. Zhang, "Research on acoustic-structure sensitivity using FEM and BEM," *Journal of Vibration Engineering*, vol. 18, no. 3, pp. 366–370, 2005.

[2] R. Astley, G. Macaulay, and J.-P. Coyette, "Mapped wave envelope elements for acoustical radiation and scattering," *Journal of Sound and Vibration*, vol. 170, no. 1, pp. 97–118, 1994.

[3] J.-C. Autrique and F. Magoulès, "Studies of an infinite element method for acoustical radiation," *Applied Mathematical Modelling*, vol. 30, no. 7, pp. 641–655, 2006.

[4] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*. Springer, 1998.

[5] L. L. Thompson, "A review of finite-element methods for time-harmonic acoustics," *The Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1315–1330, 2006.

[6] J.-C. Autrique and F. Magoulès, "Numerical analysis of a coupled finite-infinite element method for exterior Helmholtz problems," *Journal of Computational Acoustics*, vol. 14, no. 1, pp. 21–43, 2006.

[7] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[8] V. Pulkki, T. Lokki, and L. Savioja, "Implementation and visualization of edge diffraction with image-source method," in *Proceedings of the 112th Audio Engineering Society Convention*, 2002.

[9] D. O. Elorza, "Room acoustics modeling using the raytracing method: implementation and evaluation," 2005.

[10] T. Funkhouser, N. Tsingos, I. Carlbom, G. Elko, M. Sondhi, J. E. West, G. Pingali, P. Min, and A. Ngan, "A beam tracing method for interactive architectural acoustics," *Journal of the Acoustical Society of America*, vol. 115, no. 2, pp. 739–756, 2004.

[11] S. Laine, S. Siltanen, T. Lokki, and L. Savioja, "Accelerated beam tracing algorithm," *Applied Acoustics*, vol. 70, no. 1, pp. 172 – 181, 2009.

[12] A. Chandak, C. Lauterbach, M. Taylor, Z. Ren, and D. Manocha, "AD-Frustum: Adaptive Frustum Tracing for Interactive Sound Propagation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1707–1722, 2008.

[13] F. Magoulès and F.-X. Roux, "Lagrangian formulation of domain decomposition methods: a unified theory," *Applied Mathematical Modelling*, vol. 30, no. 7, pp. 593–615, 2006.

[14] F. Magoulès, K. Meerbergen, and J.-P. Coyette, "Application of a domain decomposition method with Lagrange multipliers to acoustic problems arising from the automotive industry," *Journal of Computational Acoustics*, vol. 8, no. 3, pp. 503–521, 2000.

[15] K. Subramanian and D. Fussel, "A search structure based on k-d trees for efficient ray tracing," The University of Texas at Austin, Tech. Rep. Tx 78712-1188, 1992.

[16] I. Wald, "Realtime Ray Tracing and Interactive Global Illumination," Ph.D. dissertation, Computer Graphics Group, Saarland University, 2004.

[17] I. Wald and P. Slusallek, "State of the art in interactive ray tracing," in *State of the Art Reports, EUROGRAPHICS 2001*. Manchester, United Kingdom: EUROGRAPHICS, 2001, pp. 21–42.

[18] J. Günther, S. Popov, H.-P. Seidel, and P. Slusallek, "Realtime ray tracing on GPU with BVH-based packet traversal," in *Proceedings of the IEEE Eurographics Symposium on Interactive Ray Tracing*, Sep. 2007, pp. 113–118.

[19] D. R. Horn, J. Sugerman, M. Houston, and P. Hanrahan, "Interactive k-d tree GPU raytracing," in *Proceedings of the 2007 Symposium on Interactive 3D Graphics and Games*. New York, NY, USA: ACM, 2007, pp. 167–174.

[20] R. Overbeck, R. Ramamoorthi, and W. R. Mark, "A Real-time Beam Tracer with Application to Exact Soft Shadows," in *Eurographics Symposium on Rendering*, Jun 2007.

[21] G. Cadet, S. Zambal, and B. Lcussan, "Rendering complex scenes on clusters with limited precomputation," in *Proceedings of the International Symposium on High Performance Computational Science and Engineering*, vol. 172. Springer Boston, 2005, pp. 79–96.

[22] C. Lauterbach, S.-E. Yoon, M. Tang, and D. Manocha, "Reducem: Interactive and memory efficient ray tracing of large models," *Computer Graphics Forum*, vol. 27, no. 4, pp. 1313–1321, Jun. 2008.

The Research of Q Learning-Based Estimation of Distribution Algorithm

Hu yugang
Information Department
Changzhou Textile Garment Institute
Changzhou china
Huyugang80@163.com

Abstract: **This paper focuses on the theory of estimation of distribution algorithms. First, elaborated the idea of estimation of distribution algorithms, And then for the limitations of solving complex optimization problems,proposed Q Learning-Based Estimation of Distribution Algorithm. The Q learning algorithm is introduced into evolutionary computation, through the Agent and group interaction, to achieve a probability model of adaptive updates. Test functions using six classical comparative experiment, the results show that the algorithm performance is stable, running time is short, with a strong global search ability, is an efficient solving algorithm for function optimization problems.**

Keywords- Estimation of Distribution Algorithm, Q Learning, Evolutionary Search

I THE PROPOSED ESTIMATION OF DISTRIBUTION ALGORITHM

In order to overcome the genetic algorithm because of chromosomal rearrangements lead to the chain problem, people ask if you can not use the crossover and mutation operations, but by the optimal solution set from the extracted information, and then use this information to generate new solutions of the probability distribution, which Chain Linkage Leaning. The probability model using constructive thinking into the evolution of computing the theoretical basis for estimation of distribution algorithms.

The concept of estimation of distribution algorithm first proposed in 1996, and after the year 2000 has been developing rapidly. It will build into the probability model and the sampling process of evolution to replace the traditional crossover and mutation. As is the use of probabilistic models to guide the search process, to avoid the blindness caused by chromosomal recombination and random, Thus effectively improving the efficiency of search, Fast, reliable solution to many of the traditional genetic algorithm optimization problem difficult to solve.

II Q LEARNING-BASED ESTIMATION OF DISTRIBUTION ALGORITHM

A Questions

Not difficult to find by analyzing the existing distribution of variables unrelated to the reason why estimation algorithm will often display a poor performance is due to update its probability vector normal to a single fixed strategy, not only can not guarantee that the whole evolution process of the strategy is always effective, Does not take into account the evolution of the gene locus that appears when the difference. If the probability of each gene locus corresponding values can be adaptive in the evolutionary process of updating, it will help to improve the performance of evolutionary search

In order to achieve the adaptive update the probability vector can be associated with a different gene locus Agent, and the selection probability update rules as its action. Thus, the probability value of each update to convert into Agent performs an action. If the group as a further evolution of the environment, each Agent can use reinforcement learning method and the environment interact to find the optimal movement strategy.

Q learning as a typical reinforcement learning algorithm does not need to estimate the environmental model, but an iterative calculation by optimizing the Q function to obtain the optimal movement strategy, which can be selected as the Agent of learning. It is based on this idea, the following proposed Q Learning-Based Estimation of Distribution Algorithm(QEDA).

B Algorithm design

Q Learning-Based Estimation of Distribution Algorithm with the binary coding, population size is N, code length m, the first generation of the probability vector t is denoted by p (t) = (p1 (t), p2 (t), ... pm (t)), Where pi (t) for the first i loci the probability of taking one. The basic flow algorithm consistent with Figure 1, each iteration including selection, construction (updated) probability models and sampling and other operations.

Sampling operation and PBIL, UMDA, etc. the same algorithm, using Monte Carlo method, in accordance with the probability vector N individuals randomly generated. Options selected in addition to the fitness of each generation according to the optimal sub-groups, but also select the worst sub-groups, based selectivity are r ($0 < r < 1$), the size of the sub-groups are $M=\lfloor rN \rfloor$. Updated statistical probability model, respectively, the optimal operation of first and worst in you to take a sub-group the frequency of, denoted by gi (t) and bi (t), then accordingly update the probability pi (t).

Q learning method using the update pi (t), required for each gene locus associated with an Agent, and the evolution of the corresponding groups of bits of each generation as the environment. The definition of state of the environment, should be able to distinguish the gene locus in the evolutionary process in which the different stages. Therefore, according to the gi (t) and bi (t) to define the relationship between the state.

Greater frequency threshold set θ_{high}, θ_{low} smaller frequency threshold, θ_{diff} for the frequency difference

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.97

threshold, $Agent_i$ the first t generations of the state are divided as follows.

(l)$gi(t) > \theta_{high}$.and $bi(t) > \theta_{high}$, or $gi(t) < \theta_{low}$ and $bi(t) < \theta_{low}$;

(2)$|gi(t)-bi(t)| > \theta_{diff}$;

(3) Otherwise does not meet the above criteria.

$Agent_i$ the first generation of action t set includes the following probability update rules.

(1) Action 1, the probability decreases

$Pi(t+1)=\beta pi(t)$ (0)

(2) Action 2, the probability increases

$Pi(t+1)=1-\beta[1-pi(t)]$ (1)

(3) Action 3, probability values remain unchanged

$Pi(t+1)= pi(t)$ (2)

equation (0) - (2)., i = l, 2, ..., m, β (o <β <l) to adjust the rate. Agenti interact with the environment, you can choose to perform the appropriate action to obtain the next generation of probability pi (t +1). To store Agenti corresponding to a set of Q, the definition of matrix Qi:

$Q_i=[Q_i(s_j,a_k)]_{3\times3}$ (3)

Algorithm for each iteration, each Agenti □-greedy strategy selection in accordance with action, according to the rewards and status of their conversions update the corresponding Q values. If the first t-1 on behalf of the state when the environment Agenti s, choose action α executed, the first generation of environmental state transition t to s', then press the style update Qi (s, a):

$Qi(s, a) \leftarrow Qi(s, a) + \alpha[r_i(t-1)+\gamma \max Qi(s', a') - Qi(s, a)]$ (4)

$$r_i(t-1)=\begin{cases}1, & |pi(t) - gi(t)| < |pi(t-1) - gi(t-1)| \\ -1, & \text{others}\end{cases} \quad (5)$$

Where, ri (t-1) for the Agenti in t-1 obtained on behalf of an immediate return.

Q Learning-Based Estimation of Distribution Algorithm is given below the steps. Algorithm to replace the elitist strategy group to ensure the optimal solution search is not degraded.

Algorithm1: Q Learning-Based Estimation of Distribution AlgorithmInitialization Qi, i = l, 2, ..., m zero matrix, p (1) == (0.5,0.5, ..., 0.5), t = 1;

While (termination condition is not satisfied algorithm) do

According to p (t) sampled individuals generate N-1, and t-1 together constitute the best individual on behalf of the current groups. New individual determination of i-bit value is: generate a random number $\xi \in [0,1]$, if $\xi \le pi(t)$ is taken 1, or take 0;

Calculation of N individuals of fitness function and sorting;

M-choose the best and the worst individuals, the frequency of statistics you get a value of gi (t) and bi (t), i = 1,2, ..., m;

For (i-loci associated Agenti, $1 \le i \le m$)

Recorded before found their t-1 action on behalf of the state s and a, by gi (t) and bi (t) determine the current state s';

By equation (5) calculated an immediate return, according to equation (3.14) update Qi (s, a);

Generate random numbers $\xi i \in [0,1]$, if $\xi i \le$ □, Randomly selected with equal probability of action a ', otherwise select

a '= argmaxQi (s', a');

According to equation (0) - (2) and the action a 'corresponds to the formula, Calculate the new probability value pi (t +1);

End

t←t+1;

End

C Improvement Strategies

The algorithm each update probability pi (t) time, Agenti □-greedy strategy to choose an action, that is, the greater the probability of 1-□ select the maximum Q value of the current state of the corresponding action, but with a smaller probability of randomly selected action □ . As □-greedy strategy is not always accept the best action, but increased the probability of random selection, thus contributing to Agent explore new knowledge, than the greedy strategy with better results.

However, the value of using a fixed □ has some limitations. Especially in the Ageni some time after learning, the current strategy is near optimal, if the probability of 1-□ still randomly choose an action, it will have an impact on the convergence of the algorithm. If you can gradually reduce the evolution □ values, will further enhance the Q study the performance of estimation of distribution algorithms. The use of simulated annealing (SA) algorithm MetroPolis criteria to be able to do this, it is the way by reducing the temperature to gradually reduce the probability of receiving inferior solution.

Here are guidelines for improved use of MetroPolis Q Learning-Based Estimation of Distribution Algorithm

Algorithm2: Improved Q Learning-Based Estimation of Distribution Algorithm

Initialization Qi,, i = 1,2, ..., m zero matrix, the temperature$\tau=\tau 0$, p (1) = (0.5,0.5, ... 0.5), t = 1;

While (termination condition is not satisfied algorithm) do

According to p (t) sampled individuals generate N-1, and t-1 together constitute the best individual on behalf of the current population;

Calculation of N individuals of fitness function and sorting;

M-choose the best and the worst individuals, the frequency of statistics you get a value of gi (t) and bi (t), i = 1,2, ..., m;

For (i-loci associated Agenti, $1 \le i \le m$)

Recorded before found their t-1 action on behalf of the state s and a, by gi (t) and bi (t) determine the current state s';

By (5) calculation of an immediate return, according to equation (4) Update Qi (s, a);

Come to a '= argmaxQi (s', a''), randomly select an action ar, according to the following probability to determine a ':

$$P\{a'=a^r\}=e^{\frac{Q(s,a^r)-Q(s,a^*)}{\tau}} \quad (6)$$

$$P\{a'=a^*\}=1-P\{a'=a^r\} \quad (7)$$

According to equation (0) - (2) and the action a 'new formula to calculate the corresponding probability value pi (t +1);

End

Cool: $\tau \leftarrow \lambda\tau$;

$t \leftarrow t+1$

End

Algorithm to geometric cooling strategy $\tau \leftarrow \lambda\tau$, Where $\lambda \in (0,1)$ is the temperature coefficient。As the temperature decreases , Agenti randomly selected probability of action will become increasingly smaller, When the temperature tends to 0, the strategy is equivalent to the greedy strategy.

III COMPARATIVE EXPERIMENT

A Test Functions

To evaluate the performance of Q Learning-Based Estimation of Distribution Algorithm, The following algorithm using the UMDA, PBIL algorithm, MIMIC algorithm And genetic algorithm function optimization comparative experiments. Select 6 Benchmark test functions for testing. Are the Sphere function, Quadric function, Schaffer function, Griewank function, Rosenbrock function and Rastrigin function. Benchmark functions of these different patterns, with good test performance. Which, Schaffer function, Griewank function and Rastrigin functions are multimodal function, there are a lot of local minima, generally more difficult to find the global optimum algorithm, the algorithm used to test the ability to jump out of local optimum; Rosenbrock function as a single Peak, non-convex pathological function of the flat trend in the value range, the convergence to the global optimal point is remote, can be used to evaluate the efficiency of the algorithm; Sphere function and the Quadric function is a single peak function, can test the accuracy of optimization algorithm, Examine the implementation of the algorithm performance.

B Experimental Analysis

After several tests, Q Learning-Based Estimation of Distribution Algorithm parameters are as follows: frequency threshold wereθ_{high} = 0.75, θ_{low} = 0.25, θ_{diff}== 0.35, select the rate of γ = 0.2, probability adjusted rate β = 0.9, learning factor a = 0.2, discount factor = 0.9. For comparison UMDA PBIL algorithm selection algorithm and the rate is taken 0.2, PBIL learning rate algorithm to take 0.1; MIMIC algorithm selection ratio is 0.4; genetic algorithm uses single point crossover, crossover rate 0.7, mutation rate 0.1; their Elitist strategy are used. All of the above algorithm population size are set to 50, the termination condition for the search to the global optimum or the maximum evolution generation T, take T = 200.

Taking into account all the above algorithm has a certain randomness, use them to function f1-f6 are independently tested 50 times, the experimental results shown in the table. Among them, Table 1 for each algorithm the number of global optimal value obtained, Table 2 shows the results of 50 runs on average, standard deviation and worst values, Table 3 shows the average running time of each method (unit: Seconds). Table □-QEDA and M-QEDA represent the use of strategies and MetroPolis □-greedy Q Learning-Based Estimation of Distribution Algorithm, the former taking □ = 0.1, initial temperature of the latter: τ0 = 50, temperature coefficient of λ = 0.9.Table 1 Algorithm for number of times the global optimal value obtained

Function	GA	UMDA	PBIL	MIMIC	□-QEDA	M-QEDA
Sphere	22	27	29	31	42	50
Quadric	23	17	31	24	45	50
Schaffer	11	11	27	12	40	50
Griewank	28	24	34	29	45	50
Rosenbrock	14	31	49	28	46	50
Rastrigin	19	21	35	14	44	50

Table 2 The results of the algorithm is run 50 times the mean, standard deviation and worst values

Function		GA	UMDA	PBIL	MIMIC	□-QEDA	M-QEDA
Sphere	average	1.1200E-03	9.2000E-04	8.4000E-04	4.5600E-03	3.2000E-04	0
	Standard deviation	1.0029E-03	1.0069E-03	9.9714E-04	1.9290E-02	7.4066E-04	0
	Worst value	2.0000E-03	2.0000E-03	2.0000E-03	1.2800E-01	2.0000E-03	0
Quadric	Average	0.5166	1.1365	0.1091	0.3731	0.0287	0
	Standard deviation	2.5801	3.4736	0.1407	1.1744	0.0870	0
	Worst value	18.3680	23.2470	0.2870	7.7150	0.2870	0
Schaffer	average	5.7299E-03	3.1967E-03	5.5335E-04	4.1733E-03	3.8145E-OS	0
	Standard deviation	5.0882E-03	4.1418E-03	1.5937E-03	4.9006E-03	7.7064E-05	0
	Worst value	1.2625E-02	1.4193E-02	9.5638E-03	1.8953E-02	1.9072E-04	0
Griewank	average	0.3550	0.4627	0.2582	0.3478	0.0807	0
	Standard deviation	0.4045	0.4572	0.3802	0.4151	0.2445	0
	Worst value	0.8068	1.1645	0.8068	1.0400	0.8068	0
Rosenbrock	average	5.9197	0.5817	0.0038	1.3108	0.0152	0
	Standard deviation	7.7090	1.4511	0.0266	4.0439	0.0521	0
	Worst value	19.5757	7.7538	0.1881	19.5757	0.1957	0
Rastr	average	9.3978	3.4206	0.1190	2.3572	0.0476	0

igi n	Standard deviation	14.5125	6.213 3	0.1836	5.6836	0.130 2	0
	Worst value	79.9851	20.78 51	0.3967	20.0000	0.396 7	0

Table 3 The average running time of each algorithm

function	GA	UMDA	PBIL	MIMIC	?-QEDA	M-QEDA
Sphere	1.0913	1.4806	1.3759	7.5172	1.3169	0.6519
Quadric	1.1619	2.1466	1.2916	10.0963	1.1741	0.6056
Schaffer	1.0669	0.3894	0.2441	0.9788	0.2641	0.1656
Griewank	0.8034	0.8894	0.5781	3.1812	0.6056	0.3887
Rosenbrock	1.2272	1.2347	0.1844	8.6994	1.1469	0.5784
Rastrigin	1.1497	1.9047	1.0859	13.7091	1.2369	0.9159

It can be seen, both the traditional genetic algorithm, or UMDA, PBIL and MIMIC other existing distribution algorithm, function optimization in solving these complex problems are not easy to search the global optimum value. Which, PBIL search success rate of slightly higher average of 68%, other 3, the algorithm less than 50%. Q Learning-Based Estimation of Distribution Algorithm are demonstrated excellent performance, especially after using Metropolis criterion and, for the 6 functions are 100% Benchmark global optimal value obtained. In the algorithm execution time, the dual variable associated MIMIC worst performance of the algorithm, generally longer than the other algorithms 5-10 times; and M-QEDA algorithm in addition to solving the Rosenbrock function as PBIL algorithm, but in other cases have shown The best time performance.

REFERENCES

[1] Sergios T, Konstantinos K. Pattern Recognition, Second Edition [M]. San Diego: Academic Press. 2003.

[2] Dasgupta D, ,Forrest S. Artificial immune systems and their applications [M]. Berlin: Spring-Verlag 1998

[3] Slowinski R, Hapke M. Scheduling under fuzziness [M]. New York: Physica-Verlag, 2000.

[4] Seber G A R Linear regression analysis [M]. New York: John Wiley, 1977.

[5] Deng J L. Introduction to grey system theory (J]. The Journal of Grey System, 1989, 1(1):

Chaotic Relaxation Asynchronous Iterative Parallel Algorithm and Its Implementation

LI Wenjing, YANG Wen, LI Shuju
(College of Computer Science and Information Engineering, Guangxi Teachers Education University, Nanning ,530001, China)

***Abstract*: In order to solve the system load imbalanced problem of the synchronized parallel algorithm on unconstrained optimization problems, improve the speed of solving problems. Chaotic relaxation asynchronous iterative parallel algorithm was proposed in PC cluster. First, the basic concepts of unconstrained optimization problems were introduced, and The BFGS Parallel algorithm of unconstrained optimization problems was analyzed. Then according to relaxation and asynchronous iteration theory, chaotic relaxation asynchronous iterative convergence properties and proof were given. Chaotic relaxation asynchronous iterative parallel algorithm was proposed in PC cluster. Then the algorithm is applied to unconstrained optimization problems, The BFGS chaotic relaxation asynchronous iterative parallel algorithm of unconstrained optimization problems was given. At last, the experimental results of PC cluster show that this algorithm better solved the system load imbalanced problem, improve the speed of problem solving.**

Keywords*: *PC Cluster; Unconstrained Optimization Problems; Chaotic Relaxation; Asynchronous Iterative; Parallel Algorithm

1. INTRODUCTION

Many practical problems such as industrial and agricultural production, engineering technology, transportation, production management, economic plan, national defense, finance, etc, can be abstractly transformed into optimization problems. The solution of small-scale optimization problems can be achieved by means of serial solution. However, with respect to the research and sensitivity analysis of weather and climate as well as large-scale nonlinear optimization solution of the petroleum, geological survey, astronomy, such serial algorithm has outpaced the computation time limit people can bear and is meaningless in terms of solving practical problems. Using synchronous parallel algorithm is one way to reduce the computation time, but there exists system load imbalance and operation efficiency is to be improved. Therefore, the paper proposes a disordered relaxation asynchronous iterative parallel algorithm under PC a cluster environment, and applying the algorithm to solving unconstrained optimization problems.

*Supported by National Natural Science Foundation of China (60864001), Department of Education of Guangxi Zhuang Autonomous Region(200911MS144)

2. UNCONSTRAINED OPTIMIZATION PROBLEMS AND BFGS PARALLEL ALGORITHM

2.1 Constrained optimization problem

Let the function f, g_i, h_j: $R^n \to R$, $i=1,\ldots,m_1, j=m_{1+1},\ldots,m_{1+p}$ Continuously differentiable ,Constrained optimization problem

$$\min f(x) \quad (1)$$

$$g_i(x) \geqslant 0, i \in I = \{1,\ldots,m_1\} \quad (2)$$

$$h_j(x) = 0, j \in E = \{m_{1+1},\ldots, m_{1+p}\} \quad (3)$$

The feasible domain is $D=\{x \in R^n \mid g_i(x) \geq 0, i \in I, h_j(x)=0, j \in E\}$ Which $x=(x_1,x_2,\ldots,x_n) \in R^n$ is the n-dimensional vector, f, g_i, h_j: $R^n \to R$, $i=1,\ldots,m$ as a function of x, est. (Subject to) indicate x subject to constraints (2), (3) restrictions; $f(x)$ called the objective function, (2) as inequality constraints, (3) as equality constraints, If $m_{1+p}=0$ then constrained optimization problem back into a unconstrained optimization problem [1].

Unconstrained optimization problem containing Descent algorithm、Steepest descent algorithm、Newton method、Simulation of Newton method、Conjugate gradient method direct method、Direct method and Rely on domain method and so on. Among them, through BFGS(Broyden-Fletcher-Goldfarb-Shanno) Correction formula to solve nonlinear optimization problems without restraint of algorithm , Is called BFGS method, its one of many algorithm like simulation of Newton method, Has the following three characteristics[2-4]:

(1) choose any approximation matrix B_k replace $\triangle^2 f(x), \triangle^2 f(x)$ means the Hessian matrix at the point x, make corresponding algorithm produces direction is approximation Newton direction, to ensure that the algorithm has higher convergence speed , and algorithm under certain conditions have super linear convergence speed.

(2) For all the k, the B_k is symmetrical and positive definite matrix, Thus the direction is caused by the algorithm of function f the decline in x_k direction;

(3) Matrix B_k easy calculation.

2.2 BFGS synchronous parallel algorithm and analysis

2.2.1 BFGS synchronous parallel algorithm

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.28

BFGS synchronous parallel algorithm mainly by B constitutes, synchronous parallel solution of linear equations、Synchronous parallel solving step length and Parallel solution BFGS correction formula which is composed of three parts, Its main algorithm is as follows:

Input: Initial solution vector $x^{(0)}$ and the value of a given solution accuracy ε;

Output: The optimal solution $x^{(K+1)}$.

Step1: Make p is the number of parallel processing, accuracy $\varepsilon>0$，let $k=0$，take the initial point $x^{(0)}\in R^n$, in the master node calculation norm $\|\nabla f(x)\|$，if $\|\nabla f(x)\|\leq\varepsilon$，then the algorithm terminates，the solution of the problem was $x^{(0)}$.Otherwise, make the initial symmetric positive definite matrix $B_0\in R^n\times^n$ as the unit matrix.

Step2: Using row wrapped interleaved pattern of parallel CHOLSKY decomposition algorithm solving linear equations，By synchronization iteration method solution a decline direction in the vector $d^{(k)}$

$$Ly=-\nabla f(x^{(K)}) \text{ and } ULd=y \qquad (4)$$

Step3: Through the synchronous parallel Wolfe - Powell type inexact linear searching to sure the step size α_k.

Step4:Make $k=k+1$，In each node parallel computing n/p a element $x^{(K+1)}= x^{(K)} + \alpha_k d^{(k)}$ returned after master node. The master node calculation norm $\|\nabla f(x^{(K+1)})\|$ value, if$\|\nabla f(x^{(K+1)})\|\leq\varepsilon$，then solution is $x^{(K+1)}$. Otherwise, parallel solution BFGS correction formula calculated B_{k+1}，turn to Step2.

2.2.2 Synchronous parallel algorithm analysis

In the process of realizing BFGS synchronous parallel algorithm, the main operations include:

Step 1: Solving linear equations, Find out the function f solution vectors in x_k place drop direction. Based on linear equations of the coefficient matrix are the characteristics of symmetric positive, Using CHOLESKY decomposition method synchronous parallel solving linear equations, and Proofs see literature [3]. If PC cluster have N process involved in calculation, solution vectors have n components, then the synchronous parallel average calculation time is: $n(2n+1)/2p$.

Step 2: Adopt Wolfe - Powell inexact linear search algorithm, from point x_k and its drop direction to solving the step length α_k. Due to parallel solving process, main node has repeatedly waited for each of the intermediate results returned from node process; there will be many times synchronous waiting processes. So the synchronous parallel average calculation time is: $(4n-1)/p$.

Step 3: Solving BFGS correction formula process. And divided into three parts processing, the first part is B_k, The second part is $\frac{B_k s^{(k)} s^{(k)} B_k}{s^{(k)\mathrm{T}} B_k s^{(k)}}$, The third part is $\frac{y^{(k)} y^{(k)\mathrm{T}}}{y^{(k)\mathrm{T}} s^{(k)}}$. The parallel algorithm's time is about $2n^3/p$.

This algorithm operations process has two steps to contain synchronous iterative process, synchronous iterative lead to parallel performance reduced. For synchronous parallel computing, need to set a global synchronous disabled bar, Make all process of their respective calculation in complete, must be disabled after waiting for all processes of gate point arrived. When all processes arrived, and then allows each process to start from disabled gate next iteration value calculation. Therefore, the synchronous parallel algorithm can cause fast node in the idle waiting, system load is not balanced, lead to system operation efficiency decrease. For this, proposed chaotic relaxation asynchronous iterative parallel algorithm, solve synchronous parallel deficiency.

3. CHAOTIC RELAXATION ASYNCHRONOUS ITERATIVE PARALLEL ALGORITHM

3.1 Asynchronous iterative parallel method

In the iteration parallel processing, if cancel disabled barrier (synchronous), let finish calculation node without waiting for this iteration results come from other processes, but use values that get it from last iteration, even the values which is from the former several iterations, and continue to move forward execution, and every participate in operation processes are in the same way, until the end run. So that you can overcome synchronous parallel algorithm, fast node in the idle waiting, system load is not balanced, lead to system operation efficiency decrease.

Definition 1 In the iteration parallel processing, many processes without waiting for others this iteration result, the method which is directly use the iteration value with first or the last few times and continue forward iterative calculation, called asynchronous iterative parallel algorithm.

3.2Relaxation iterative and condition of convergence

Considering the convergence of asynchronous iterative parallel algorithm, when the process using before several times iterative value must satisfy the chaotic relaxation convergence conditions. Relaxation iteration method, formula and convergence conditions as follows:

Definition 2 set the equations

$$Ax=b \qquad (5)$$

$A\in R^n\times^n$ Is nonsingular matrices, and let $a_{ii}\neq0(i=1,\ldots,n)$, A decomposed to

$$A=D-L-U \qquad (6)$$

There is following formula:

$$\begin{cases} x_i^{(k+1)}=\dfrac{\omega}{a_{ii}}\left(b_i-\sum\limits_{j=1}^{i-1}a_{ij}x_j^{k+1}-\sum\limits_{j=i+1}^{n}a_{ij}x_j^{(k)}\right) \\ x^{(k)}=(x_1^{(k)},x_2^{(k)},\cdots,x_n^{(k)})^T,(k=0,1,\cdots;i=1,2,\cdots,n) \end{cases} \qquad (7)$$

When the flabby factor is $\omega<1$, called low relaxation method, when it flabby factor is $\omega>1$, called over relaxation method.

Theorem 1 Use relaxation iteration method solving linear equations （5）,the sufficient and necessary conditions are:

$$\rho(L_\omega)<1 \qquad (8)$$

There is $L_\omega=(D-\omega L)^{-1}((1-\omega)D+\omega U)$, says matrix L_ω for relaxation iteration matrix. $\rho(L_\omega)<1$ represent iterative matrix of the spectral radius less than 1. Proofs of the theorem see reference [4].

Theorem 2 assuming solutions the linear equations (5) of relaxation iteration method is convergence, so $0<\omega<2$. Theorem 2 illustrates when the use of relaxation iteration method for solving linear equations, the value of relaxation factor must be (0, 2) range to converge. But when the matrix A is symmetric positive definite matrix, there is $0<\omega<2$, a certain convergence of relaxation method.

Deduction If matrix A of equations (5) is symmetric positive definite and $0<\omega<2$, then solutions the linear equations (5) of relaxation iteration method is convergence.

Theorem 3 There is a fixed positive integer s, Makes when the request $k \in \{k \mid k> = s$, and $k <= n$, n is the maximum number of iterations$\}$ iterations, Process cannot use any value from formation before the *k-s* times iteration.

Proof: Assuming S is a fixed positive integer, the process can use *k-s* times before any value formation in the iteration. From the formula (6) and formula (7) can launch below iterative formula:

$$x^{(K+1)}=L_\omega x^{(K)}+f$$

There is $L_\omega=(D-\omega L)^{-1}((1-\omega)D+\omega U)$, $f=\omega(D-\omega L)^{-1}b$. According to the title mean there is $x^{(K+1)}=L_\omega x^{(k-s)}+f$, then $L_\omega(x^{(K)}-x^{(k-s)})=0$, Thus, we obtain $s=0$ and s is a positive integer. So we conclude that contradictions.

Definition 3 Meet the Theorem 3 of convergence conditions with relaxation iteration method, we called non-relaxation asynchronous iterative parallel algorithm.

3.3Algorithm Design Equations

2.3.1 Design Strategy

We assume that there are P processors in PC cluster environment, and the matrix of linear equations $A \in R^{n\times n}$, $a_{ii} \neq 0(i=1,\ldots,n)$ is a square matrix of order n, then according to the above definitions and theorems, the design strategies of chaotic relaxation asynchronous iterative parallel algorithm are:

(1) Data divided. Per processor deals with n/p components of solution vector averagely (let n can been divisible by p), for example, processor P_0 is responsible for calculating value from x_0 to $x_{(n/p)-1}$, P_1 is from $x_{n/p}$ to $x_{(2n/p)-1}$, and so on.

(2) As long as the result is not met the iteration terminal condition$|x_i^{(K+1)}-x_i^{(K)}|<\varepsilon$ when every solution vector is calculated by any process or processor, we continue to calculate the next solution component or next iteration with any value directly which is formed in the ($j \in \{j \mid j \geq i-s$, and $j<i\}$) iteration by current processor other processors. There's no need to wait for the current result of other processes.

(3) The component of the solution obtained by iteration every time needs to set one "time stamp" which is used to show which time the value is generated. We save the solution vectors in the form of two-dimensional array.

2.3.2 Parallel algorithm

The chaotic relaxation asynchronous iterative parallel algorithm is as follows:

Input: coefficient matrix $A[n][n]$, vector $b[n]$, initial solution vector $x^{(0)}$, the maximum iteration times, the precision values ε.

Output: solution vector $x^{(k+1)}$

Step1：The main process P_0 reads matrix array, vector and other data from the file, and input the initial solution vector, the number of iterations, the precision values ε by keyboard. Then broadcast these data to every subordinate process.

Step2: The subordinate process P_i receive data from the main process

(a) For (k=1; $i<$ *the maximum iteration times*;k++)

(b) For (i=*low*; $i<=$*high*; i++)// the low and high show respectively the minimum and maximum subscripts of all components which is handled by the subordinate process//;

{

(i)Calculate the value which is i-vector component of $x_i^{(K+1)}$ interacted for the (k+1) times according to the iterative formula (8), and broadcast it to every process.

(ii) Determine whether $x_i^{(K+1)}$ met the terminate condition: $\max\limits_{1\leq i\leq n} | x_i^{(K+1)}- x_i^{(K)}|<\varepsilon$, if met ,the component is used as a mark of end , it will no longer operate for the next iteration. If not, store the value.

}

Step3: The subordinate process P_i sent the result to the main process P_0, P_0output the solution vector.

2.3.3 Achieve Parallel algorithm

We by the following solutions of linear equations, verify the feasibility and effectiveness of the algorithm, for example, solving linear equations:

$$\begin{cases} -4x_1+x_2+x_3+x_4=1 \\ x_1-4x_2+x_3+x_4=1 \\ x_1+x_2-4x_3+x_4=1 \\ x_1+x_2+x_3-4x_4=1 \end{cases}$$

The precise solution for $x^*=(-1,-1,-1,-1)^T$.

When take slack factor $\omega=1.3$, precision value is $\varepsilon=0.46\times10^{-5}$, Single sequential program 11th iterations results is

$x^{(11)}=(-0.99999646,-1.00000310,-0.99999953,-0.99999912)^T$

PC cluster environment, parallel platform consisting of 4 PCs, application of chaotic relaxation asynchronous iterative parallel algorithm calculation of 11th iterations results is $x^{(11)}$=(-0.99987953,-1.00010247,-0.99989932,-0.99997531)^T.

These results indicate that in the case of the same number of iterations, non- relaxation asynchronous iterative parallel algorithm solution accuracy obtained was slightly lower, convergence is relatively slow. There are two main reasons:

(1)Fewer number of variables that solving, Fail to reflect the advantages of parallel; (2)When the parallel algorithm are solving, need to send and receive iteration value each other, when no receives this iteration value, used before twice iterations value. In the same iterations, the convergence speed to be slower than the Single. However, if the solution is large linear equations, the advantage of parallel algorithm will manifest. Through the example calculation, although slightly slower convergence of the algorithm, but demonstrate that the algorithm is feasible and effective.

4. THE APPLICATION OF THE METHOD TO THE UNCONSTRAINED OPTIMIZATION PROBLEMS

4.1 Improvement for BFGS synchronized parallel algorithm

BFGS synchronized parallel algorithm include synchronized parallel solution of linear equations, synchronized parallel reducing the step size, and parallel computing the amendment formulas of BFGS. Synchronization computing is used very obvious in the first and second steps, the parallel computing of BFGS modification formula include the multiplication of matrix and vector, which implicit parallel computing too. Therefore, we mix the relaxation and asynchronous iteration parallel algorithm and strategy, in the many PC cluster, we designed the algorithm as follow:

Input: initial solution vectors $x^{(0)}$ and the given precision of the solution ε.

Output: optimized solution $x^{(k+1)}$.

Step 1: The master process P0 receive initial solution vectors $x^{(0)} \in R^n$ and the precision ε>0,let k=0,calculate norm $\| \triangle f(x)^{(0)} \|$,if $\| \triangle f(x)^{(0)} \| \leq \varepsilon$, terminate the algorithm, and the result is obtained $x^{(0)}$.Otherwise, resets initial symmetric positive definite matrix $B_0 \in R^{n \times n}$ to an identity, then pass B_0, $x^{(0)}$,ε into all of the child process Pi（i=1,2,…,p）.

Step 2: each child process receives the data from master process, and employ the row wrapped interleaved pattern of data concurrently to solve the linear simultaneous equations:

$$B_k d + \nabla f(x^{(k)}) = 0$$

Coefficient matrix B element of a set.

Step 3: The two equations from where (4) use chaotic relaxation asynchronous iteration parallel algorithm to solve the vector of direction of descent $d^{(k)}$.

Step 4: Employing the non-exactitude linear search of chaotic relaxation and asynchronous iteration parallel Wolfe-Powell to get the step size.

Step 5: Let k=k+1, after calculating n/p elements in each node by chaotic relaxation and asynchronous iteration parallel algorithm, send it back to the principal nodes. Calculate the solution of norm $\| \square f(x)^{(k+1)} \|$ in principal nodes，if$\| \square f(x)^{(k+1)} \| \leq \varepsilon$，the result is $x^{(k+1)}$.Otherwise，employ BFGS amendment formula to solve the solution of B_{k+1} by chaotic relaxation and asynchronous iteration parallel algorithm，then go to Step 2.

4.2 BFGS relaxation asynchronous iterative parallel algorithm analysis

This article mainly analysis from the computational complexity. Let B is the n order matrix, have P processors(node),then each processor the computing task is n / p rows, taking into account B for the n order symmetric positive definite matrix, Using row wrapped interleaved divide method used to divided the processor of tasks, processor load is relatively balanced.

When in the solution of linear equations (4) and by Wolfe-Powell type inexact line search step, running chaotic relaxation asynchronous iterative parallel algorithm to calculation. The whole time complexity of the parallel algorithm and as follows three main components:

(1) The time of to solve the linear equations vector: synchronous parallel algorithm of time is T= n (2n+1)/2p; The relaxation asynchronous parallel algorithm calculation time is Tp =n (n+3)/2p.

(2)The time of the computation search step: parallel algorithm computing time is T= (4n-1)/p; the relaxation asynchronous iterative parallel algorithm calculation time is Tp2= (2n+1)/p.

(3) The time of calculation BFGS correction formula: parallel algorithm computing time is $T \approx 2n^3/p$; The relaxation asynchronous iterative parallel algorithm calculation time is $Tp \approx n^3/p$.

From the above analysis, relaxation asynchronous parallel algorithm calculation time less than synchronous parallel, if the meta-variable function f is large enough, the computing speed of BFGS chaotic relaxation asynchronous parallel algorithm is faster than Synchronous Parallel.

5. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental environment including four sets RAM256 MB CPU2.5 GHz, isomorphism PC cluster, Linux8.0+MPI, C programming language; n is a function of the size of variable dimension, p that the number of nodes. The result was shows on Table1 and Table2.

Table 1 Question 1 of the computation time (seconds)

n	p=2		p=4	
	Synchroniz	Asynchron-	Synchroniz	Asynchron-

	-ation	ous	-ation	ous
400	8.347 211	7.583247	6.239 653	5.026943
1000	16.554 893	14.83796	13.407 536	10.403583
2000	43.459 147	39.72302	29.534 102	25.65307

Table 2 Question 2 of the computation time (seconds)

n	*p*=2		*p*=4	
	Synchroniz -ation	Asynchron- ous	Synchroniz -ation	Asynchron- ous
400	33.813 523	31.553472	23.144 065	19.53784
1000	47.213 906	41.902478	31.125 713	27.25473
2000	66.037 254	60.127315	45.041 233	38.94573

Question 1 of the test function as follows:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n/2} \{100(\boldsymbol{x}_{2i} - \boldsymbol{x}_{2i-1}^2)^2 + (1 - \boldsymbol{x}_{2i-1})^2\}$$

Question 2 of the test function as follows:

$$f(\boldsymbol{x}) = (\sum_{i=1}^{n} i\boldsymbol{x}_i^2)^2$$

By checking results that relaxation asynchronous parallel algorithm calculation time less than synchronous parallel, and the theoretical calculation of the number of steps less consistent, solve the synchronization problem of load imbalance in parallel; improve the speed to solve the problems.

6.CONCLUSION

Based on the unconstrained optimization problems of BFGS synchronous parallel algorithm and system load disequilibrium problem caused by the fact that the synchronous iterative parallel iterative process must synchronously wait, this paper puts forward the asynchronous iteration method using disorder flabby, solving unconstrained optimization problems BFGS algorithms, linear equations, and the step length reconciliation BFGS correction formula synchronization problems existing in the construction of without constraint optimization problem. It builds BFGS disorder parallel algorithms, and this algorithm enhances solving speed of large-scale unconstrained optimization problems. The application of constrained nonlinear optimization problems are more complicated than that of the unconstrained optimization problems, thus the next step is to focus on the main research work of applying algorithm to the improvement of the solving speed of this kind of problems.

REFERENCES

[1] Dai Yuhong. Convergence Properties of the BFGS Algorithm[J]. SIAM Journal on Optimization, 2002, 13(3): 693-701.

[2] Li Donghui, Fukushima M. A Modified BFGS Method and Its Global Convergence in Nonconvex Minimization[J]. Journal of Computational and Applied Mathematics, 2001, 129(1): 15-35.

[3] Li Donghui, Fukushima M. On the Global Convergence of the BFGS Method for Nonconvex Unconstrained Optimization Problems[J]. SIAM Journal on Optimization, 2000, 11(4): 1054-1064.

[4] Li Dong Hui, Tong Xiao Jiao, Wan Zhong. Number Optimization (M). BeiJing : Science Publishing House, 2005: 167-175.

[5] Li Wenjing, Wang Ruliang, Liao Weizhi. The BFGS Parallel Algorithm of Unconstrained Optimization Problems and Its Implementation [J]. Computer Engineering, 2009, 35(15):58-63.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Coupling the Parareal algorithm with the Waveform Relaxation method for the solution of Differential Algebraic Equations

Thomas Cadeau and Frédéric Magoulès
Applied Mathematics and Systems Laboratory
Ecole Centrale Paris
Châtenay-Malabry, France
Email: frederic.magoules@hotmail.com

***Abstract*—In this paper, a coupling strategy of the Parareal algorithm with the Waveform Relaxation method is presented for the parallel solution of differential algebraic equations. The classical Waveform Relaxation (in space) method and the Parareal (in time) method are first recalled, followed by the introduction of a coupled Parareal-Waveform Relaxation method recently introduced for the solution of partial differential equations. Here, this coupled method is extended to the solution of differential algebraic equations. Numerical experiments, performed on parallel multicores architectures, illustrate the impressive performances of this new method.**

***Keywords*-Differential Algebraic Equations; Domain decomposition method; Waveform Relaxation method; Parareal algorithm; Parallel and distributed computing**

I. Introduction

The solution of Differential Algebraic Equations (DAE) appears in various industrial applications such as systems or circuits stability simulation, chemical reaction, etc. For instance, in systems analysis, large interconnected networks lead to the solution of large DAE systems of hundreds of thousands of variables, while industrial softwares are currently designed for the calculation of systems of tens of thousands of variables. In circuits analysis, the same software limitations occurs, but is even more critical since DAE systems of millions of variables are needed to obtain an accurate solution. With the significant development of multiprocessors machines, the use of new algorithms designed for parallel and distributed processing appears very promising.

In this paper, after a brief presentation of the mathematical problem to solve and the linearization of the associated equations, the Waveform Relaxation in space method [1] is presented. The Parareal in time method [2] is then introduced, followed by a coupled Parareal-Waveform Relaxation method, originally introduced in [3] for Partial Differential Equations (PDE). This method is here extended to DAE in a parallel computing paradigm. This new method leads by construction to a higher speedup compared to the Waveform Relaxation in space method and the Parareal in time method. Numerical experiments, performed on parallel multicores architectures, illustrate the performance of the proposed method.

II. Differential Algebraic Equations

The mathematical system to solve is a DAE system of the form:

$$\left\{ \begin{array}{rcl} y' & = & f(y,z,t),\ t\ \in [0,T] \\ 0 & = & g(y,z,t) \end{array} \right. \tag{1}$$

where f and g are functions of the differential variables, y, the algebraic variables, z, and the time variable, t. Due to some properties of the physical problem, the functions f and g are non-linear functions, non differentiable, neither continuous. The system is considered as a rough problem, not only due to the discontinuities, but also because of its stiffness. The interesting and important point is that the functions f and g are smooth enough between two discontinuities (like in systems or circuits), which allows to consider linear approximation of f and g in such interval. Fixing the functions f_0 and g_0 at $t = 0$, the system to solve becomes the following:

$$\left\{ \begin{array}{rcl} y' & = & M\delta y + N\delta z + f_0,\ t\ \in [0,T] \\ 0 & = & P\delta y + Q\delta z + g_0, \end{array} \right. \tag{2}$$

with $\delta y = y - y_0$, $\delta z = z - z_0$, where y, z are the states at time $t \in [0,T]$, and y_0, z_0 are the states at time $t = 0$. The integration scheme used for the differential equations is a backward differentiation formula (BDF), using a trapezoidal rule. The implicit system to solve can be written as:

$$A \left(\begin{array}{c} \delta y_{n+1} \\ \delta z_{n+1} \end{array} \right) = B \left(\begin{array}{c} \delta y_n \\ \delta z_n \end{array} \right) + b \tag{3}$$

where:

$$A = \left(\begin{array}{cc} I - \frac{h}{2}M & -\frac{h}{2}N \\ P & Q \end{array} \right),$$

$$B = \left(\begin{array}{cc} I + \frac{h}{2}M & \frac{h}{2}N \\ 0 & 0 \end{array} \right), \quad b = \left(\begin{array}{c} hf_0 \\ -g_0 \end{array} \right).$$

Solving the system Eq. (3) implies a matrix-vector multiplication by the matrix B and the solution of a linear system with the matrix A. The matrices A and B, and the vector b depend only of the timestep h, chosen at the begin of the computation. Consequently, A, B and b are fixed for the simulation and do not require additional computation. Once the right-hand side (rhs) of Eq. (3) evaluated, a direct

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.34

method is used to solve the linear system with the matrix A. The factorization of the matrix A is done at the beginning of the computation and only a forward/back substitution is performed at each timestep. Finally, the solution is rebuilt knowing the initial values:

$$\begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} \delta y_{n+1} \\ \delta z_{n+1} \end{pmatrix} + \begin{pmatrix} y_0 \\ z_0 \end{pmatrix}. \quad (4)$$

The following algorithm details the different steps of the sequential solver based on the lower/upper factorization and the corresponding forward/back substitution.

Lower/Upper factorization of matrix A
while $t_n < T$ **do**
  Compute the rhs, Eq. (3)
  Forward/back substitution, Eq. (3)
  Update the solution, Eq. (4)
end while

III. Waveform Relaxation Method

Domain Decomposition methods [4]–[7] are well suited for parallel computations. Indeed, the division of a problem into smaller sub-problems, through artificial subdivisions of the domain/matrix, is a means for introducing parallelism. Domain decomposition strategies include in one way or another the following ingredients: (i) a decomposer to split a domain/matrix into sub-domains/sub-matrices using different heuristics; (ii) local solvers (direct or iterative, exact or approximate) to find solutions for the sub-problems for specific boundary conditions on the interface; (iii) interface conditions [8], [9] (weak or strong) enforcing compatibility and equilibrium between overlapping or non-overlapping sub-domains/sub-matrices; (iv) a solution strategy for the interface problem [10], [11]. The differences between the methods lie in how those ingredients are actually put to work and how they are combined to form an efficient solution strategy for the problem at hand.

To ensure good performances of domain decomposition methods, one first point is to find an appropriate partitioning of the domain/matrix. The configurations to avoid are: (i) a bad distribution of the variables which leads to singular sub-matrices or which affects the convergence; (ii) a bad loadbalancing of the variables between the sub-domains/sub-matrices which causes strong unbalance computation times between the processors; (iii) large interfaces size which leads to high communications time. Given a sub-domain/sub-matrix named (s), the indexes of the variables belonging to the internal area of the sub-domain/sub-matrix or to the overlapping area of the neighboring sub-domain/sub-matrix are named (i_s). The indexes of the variables belonging to the neighboring sub-domain/sub-matrix in contact with the indexes of the variables (i_s) are named (b_s). Each sub-system s depends on the variables $(\delta_y^{(b_s)}, \delta_z^{(b_s)})^t$ which are computed by the neighboring sub-system. The global system Eq. (2) is projected on the partitioning of the domain/matrix and Dirichlet conditions are added on the interface between the sub-domains/sub-matrices. With these notations, each local sub-problem can be written as:

$$\begin{pmatrix} y^{(i_s)'} \\ 0 \end{pmatrix} = J^{(i_s)} \begin{pmatrix} \delta y^{(i_s)} \\ \delta y^{(b_s)} \\ \delta z^{(i_s)} \\ \delta z^{(b_s)} \end{pmatrix} + \begin{pmatrix} f_0^{(i_s)} \\ g_0^{(i_s)} \end{pmatrix} \quad (5)$$

where $J^{(i_s)} = \begin{pmatrix} M^{(i_s)} & M^{(b_s)} & N^{(i_s)} & N^{(b_s)} \\ P^{(i_s)} & P^{(b_s)} & Q^{(i_s)} & Q^{(b_s)} \end{pmatrix}$.

Using the same numerical scheme than for the sequential solver (algorithm described on Section II), each sub-system can be written as follows:

$$A^{(i_s)} \begin{pmatrix} \delta y_{n+1}^{(i_s)} \\ \delta z_{n+1}^{(i_s)} \end{pmatrix} = B^{(i_s)} \begin{pmatrix} \delta y_n^{(i_s)} \\ \delta z_n^{(i_s)} \end{pmatrix} + b^{(i_s)} \quad (6)$$

where $A^{(i_s)}$, $B^{(i_s)}$, $b^{(i_s)}$ are defined by:

$$A^{(i_s)} = \begin{pmatrix} I^{(i_s)} - \frac{h}{2} M^{(i_s)} & -\frac{h}{2} N^{(i_s)} \\ P^{(i_s)} & Q^{(i_s)} \end{pmatrix},$$

$$B^{(i_s)} = \begin{pmatrix} I^{(i_s)} + \frac{h}{2} M^{(i_s)} & \frac{h}{2} N^{(i_s)} \\ 0 & 0 \end{pmatrix},$$

$$b^{(i_s)} = \begin{pmatrix} h f_0^{(i_s)} \\ -g_0^{(i_s)} \end{pmatrix} + \begin{pmatrix} hM^{(b_s)} & hN^{(b_s)} \\ -P^{(b_s)} & -Q^{(b_s)} \end{pmatrix} \begin{pmatrix} \delta y_{n+1}^{(b_s)} \\ \delta z_{n+1}^{(b_s)} \end{pmatrix}.$$

Knowing $(\delta y_{n+1}^{(b_s)}, \delta z_{n+1}^{(b_s)})^t$ the sub-problem can be solved directly by the forward/back substitution. The difficulty is that $(\delta y_{n+1}^{(b_s)}, \delta z_{n+1}^{(b_s)})^t$ is not known and has to be computed by the other sub-systems. Thus, the Waveform Relaxation algorithm introduces an iterative scheme, with an iteration number k, as:

$$A^{(i_s)} \begin{pmatrix} \delta y_{n+1}^{(i_s),k+1} \\ \delta z_{n+1}^{(i_s),k+1} \end{pmatrix} = B^{(i_s)} \begin{pmatrix} \delta y_n^{(i_s),k} \\ \delta z_n^{(i_s),k} \end{pmatrix} + b^{(i_s),k} \quad (7)$$

where

$$b^{(i_s),k} = \begin{pmatrix} hM^{(b_s)} & hN^{(b_s)} \\ -P^{(b_s)} & -Q^{(b_s)} \end{pmatrix} \begin{pmatrix} \delta y_{n+1}^{(b_s),k} \\ \delta z_{n+1}^{(b_s),k} \end{pmatrix} + \begin{pmatrix} h f_0^{(i_s)} \\ -g_0^{(i_s)} \end{pmatrix}.$$

Each local sub-problem is allocated to a different processor and is solved in parallel. Communications between processes occur at each iteration since process s needs to receive $\delta y_{n+1}^{(b_s),k}$ and $\delta z_{n+1}^{(b_s),k}$. The set of these timesteps is named a window. The time interval $[0, T]$ is split into W windows $[T_w, T_{w+1}]$:

$$0 = T_0 < T_1 < ... < T_w < T_{w+1} < ... < T_W = T.$$

In this study, the window size is fixed to S timesteps. Thus $T_w = t_{S.w} = w.S.h$ and $(\delta y_{S.w}^{(i_s)}, \delta z_{S.w}^{(i_s)})^t$ is the solution at time T_w. Knowing $\delta y_n^{(b_s),k}$ and $\delta z_n^{(b_s),k}$ for $n \in \{w.S, ..., (w+1).S\}$, the process s computes on its own $\delta y_n^{(i_s),k+1}$ and $\delta z_n^{(i_s),k+1}$ for $n \in \{w.S, ..., (w+1).S\}$. The iterative scheme is now defined. To initialize the iterative

algorithm, initial boundary values need to be defined. To find coherent values without additional computational cost, the initial guess is fixed at the last known value:

$$\begin{pmatrix} \delta y_n^{(i_s),0} \\ \delta z_n^{(i_s),0} \end{pmatrix} = \begin{pmatrix} \delta y_{w.S}^{(i_s)} \\ \delta z_{w.S}^{(i_s)} \end{pmatrix},\ n \in \{w.S, ..., (w+1).S\}. \tag{8}$$

At the convergence, the solution is rebuilt, as in the sequential solver:

$$\begin{pmatrix} y_{n+1}^{(i_s)} \\ z_{n+1}^{(i_s)} \end{pmatrix} = \begin{pmatrix} \delta y_{n+1}^{(i_s)} \\ \delta z_{n+1}^{(i_s)} \end{pmatrix} + \begin{pmatrix} y_0^{(i_s)} \\ z_0^{(i_s)} \end{pmatrix} \tag{9}$$

The following algorithm details the steps of the Waveform Relaxation solver with a local solver similar to the sequential algorithm on Section II. The set of indexes (i_s, b_q) is the intersection set of the indexes i_s of the sub-domain/sub-matrix s and of the indexes b_q of the sub-domain/sub-matrix q, i.e. $i_s \cap b_q$.

Lower/Upper factorization of matrix $A^{(i_s)}$
while $t_n < T$ i.e. $w < W$ **do**
  Initialise with Eq. (8)
  while WR has not converged, iterate on k **do**
    Recv $\begin{pmatrix} \delta y_{n+1}^{(i_q,b_s),k} \\ \delta z_{n+1}^{(i_q,b_s),k} \end{pmatrix}$,
    and Send $\begin{pmatrix} \delta y_{n+1}^{(i_s,b_q),k+1} \\ \delta z_{n+1}^{(i_s,b_q),k+1} \end{pmatrix}$,
    $\forall n = \{S.w, ..., S.(w+1)\}$
    while $t_n < T_{w+1} = t_{S.(w+1)}$ **do**
      Compute the rhs, Eq. (7)
      Forward/back substitution, Eq. (7)
    end while
  end while
  Update the solution, Eq. (9)
end while

IV. Parareal in Time Method

The Parareal in time method [2] aims to split the time domain and to solve each associated sub-problem independently. The time domain $[0, T]$ is split into time windows $[T_w, T_{w+1}]$:

$$0 = T_0 < T_1 < ... < T_w < T_{w+1} < ... < T_W = T.$$

The objective is to solve the initial problem on each time window, in parallel. To ensure good load balancing, each time window contains the same number S of timestep h, i.e. $T_{w+1} - T_w = \Delta T = S.h$. Obviously, to solve the problem on $[T_w, T_{w+1}] = [t_{S.w}, t_{S.(w+1)}]$, the initial solution at time $T_w = t_{S.w}$ should be known.

Let's define $F_{\Delta T}$ a propagator which allows to compute the solution on the time window $[T_w, T_{w+1}]$, knowing the initial solution at time T_w. This propagator is based on the sequential solver, called the Fine propagator in the following.

Let's define $C_{\Delta T}$ a cheap propagator which allows to compute an approximation of the solution at time T_{w+1} knowing an approximation of the solution at time T_w. This propagator is based on the sequential solver, but with $h = \Delta T$, called the Coarse propagator in the following.

The matrix associated to $F_{\Delta T}$ (resp. $C_{\Delta T}$) is noted A_F (resp. A_C). The coarse propagator is computed sequentially, this allows to have approximation of all $(\delta y_{S.w}, \delta z_{S.w})^t$. These approximate solutions allow the fine solver to compute the solution on each time window in parallel. Then additional computation of the coarse solver, using the values computed by the fine solver, enable to introduce the iteration scheme, with an iteration number k, of the Parareal in time algorithm:

$$\begin{cases} \begin{pmatrix} \delta y_{n+1}^0 \\ \delta z_{n+1}^0 \end{pmatrix} &= C_n^0 \\ \begin{pmatrix} \delta y_{n+1}^{k+1} \\ \delta z_{n+1}^{k+1} \end{pmatrix} &= C_n^{k+1} + F_n^k - C_n^k \end{cases} \tag{10}$$

where

$$C_n^{k+1} = C_{\Delta T}(\begin{pmatrix} \delta y^{k+1} \\ \delta z_n^{k+1} \end{pmatrix}), F_n^{k+1} = F_{\Delta T}(\begin{pmatrix} \delta y_n^k \\ \delta z_n^k \end{pmatrix}).$$

The following algorithm details a parallel implementation of the Parareal in time method. Note that for the first process (called 0), *Recv* is not a reception from another process but just the initialization from the state at t_0. Note also that *Send* of the last process is not used. Each computation of $F_{\Delta T}()$ and $C_{\Delta T}()$ uses forward/back substitution, to solve Eq. (3).

Lower/Upper factorization of A_C
Lower/Upper factorization of A_F
Recv $\begin{pmatrix} \delta y_{S.w}^0 \\ \delta z_{S.w}^0 \end{pmatrix}$ form previous process
Compute $\begin{pmatrix} \delta y_{S.(w+1)}^0 \\ \delta z_{S.(w+1)}^0 \end{pmatrix} = C_{S.w}^0$
Send $\begin{pmatrix} \delta y_{S.(w+1)}^0 \\ \delta z_{S.(w+1)}^0 \end{pmatrix}$ to following process
while Parareal has not converged, iterate on k **do**
  Compute $\begin{pmatrix} \delta \hat{y}_{S.(w+1)}^k \\ \delta \hat{z}_{S.(w+1)}^k \end{pmatrix} = F_{S.w}^k$
  Recv $\begin{pmatrix} \delta y_{S.w}^{k+1} \\ \delta z_{S.w}^{k+1} \end{pmatrix}$ form previous process
  Compute $\begin{pmatrix} \delta \tilde{y}_{S.(w+1)}^{k+1} \\ \delta \tilde{z}_{S.(w+1)}^{k+1} \end{pmatrix} = C_{S.w}^{k+1}$
  Update $\begin{pmatrix} \delta y_{S.(w+1)}^{k+1} \\ \delta z_{S.(w+1)}^{k+1} \end{pmatrix}$ with Eq. (10)
  Send $\begin{pmatrix} \delta y_{S.(w+1)}^k \\ \delta z_{S.(w+1)}^k \end{pmatrix}$ to following process
end while
Update the solution, Eq. (4)

V. Coupling Parareal and Waveform Relaxation Methods

The Waveform Relaxation method and the Parareal in time method have a limited speedup. Indeed, on one hand, for the Waveform Relaxation method, the solution cost of the local sub-domains should be high and the communications between the sub-domains cost should be small. It is thus important to have a large number of sub-domains, in such a way that the number of variables within each sub-domain is large enough. In other words, this means that to split the initial domain into sub-domains with few variables within each sub-domain leads to a lower efficiency. On the other hand, the Parareal in time method has a speedup bounded by the number of processes divided by two. Coupling these two methods [3] allows to combine the good speedup of the Waveform Relaxation, with the maximum possible number of sub-domains i.e. before its speedup decreases, and to continue to increase its speedup by using the Parareal in time method.

The coupled Parareal-Waveform Relaxation method is based on the Parareal in time method. The Fine propagator $F_{\Delta T}$ is based on the Waveform Relaxation method. With the same time windows than the Parareal in time algorithm $[T_w, T_{w+1}]$, and with the same timestep h than the sequential method. The Coarse propagator $C_{\Delta T}$ is based on the same method, and with both the timestep and the window size equal to the Parareal time windows: $h = T_{w+1} - T_w$. There is no reason for the Fine and the Coarse propagators to converge at each iteration of the Parareal method. They only need to converge at the last iteration of the Parareal method. This is the reason why the number of iterations of the propagators is bounded to a small number, 2 or 3. These new propagators with limited number of iterations are denoted $\tilde{F}_{\Delta T}$ and $\tilde{C}_{\Delta T}$.

The working variables in each Waveform Relaxation are written as follows: $u_{n,k}^{(i_s),l}$ with s the sub-domain number, n the index of the timestep, k and l the iteration numbers of the Parareal and Waveform Relaxation propagator respectively. The Parareal solutions are written as follows: $u_{n,k}^{(i_s)}$. The difference with the Parareal Algorithm (Section IV) is that the communication *Send/Recv* are done from the previous and to the following processes group or processes communicator. Each communicator computes the Fine and Coarse propagator in parallel following the algorithm described on Section III. The Fine Waveform Relaxation initial guess of k^{th} coupled Parareal-Waveform Relaxation is not set to a constant initialization but to an improved approximation which does not require additional computational cost, $n \in \{w.S, ..., (w+1).S\}$:

$$\begin{pmatrix} \delta y_{n,k+1}^{(i_s),0} \\ \delta z_{n,k+1}^{(i_s),0} \end{pmatrix} = \begin{pmatrix} \delta y_{w.S,k+1}^{(i_s)} \\ \delta z_{w.S,k+1}^{(i_s)} \end{pmatrix} + \tilde{F}_{n,k}^{(i_s)} - \begin{pmatrix} \delta y_{w.S,k}^{(i_s)} \\ \delta z_{w.S,k}^{(i_s)} \end{pmatrix}. \tag{11}$$

The following algorithm shows a parallel implementation of the coupled Parareal-Waveform Relaxation method using the Waveform Relaxation algorithm introduced in Section III and initialized by Eq.(11).

Lower/Upper factorization of A_C
Lower/Upper factorization of A_F
Recv $\begin{pmatrix} \delta y_{S.w}^{0} \\ \delta z_{S.w}^{0} \end{pmatrix}$ form previous process
Compute $\begin{pmatrix} \delta y_{S.(w+1)}^{0} \\ \delta z_{S.(w+1)}^{0} \end{pmatrix} = \tilde{C}_{S.w}^{0}$ by WR
Send $\begin{pmatrix} \delta y_{S.(w+1)}^{0} \\ \delta z_{S.(w+1)}^{0} \end{pmatrix}$ to following process
while Parareal has not converged, iterate on k **do**
  Compute $\begin{pmatrix} \delta \hat{y}_{S.(w+1)}^{k} \\ \delta \hat{z}_{S.(w+1)}^{k} \end{pmatrix} = \tilde{F}_{S.w}^{k}$ by WR
  Recv $\begin{pmatrix} \delta y_{S.w}^{k+1} \\ \delta z_{S.w}^{k+1} \end{pmatrix}$ form previous process
  Compute $\begin{pmatrix} \delta \tilde{y}_{S.(w+1)}^{k+1} \\ \delta \tilde{z}_{S.(w+1)}^{k+1} \end{pmatrix} = \tilde{C}_{S.w}^{k+1}$ by WR
  Update $\begin{pmatrix} \delta y_{S.(w+1)}^{k+1} \\ \delta z_{S.(w+1)}^{k+1} \end{pmatrix}$ with Eq. (10)
  Send $\begin{pmatrix} \delta y_{S.(w+1)}^{k} \\ \delta z_{S.(w+1)}^{k} \end{pmatrix}$ to following process
end while
Update the solution, Eq. (4)

VI. Numerical Results

The problem considered in this section consists to solve a DAE system composed of 100.000 differential and algebraic variables, during a physical simulation time equal to 6.4 seconds. This time corresponds to the time between two events, i.e. where the linearization of the functions f and g is valid. Despite the space dimension of the DAE system might looks small, the reader should keep in mind that after the time discretization, this problem is impossible to solve in a near real time simulation without parallel computing.

To solve this problem, the system (in space) is split into 2, 4, 8, 16, 32, and 64 sub-systems with a mesh partitioning software [12]–[14]. In this paper, the METIS software has been used. The Waveform Relaxation method is applied on these sub-systems. Each sub-system is allocated to one different processor. As illustrated in Figure 1, the speedup of the Waveform Relaxation method increases till 16 sub-systems. After reaching a peak, considering more sub-systems does not increase the speedup, but decrease it. This is coherent with the analysis performed in the previous sections, since when more than 16 sub-systems are considered on this problem (100.000 variables), each sub-system contains too few variables to ensure a good efficiency of the method. In order to improve the speed-up, the coupled Parareal-Waveform Relaxation method is

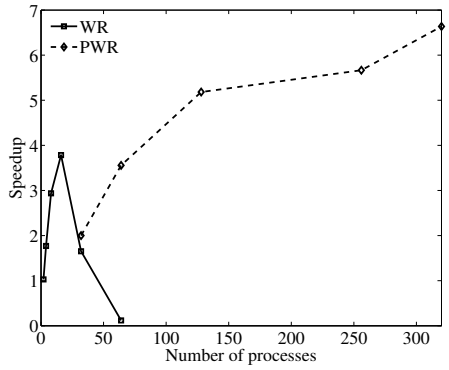


Figure 1. Comparison of the speedup of the Waveform Relaxation method and the coupled Parareal-Waveform Relaxation method.

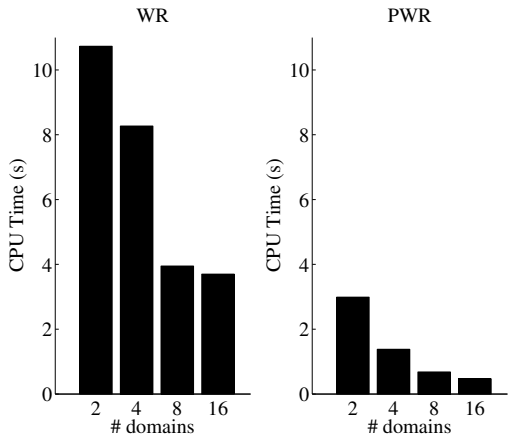


Figure 2. Comparison of the computational time of the Waveform Relaxation method and the coupled Parareal-Waveform Relaxation method.

now applied. For this purpose, the system (in space) is split into 16 sub-systems (in space) and into several windows (in time), forming a space-time grid. Each windows always contains 10 timesteps. Each sub-system, i.e. a cell of this space-time grid, is allocated to one different processor. The coupled Parareal-Waveform Relaxation method illustrated in Figure 1 shows that the speedup continues to increase when considering more processors.

The results reported Figure 2 illustrate the respective computational time of the Waveform Relaxation method and the coupled Parareal-Waveform Relaxation method upon the number of sub-systems (in space) when the system is split into 32 windows (in time). This figure clearly shows the extraordinary reduction of the computational time when using the coupled Parareal-Waveform Relaxation method.

VII. Conclusion

In this paper a new coupled Parareal-Waveform Relaxation method has been presented for solving Differential Algebraic Equations. This method outperforms the existing Waveform Relaxation method and the Parareal in time methods, by taking benefits of these two methods. The numerical results obtained illustrate the impressive efficiency of this coupled Parareal-Waveform Relaxation method, which allows to take full advantage of parallel multicores architectures.

References

[1] J. White and A. Sangiovanni-Vincentelli, *Relaxation Techniques for the Simulation of VLSI Circuits*. Kluwer Academic Publishers, 1987.

[2] J.-L. Lions, Y. Maday, and G. Turinici, "Résolution d'EDP par un schéma en temps pararéel," *Comptes Rendus de l'Académie des Sciences - Série I - Mathematics*, vol. 332, no. 7, pp. 661–668, 2001.

[3] R. Gueta and Y. Maday, "Coupling the Parareal algorithm with overlapping and nonoverlapping domain decomposition methods," ser. Lecture Notes in Computational Science and Engineering. Springer, (in press).

[4] B. Smith, P. Bjorstad, and W. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*. Cambridge University Press, UK, 1996.

[5] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*. Oxford University Press, Oxford, UK, 1999.

[6] F. Magoulès and F.-X. Roux, "Lagrangian formulation of domain decomposition methods: a unified theory," *Applied Mathematical Modelling*, vol. 30, no. 7, pp. 593–615, 2006.

[7] A. Toselli and O. Widlund, *Domain Decomposition methods: Algorithms and Theory*. Springer, 2005.

[8] Y. Maday and F. Magoulès, "Absorbing interface conditions for domain decomposition methods: a general presentation," *Computer Methods in Applied Mechanics and Engineering*, vol. 195, no. 29–32, pp. 3880–3900, 2006.

[9] P. Collino, G. Delbue, P. Joly, and A. Piacentini, "A new interface condition in the non-overlapping domain decomposition," *Computer Methods in Applied Mechanics and Engineering*, vol. 148, pp. 195–207, 1997.

[10] Y. Maday and F. Magoulès, "Optimal convergence properties of the FETI domain decomposition method," *International Journal for Numerical Methods in Fluids*, vol. 55, no. 1, pp. 1–14, 2007.

[11] M. Gander, L. Halpern, F. Magoulès, and F.-X. Roux, "Analysis of patch substructuring methods," *International Journal of Applied Mathematics and Computer Science*, vol. 17, no. 3, pp. 395–402, 2007.

[12] G. Karypis and V. Kumar, *METIS: A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices Version 4.0*, 1998.

[13] C. Walshaw, *The serial JOSTLE librairy user guide : Version 3.0*, 2002.

[14] F. Pellegrini, *Scotch and LibScotch 5.1 User's Guide*, 2008.

The Research Of Feature Selection Of Text Classification Based On Integrated Learning Algorithm

Xia Huosong
Department of College of
Economics and Management
Wuhan Textile University
Wuhan, Hubei , China
e-mail:xiahuosong8@gmail.com

Liu Jian
Department of College of
Economics and Management
Wuhan Textile University
Wuhan, Hubei, China
e-mail:liujian0906018@163..com

***Abstract*—Feature selection is a very important step in text classification , which affects the accuracy and validity in text classification. Base on four classic feature selection algorithms of text mining , this paper has established a kind of integrated learning algorithm which applies with feature selection in text classification ,in order to improve the accuracy of text classification. Test results show that the algorithm improves the accuracy of text classification.**

Keywords- Text Classification ; The Weights Of Feature ; Integrated Learning Algorithm(ILA); Feature selection.

I. INTRODUCTION

Text classification is the problem of automatically assigning predefined categories to free text documents[1]. While computer con not directly identify natural language, which need to be converted to the forms that computer can identify, and at present the widespread use of text representation model is vector space model(VSM). In VSM, a major characteristic , or difficulty, of text classification problems is the high dimensionality of the feature space. The native feature space consists of the unique terms(words or phrases) that occur in documents , which can be tens or hundreds of thousands of terms for even moderate-sized text collection. This is prohibitively high for many learning algorithms, and the problem greatly reduces the accuracy and validity of text classification. The contribution of each feature term for text classification is different , so according to the contribution of each feature term , how to use a efficient algorithm to selection the most important feature terms in terms of contribution is very crucial . Several common feature selection algorithms are document frequency(DF)[2], information gain(IG)[3], mutual information (MI)[4], A x2 statistic(CHI)[5,6], and term strength(TS) and so on. In paper[4], DF, IG and CHI achieved a higher accuracy. DF is the most simple feature selection algorithm, and is applicable to massive documents classification. DF considers the contribution of the high-frequency terms, but it ignore the contribution of rare-frequency terms which maybe very important; the advantage of IG is that it consider the contribution of the terms that do not appear in the text but maybe very important for the text classification.; discerning the low frequency terms , CHI is not so accurate , but overall it is the most accurate feature selection algorithm; the great disadvantage of MI is that it does not consider the frequency of the terms that appear in the text, which leads the evaluation function of MI to incline to choose rare-frequency terms. In view of the advantages and disadvantages of the above algorithms ,a large number of scholars focus the improvement of the algorithms. In paper[7], integrating the model of statistical machine learning and integrated learning algorithm , accomplishes automatic selection for feature terms in text classification. In paper[8] , a method of feature selection based on fuzzy relation between terms and classes is proposed . In paper[6], through introducing and comparing eight feature selection algorithm in text classification , a new method of feature selection based on Class-Discriminating Words(CDW) is proposed. The algorithms above have different features, but they only focus the certain aspect in feature selection, not a comprehensive consideration . In this paper , through the research of the four common feature selection algorithms- DF , GI , MI , CHI -, making full use of advantages of the four classic algorithms, an integrated learning algorithm is proposed , in order to improve the accuracy of text classification.

II. DESCRIPTION OF FOUR COMMON ALGORITHM

A. DF Algorithm

Document frequency is the number of documents in which a term occurs. We computed the document frequency for each unique term in the training corpus and removed from the feature space those terms whose document frequency was less than some predetermined threshold . The basic assumption is that rare terms are either non-informative for category prediction, or not influence in global performance. In either case removal of rare terms reduces the dimensionality of the feature space. Improvement in categorization accuracy is also possible, if rare terms happen to be noise terms .

DF is the simplest technique for vocabulary reduction. It easily scales to very large corpora, with a computational complexity approximately linear in the number of training documents . However , it is usually considered an ad hoc approach to improve efficiency, not a principled criterion for selection predictive features. Also, DF is typically not used for aggressive term removal because of a widely received assumption in information retrieval. That is low-DF terms are assumed to be relatively informative and therefore should not be removed aggressively. We will re-examine this assumption with respect to text categorization tasks.

B. IG Algorithm

Information gain is frequently employed as a term-goodness criterion in the field of machine learning. It measures the number of bits of information obtained for

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.95

category prediction by knowing the presence or αβσενχε of a term in a document. The information gain of term t is defined to be :

$$IG(t) = -\sum_{i=1}^{m} p(c_i)\log P(c_i) + P(t)\sum_{i=1}^{m} P(c_i|t)\log P(c_i|t) + P(t^-)\sum_{i=1}^{m} P(c_i|t^-)\log P(ci|t^-) \quad (1)$$

Given a training corpus , for each unique term we computed the information gain , and removed the feature space those information gain was lea than some predetermined threshold. The computation includes the estimation of the conditional probabilities of a category given a term , and the entropy computation in the definition.

C. MI Algorithm

Mutual information is criterion commonly used in statistical language modeling of word associations and related applications. If one considers the two-way contingency table of a term t and a category c , where A is the number of times t and c co-occur, B is the number of time the t occur without c ,C is number of documents, then the mutual information criterion between t and c is defined to be :

$$MI(t) = \sum_{i=1}^{m} P(c_i)\log\frac{P(t,c_i)}{P(t)P(c_i)} \quad (2)$$

A weakness of mutual information is that the score is strongly influenced by the marginal probabilities of terms. For terms with an equal conditional probability, rare terms will have a higher score than common terms. The scores, therefore, are not comparable across terms of widely differing frequency.

D. CHI

The x2 statistic measures the lack of independence between t and c and can be compared to the x2 distribution with one degree of freedom to judge extremeness. Using the two-way contingency table of a term t and a category c ,where A is the number of times t and c co-occur . B is the number of times the t occurs without c. C is the number of times c occur without t. D is the number of times neither c nor t occurs , and N is the total number of documents, the terms-goodness measure is defined to be :

$$X^2(t,c) = \frac{N\times(AD-CB)^2}{(A+C)\times(B+C)\times(A+B)\times(C+D)} \quad (3)$$

The x2 statistic has a natural value of zero if t and c are independent. We computed for each category the x2 statistic between each unique term in a training corpus and that category , and then combined the category-specific scores of each term in two scores. The computation of CHI scores has a quadratic complexity , similar to MI and IG. A major difference between CHI and MI is that x2 is a normalized value ,a and hence x2 values are comparable across terms for the same category. However , this normalization breaks down (can no longer be accurately compared to the x2 distribution), which is the case for low frequency terms. Hence , the x2 statistic is known not to be reliable for low-frequency terms.

III. DESCRIPTION OF INTEGRATED LEARNING ALGORITHM

The integrated learning algorithm is defined to be:

$$Q(t_i) = aQDF(t_i) + bQMI(t_i) + cQIG(t_i) + dQCHI(t_i) \quad (4)$$

Q(ti) is the score of term i in document t. The accuracy degree of DF is expresses as a; The accuracy degree of MI is expresses as b; The accuracy degree of GI is expresses as c; The accuracy degree of CHI is expresses as d. And QDF indicates the score of term i in document t through DF; QMI indicates the score of term i in document t through MI; QGI indicates the score of term i in document t through GI; QCHI indicates the score of term i in document t through CHI .

IV. DATA TEST AND THE ANALYSIS OF THE RESULTS

A. Data Set

The experimental corpus in this paper adopts on –line Tan Songo balance corpus[9]. The general statistical features are as follow: the scale of the corpus is 4000 ; among the corpus , passive corpus accounts 50%(2000), and the positive corpus accounts 50%(2000). Removal of markings and stop words, this paper uses the open-source code of ICTCLAS which is developed by the Chinese Academy of Sciences Institute of Computing Technology. The texts which are processed constitute the text feature vectors.

B. Classifier

To assess the effectiveness of the feature selection algorithm, in this paper , we select SVM(Support Vector Machine) to construct classifier. All SVM classifiers adopt RBF kernel function , through the cross-validation way to find the best c, g in order to construct the optimal SVM classifier. We use the SVM form professor Lin Zhiren who teaches in Tai Wan National University[10]. The training set accounts 50%; the test set accounts 50%.

C. Asserssment Criteria

We apply feature selection to documents in the processing of SVM. The effectiveness of a feature selection method is evaluated using the performance of SVM on the preprocessed documents. We use the standard definition of recall and precision as performance measures:

Recall=categories found and correct/total categories correct

Precision=categories found and correct/total categories found

F1= categories found and correct/ total categories correct+ total categories found

So we define the six measure indexes:

Positive recall (PR) ; positive precision (PP);

Passive recall (NPR); passive precision (NPP);

Positive and passive F1(PNF1).

D. Results and the Analysis

In the experiment , we use MATLAB to accomplish the documents preprocessing and the feature selection algorithm,

and 4000 –dimension al features are selected , and then compute the weight of the feature items, and establish Vector Space Model of the features .
The results are as follows (table I):

TABLE I THE COMPARISON OF THE RESULTS OF TEXT CLASSIFICATION THOUGH THE FOUR COMMON METHODS AND INTEGRRATED LEARNING ALGORITHM

Algorit -hms	Measure Indexes				
	PR	***PP***	***NPR***	***NPP***	***NPF1***
DF	70.80%	74.54%	78.14%	72.60%	74.44%
MI	76.42%	73.53%	79.83%	77.17%	76.47%
IG	78.88%	77.33%	78.82%	79.87%	78.75%
CHI	81.63%	79.35%	82.70%	80.95%	81.47%
ILA	84.17%	81.39%	84.12%	80.11%	83.11%

From the table we can see as follows:

(1) Through integrated learning algorithm ,the accuracy of text classification is improved , but the degree of improvement is shallow;
(2) The gradient between the recall and precision of the integrated learning algorithm is obvious .

V. CONCLUSION

This is an evaluation of feature selection methods in dimensionality reduction for text categorization at all the reduction levels of aggressiveness, form using the full vocabulary(except stop words) as the feature space, to removing 98%of the unique terms. We found the integrated learning algorithm and CHI most effective in aggressive term removal without losing categorization accuracy in our experiments with SVM. But through integrated learning algorithm ,the accuracy of text classification is improved , but the degree of improvement is shallow, and the gradient between the recall and precision of the integrated learning algorithm is a little big . So ,it is the new direction that we should proceed our research to construct a more perfect feature selection algorithm . And the availability of a simple but effective means for aggressive feature space reduction may significantly ease the application of more powerful and computationally intensive learning methods, such as neural network , to very large text categorization problems which are otherwise intractable.

REFERENCES

[1] C. Apte , F , Damerau and S . Weiss. Text mining with decision rules and decision trees[C].In Proceeding of the Conference on Automated Learning and Discovery Workshop 6: Learning form Text and the web, Carnegie-Mellon University, June, 1998.

[2] Dumais S T , Platt J , Heckerman D , et al. Inductive learning algorithms and representations for text categorization. Technical Report , Microsoft Research, 1998.

[3] Shang Wenqian , Huang Houkuan, Zhu Haibin, et al. A novel feature selection algorithm for text categorization[J]. Expert Systems with Application, 2007, 33:1-5.

[4] Y Yang, J. Q Pedersen. Feature selection in statistical learning of text categorization. In the 14th Int Conf on Machine Learning, 1997:412-420.

[5] Chen Tao, Xie Yangqun. Literature Review of Feature Dimension Reduction in Text Categorization[J]. Journal of The China Society For Scientific and Technical Information, 2005, 24(6):690-695.

[6] Zhou Qian, Zhao Mingsheng, Hu Min. Study of Feature Selection in Chinese Text Categorization[J]. Journal of Chinese Information Processing, 2004, 18(3):17-23.

[7] Zhang Chengzhi. Research of Automatic Indexing Based on Integrated Learning algorithm[J]. Journal of The China Society For scientific and Technical Information,2010-02 29(1):3-8.

[8] Zhen Zhilong, Han Lixin, Lu Dianlong. Feature selection based on fuzzy relation for text categorization[J]. Journal of the china society for scientific and technical information, 2008-12: 851 - 856.

[9] Tan Songbo. Chinese emotional mining corpora. http://www.searchforum.org.cn/tansongbo/corpus-senti.htm

[10] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001.Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Acknowledgement

This research is supported by HuBei Province Fund of China (2010064, Z20091701, 2008d062, 2008244, 2007097, 2010DEA025）), Wuhan Science and Technology Bureau(200940833384-02),WTU Fund (ZD20100202), CNTAC fund(2007082). We deeply appreciate the suggestions from fellow members of Xia's project team and Research center of Enterprise Decision Support,Key Research Institute of Humanities and Social Sciences in Universities of Hu Bei Province.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

A QoS-based Routing Algorithm in Ad Hoc Network

Zheng Sihai, Li Layuan, Li Yong
College of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
E-mail: sihaizheng@sina.com

***Abstract*—With the development of Ad Hoc Network, the demand of the users becomes more and more than ever before, some business requires the ability not only to communicate, but also to guarantee delay and bandwidth. This paper puts forward a QoS-based multipath routing protocol QMPSR. It takes bandwidth and delay constraint into account, it can find several paths to provide QoS guarantee. Simulation shows that QMPSR protocol is better than QoS-MSR. Its packet delivery ratio is higher, average delay is less; even the increased routing overhead caused is not too much.**

Keywords-Ad Hoc Network; routing protocol; average delay; QoS

I. INTRODUCTION

Ad Hoc Network is a set of wireless mobile nodes which communicates with each other using multi-hop wireless links, and there are no any existing network infrastructure or centralized administration. Each node moves arbitrarily, thus network topology changes frequently and unpredictably. Moreover, bandwidth, battery life and physical security are all limited in actual situations. Because of these constraints, it is much difficult to design routing protocol in Ad Hoc Network. With the development of Ad Hoc Network, the demand of the user becomes more and more than ever before, some business requires the ability not only to communicate, but also to guarantee delay and bandwidth, etc. Therefore, it is more and more important to provide QoS guarantee for multimedia application.

At present, some papers on multipath routing show that providing more than one path can better support high mobility network with low end-to-end delay, high packet delivery ratio and lower control overhead. Ref. [1] proposes a QoS metric called end-to-end reliability with multipath discovery (MP-DSR), but it is not defined for real-time traffic with QoS constraints. Ref. [2] proposes an interfering-aware QoS multipath routing for QoS multimedia applications in Ad Hoc Network (IMRP). Though it can compute disjoint paths based on bandwidth availability and the link stability computed by the nodes, it has not mentioned about the method used for estimation of bandwidth and stability values. Ref. [3] proposes a distributed multi-path dynamic source routing protocol (QoS-MSR) to improve QoS support with respect to end-to-end reliability. In a sort of ideal model, but it doesn't consider the congestion of nodes to QoS routing.

In the rest of this paper, we first discuss the notations and assumptions in section Ⅱ. Then, research on QoS model in section Ⅲ. QMPSR routing algorithm is analyzed in section Ⅳ. The results of simulation are shown in Section Ⅴ. At last, we offer conclusions in Section Ⅵ.

II. NOTATIONS AND ASSUMPTIONS

Definition 1: Bandwidth B is data transmission rate; its unit is bits/sec.The average transmission required by multimedia files is as follow [4]:

$$B_f = F \times R \times C \quad (1)$$

Where F is the number of frames which have been played in per unit time (frame/sec), R is image resolution; C is the rate codes are compressed.

Definition 2: End to end delay T is the time in which the packets are sent from the source node to destination, its unit is second.

$$T = \sum (T_t(I) + T_p(I)) + T_{prop}(S,D) \quad (2)$$

Where I is the members nodes in the selected routing, S is source node, D is destination node, T_p is the processing delay, T_{prop} is the propagation delay in wireless channel, they further can be expressed as follow [5] :

$$T_t(I) = S_{pkt}/B_a(I), \quad (3)$$

$$T_p(I) = T_w(I) + T_d(I). \quad (4)$$

Where S_{pkt} is the size of packet, $B_a(I)$ is actual effective bandwidth can be used by node I, $T_p(I)$ is the packet processing time at node I, it also includes $T_w(I)$ and $T_d(I)$. $T_w(I)$ is the waiting time of node I in the queue, $T_d(I)$ is the real time data packets are processed.

Definition 3: J is delay jitter, it represents the size of the delay variation, its value is equal to the variance of delay T.

Definition 4: Packet loss rate η is expressed as follow:

$$\eta = 1 - P_r/P_s \quad (5)$$

Where P_r is the total packets the destination nodes have received, P_s is the total packets the source nodes have sent. Packet loss rate is primarily determined by the state of the buffer. There will loss some packets when data exceed the capacity of the buffer.

III. QOS MODEL

QoS routing is an NP-complete problems, it will lead to significant computation if we want to meet these restrictions simultaneously. This obviously does not pay in practice. This article hopes to get a simplified QoS model, which can not

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.13

only achieve a certain QoS guarantee, but also easy to implement the agreement.

A. QoS Service Model

The QoS based model proposed in the current can be divided into two categories [6]: QoS guarantees based on data flow and hierarchical service. The QoS guarantees based on data flow is evolved by the intServ model, its main idea is to reserve resources, and it should provide QoS guarantees required by each data flow which are admitted. The hierarchical service is improved from DiffServ model, the main idea is to provide a high level of QoS guarantee for data flow if the network resources are sufficient, otherwise, only do best effort.

SWAN is a typical hierarchical service model [7]. In this system, UDP is treated as real-time business, so it can obtain a higher level of QoS guarantee, and TCP receives only best-effort service. When data flow ask to access the network, source nodes first initiate a network resource query, if the network resources meet the transmission requirements, data flow will be normal access, and otherwise it will be marked as "best effort". Those best effort data should be transmit through the traffic shaper, which make its rate under control to adapt to the current state of network load received the feedback from MAC layer. The inadequacies of this approach are much dependent on the feedback from MAC layer. In addition, those data flow which access to network later can not obtain a high level of QoS guarantee because the resources have been occupied.

B. Theorem and Proof

Theorem 1: When the function F_l is additive, Multiplicative or minimized, the model is a linear programming problem.

Proof: Linear programming (LP) is a mathematical method for determining a way to achieve the best outcome (such as maximum profit or lowest cost) in a given mathematical model for some list of requirements represented as linear relationships.

More formally, linear programming is a technique for the optimization of a linear objective function, subject to linear equality and linear inequality constraints. Given a polytope and a real-valued affine function defined on this polytope, a linear programming method will find a point on the polytope where this function has the smallest (or largest) value if such point exists, by searching through the polytope vertices. Linear programs are problems that can be expressed in canonical form:

$$\text{maximize } c^T x$$
$$\text{subject to } Ax \leqslant b$$

Where x represents the vector of variables (to be determined); c and b are vectors of (known) coefficients and A is a (known) matrix of coefficients. The expression to be maximized or minimized is called the objective function ($c^T x$ in this case). The equations $Ax \leqslant b$ are the constraints which specify a convex polytope over which the objective function is to be optimized.

According to lemma 1 and lemma 2, the function F_l is additive, Multiplicative or minimized, the model is a linear programming problem, and if $k \geqslant 2$, it is an NP-complete problem.

Theorem 2: The packet delivery ratio of QMPSR is higher than QoS-MSR, and the time delay of QMPSR is least.

Proof: Some researchers have proposed some algorithms to optimize multi-constrained QoS routing. These algorithms can be divided into heuristic algorithms, approximation algorithms and multi-constrained QoS routing algorithm based on scheduling strategy. Heuristic algorithm can reduce the time complexity, but it can not guarantee to find a feasible transmission path even if there are. Approximation algorithm can find the approximate solution of the optimal path, however, the time complexity of these algorithms are often higher than the heuristic algorithm. Approximation algorithm can also be divided into pseudo-polynomial approximation algorithms and polynomial approximation algorithm. The difference is the computational complexity of pseudo-polynomial depends on not only the size of the network, but also the size of the network link parameters. The multi-constrained QoS routing algorithm based on scheduling strategy can solve multiple constrained QoS routing problem, however, it requires a particular scheduling strategy, which makes these algorithms have some limitations in reality.

QoS routing is an NP-complete problems, it will lead to significant computation if we want to meet these restrictions simultaneously. This obviously does not pay in practice. QoS-MSR can improve QoS support with respect to end-to-end reliability in a sort of ideal model, but it doesn't consider the congestion of nodes to QoS routing. QMPSR uses a simplified QoS model, which can not only achieve a certain QoS guarantee, but also easy to implement the agreement. So its performance is better. □

Theorem 3: When the part of links of G is unidirectional, the complexity of QMPSR is o (n^2).

Proof: We model the network as a graph $G = (N, L)$, where N is the set of vertices and L is the set of edges. Some of the edges are assumed to be directed. Every vertex (also referred to as a node) is reachable from every other vertex. Thus, every node in the network can send packets to every other node in the network. A node, on receiving a packet from some other node, can determine the length of the path taken by that packet. Let each packet start from the source x with its Time To Live (TTL) field initialized to TTL-max. All nodes have agreed a priority on the value of TTL-max. Each intermediate node z and the destination y on receiving the packet decrements the TTL field by one. Let us refer to the resultant value as TTL-receive. When the packet arrives at the destination node the length of the path traversed by the packet thus far is equal toTTL-max - TTL-recv.

Let path(ab) be the shortest path from node a to node b. As some links are unidirectional, path(ab) may be different from path(ba).

Let path($av_1 v_2 \ldots v_k$b) be the shortest path from a to b that passes through vertices $v_i : 1 \leq i \leq k$ such that v_i precedes v_j if $i < j$.

Let length (path (x)) be the number of wireless links in path (x), where x is a sequence of vertices.

Let directed path (*ab*) be the path (*ab*), which is said to be a directed path if it has at least one directed link.

We assume that path(*cdef*) is the shortest path from *c* to *f*. Let:

$X = \{x: x$ is a node on path $(cd)\}$, and

$Y = \{y: y$ is a node on path $(ef)\}$

As path (*cdef*) is the shortest path from *c* to *f*, for all *x* and *y*, path(*xy*) goes through vertices *d* and *e*. Therefore, every node *p* on path (*yfcx*) has to propagate length(path(*xy*)); $\forall x \in X$; $\forall y \in Y$. As sets *X* and *Y* can be as large as *N*, $|X| = O(n)$ and $|Y| = O(n)$, where $n=|N|$. Then the complexity of QMPSR is $o(n^2)$. □

IV. THE PROCESS OF ROUTING

A. Routing Discovery

The routing discovery process is shown in Fig. 1.

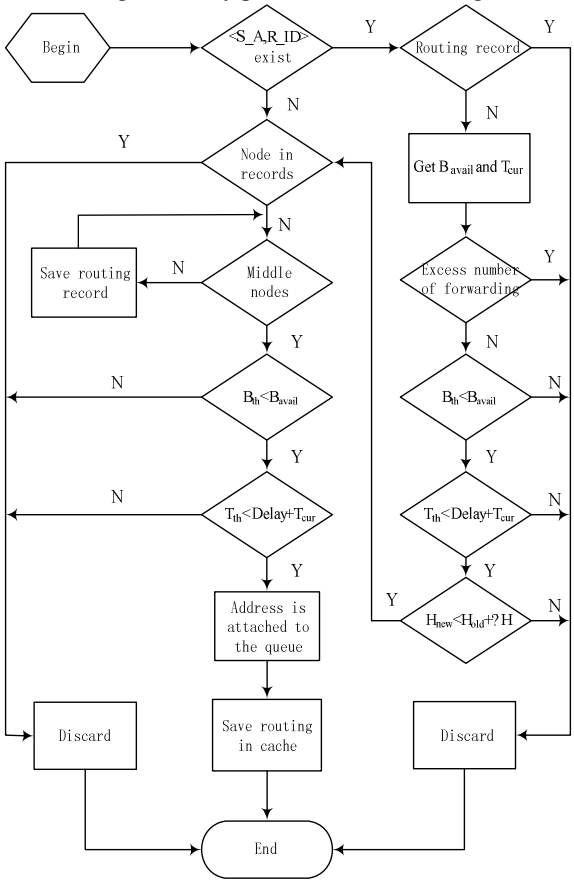


Fig.1. Routing Discovery Process

B. Load Distribution

If the weight of load distribution in *k*-path is W_k, then,

$$W_k \propto B_k \quad (k=1,2,3,\cdots,n) \tag{6}$$

B_k is the available bandwidth in *k*-path. In addition,

$$B_U = B_1 + B_2 + \cdots + B_n \tag{7}$$

$$W_k = B_k / B_U \tag{8}$$

Load distribution algorithm is as follow,

$$W_k = \min([B_k / B_U], M) \times R \tag{9}$$

M is the maximum value of W_k, Its role is to limit W_k, if W_k is over the range, multipath would degenerate into single path. *R* is used to the granularity of load distribution on each path. When the load is distributed to the multipaths, *Round Robin Scheduling Algorithm* [8-9] is used.

V. SIMULATION ANALYSIS

In order to fully examine the performance of QMPSR, we had conducted simulation experiment. The scene is 1000m × 1000m; there are 50 nodes, they are also randomly distributed. The connections are ten pairs. Nodes can randomly move. Its maximum speed is 15m/s; Data source and packet size are the same as the first set, but the speed of sending packets is 300.0.

Fig.2 shows packet delivery ratio of QMPSR is about 5% higher than that of QoS-MSR, but that is about 15% higher in first set. There are more nodes and connections in the second set, the communication between nodes is very frequent, it is more difficult to choose multipath, therefore, the difference is not much obvious.

Fig.3 shows the delay of QMPSR is less than that of QoS-MSR. It indicates QMPSR can speed up the delivery of data and reduce the delay. However, the delay increases too much in the second set. The hops of communication links are more; therefore, it has a corresponding increase in average delay. But the advantage of QMPSR is still very clear.

Fig.4 shows the routing overhead of QMPSR is less than that of QoS-MSR. But overhead of the two protocols are more than that of the first set. It indicates the performance have declined with the expansion of the network.

VI. CONCLUSION

This paper focuses on a new QoS multipath routing protocol QMPSR. It can find several paths provided QoS guarantee. Also, it can disassemble QoS demands to transmit application data stream using different node disjointed path simultaneously. Simulation shows that QMPSR protocol is better than QoS-MSR. Author will focus on how to load balance more effectively in future work, which can solve the problem of out of order caused by multipath.

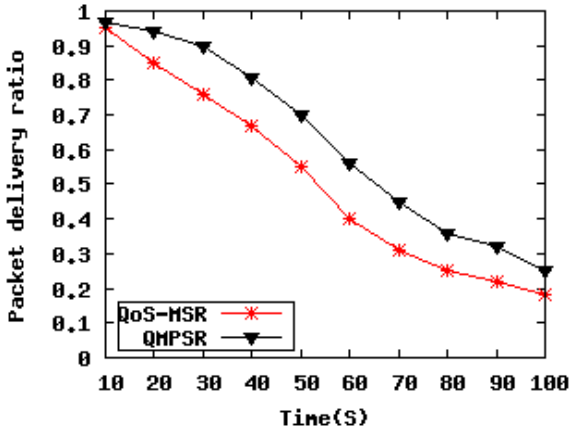


Fig.2 Average packet delivery ratio of 50 nodes

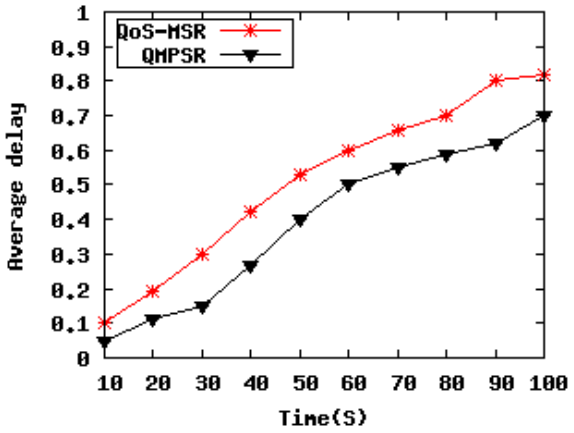


Fig.3 Average end to end delay of 50 nodes

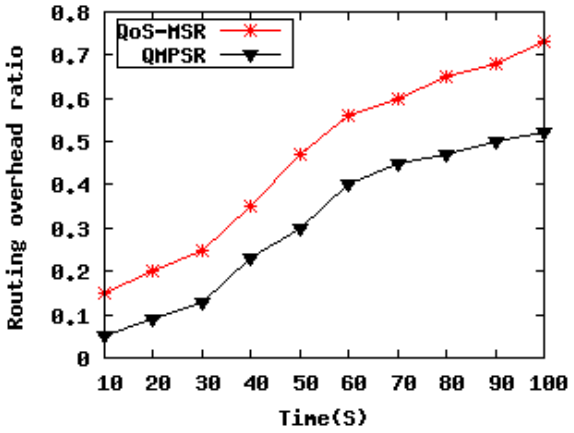


Fig.4 Routing overhead of 50 nodes

ACKNOWLEDGMENT

This paper is supported by National Natural Science Foundation of China (No: 60672137, 60773211, 60970064), Open Fund of the State Key Laboratory of Software Development Environment (No: SKLSDE-2009KF-2-02), New Century Excellent Talents in university (No: NCET-08-0806), Specialized Research Fund for the Doctoral Program of Higher Education of China (No：20060497105).

REFERENCES

[1] Wu Huayi; Jia Xiaohua. QoS multicast routing by using multiple paths/trees in wireless ad hoc networks. Ad Hoc Networks, 2007,5(5): 600~612.

[2] Mateen Wajahat; Raza Saqib. Adaptive multi-path on-demand routing in mobile ad hoc networks. Proceedings of ISORC, 2005:237~244.

[3] Mike Burmester, Tri van Le and Alec Yasinsac. Weathering the storm: managing redundancy and security in ad hoc networks. Proceedings of the 3rd International Conference on Ad Hoc Networks & Wireless, Vancouver, British Columbia, 2004,7: 96-107.

[4] Ivan Stojmenovic , Mahtab Seddigh , Jovisa Zunic. Dominating Sets and Neighbor Elimination-Based Broadcast Algorithms in Wireless Networks. IEEE Transactions on Parallel and Distributed Systems, 2002, 1 (13): 14-25.

[5] Jie Wu, Fei Dai. A Generic Distributed Broadcast Scheme in Ad Hoc Wireless Networks. IEEE TRANSACTIONS ON COMPUTERS. 2004, 10 (53): 1343-1354.

[6] Li Layuan, Li Chunlin. A QoS multicast routing protocol for dynamic group topology, Information Science. 2004, 169(1/2):113-130.

[7] Li Layuan, Li Chunlin. A distributed QoS-Aware multicast routing protocol, Acta Informatica, Spring-Verlag Heidelberg, Vol 40/3, pp.211-233, November 2003.

[8] Jooyandeh Mohammadreza1,Mohades Ali1.Uncertain Voronoi diagram.Information Processing Letters, v 109, n 13, p 709-712, June 15, 2009.

[9] Yin Liuguo,Wang Changmian.An energy-efficient routing protocal for event-driven dense wireless sensor networks.International Journal of Wireless Information Networks,v 16,n 3,p 154-164,September 2009.

Research on Clustering Routing Protocol of MANET

Zheng Sihai, Li Layuan, Li Yong, Yuan Junchun
College of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
E-mail: sihaizheng@sina.com

***Abstract*—In general, each cluster consists of a cluster head and several member nodes in clustering routing protocol of MANET. Its advantages are obvious, such as better scalability, less control overhead, and it also reduce the number of nodes sharing the same channel. But it needs a clustering algorithm to maintain the hierarchy so that it can adapt well to changes of network topology, and the routing elected by the clustering algorithm is not always optimal. This paper proposes a new clustering routing protocol EARP. New protocol takes into account the mobility and residual energy; it significantly improves network performance. The results of simulation show EARP is much better than CBRP routing protocol.**

Keywords-*MANET; clustering routing protocol; average delay; weight*

I. INTRODUCTION

In MANET, the routing protocols based on cluster can reduce the number of nodes involved in routing calculation and the size of routing table. For a large scale network, the clustering routing protocol is a better choice. So far, the clustering routing protocol has been researched a lot, and researcher had made many achievements.

CBRP [1] (Cluster based Routing Protocol) is a distributed, scalable and efficient routing protocol. It uses hierarchical mechanism to reduce flooding packets in the process of on-demand routing discovery. Also, local repair mechanism is used to increase the packet delivery rate, reduce delay and routing overhead. Each routing is optimized by shortening mechanism. At first, CBRP send flooding probe packets to all cluster head to seek destination node, then destination node sends back response packets to the source node using reverse loose source path, the cluster head along the path will count the appropriate routing. The upstream forwarding nodes in failure link use topology information within two-hop to repair routing [2] and destination node will inform source node a message.

ZHLS [3] (Zone Based Hierarchical Link State Routing) is another routing algorithm based on partition. It includes two kinds of routing message, which are interior message and external message. The former is used to provide the connection information between the nodes in interior district and only is broadcast in the district. The latter is used to provide the connection information between the nodes in external district and is broadcast in whole network, which reduces the communication and storage overhead. As the interior message does not broadcast to other district, the impact on node-level information caused by mobility is limited to local district. The messages are broadcast to other districts only when the connections with other districts are changed since mobility, but the connections between districts is relatively stable. ZHLS defines two topologies: node-level and district-level [4]. If there is at least one physical link between two intervals, then there will be one virtual link between them. District-level topology consists of virtual links. Each node uses the node-level topology information broadcast in local district and district-level topology information broadcast in whole network to construct routing table.

In the rest of this paper, we first research on AOW clustering algorithm in section Ⅱ. Then, EARP routing protocol is analyzed in section Ⅲ. The experiments and analysis are shown in section Ⅳ. At last, we offer conclusions in Section Ⅵ.

II. AOW ALGORITHM

AOW [5] (Adaptive On-demand Weighting) algorithm is a typical clustering algorithm. When choosing cluster head, it will take into account multiple factors, and make a reasonable compromise based on actual needs and operating environment. Therefore, it uses combined weighted algorithm to select cluster heads, which can improve overall performance of network. Each node is assigned a weight showing whether the node is suitable to act as cluster head. Weight can be expressed as equation (1).

$$W= a\times mo+b\times de+c\times po+d\times en+x \quad (1)$$

Where sign *mo* denotes the mobility of node, *de* denotes the degree of node, *po* denotes the transmission power of node, *en* is the residual energy and *x* shows the other possible factors that may impact on the weight, such as the power of processor and storage space, etc.. Those parameters, *a*, *b*, *c* and *d*, are weighting factor, they are determined by specific application and network environment and they should be normalized.

In AOW algorithm, how to form a cluster structure is very flexible. It can form an overlapping cluster or non-overlapping cluster according to the different clustering strategy, so the control overhead of AOW is less and it is more universal [6]. But AOW algorithm has some inherent weakness. For example, its overhead is much more in the process of clustering and the index of node mobility is less reasonable, etc.. To solve the those problem, we add several constraints into AOW, which can find more appropriate

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.77

parameters to describe system and cost less overhead in the process of clustering. This paper will introduce a new clustering algorithm in detail.

III. EARP ROUTING PROTOCOL

A. Size of Cluster

If there are too much member nodes in a cluster, the cluster head must manage a great number of packets simultaneously, which may cause congestion. On the contrary, if there are too few members in a cluster, the bandwidth of cluster head can not be fully utilized. In order to get best throughput, we can derive the optimal size of cluster using equation (2).

$$R = \theta\left(\frac{W}{\sqrt{n}}\right) \tag{2}$$

Where W is the channel bandwidth. In EAOW algorithm, the cluster head will construct a high-level backbone network, in which the traffic of a cluster is the sum of all traffics. If N is the total number of all nodes, m is the number of cluster, and then the number of nodes in each cluster is N/m. The bandwidth within cluster is denoted by w1, the bandwidth between clusters is denoted by w2, the traffic of each member in cluster is expressed as equation (3).

$$R_{local} = \theta\left(\frac{w_1}{\sqrt{N/m}}\right) \tag{3}$$

The traffic of each backbone member node is expressed as equation (4).

$$R_{backbone} = \theta\left(\frac{w_2}{\sqrt{m}}\right) \tag{4}$$

The cluster head is in charge of the communication between inside and outside cluster, which has two interfaces. That is, it is not only a member of cluster, but also a node work in backbone network. Assuming network traffic is evenly distributed, the traffic which cluster head had exchanged to outside cluster is (m-1)/m of total traffic of cluster head. To avoid congesting, we can derive the following inequality from equation (3) and (4).

$$\frac{m-1}{m} R_{local} < R_{backbone} \tag{5}$$

Now, our goal is to derive the optimal size of cluster. If m=M, where M is the optimal size, then ((m-1)/m)R_{local} will be the maximum, which is in line with equation (10). If m<M, some bandwidth will be wasted. On the contrary, the backbone can not cover all business, which may cause seriously congestion.

According to equation (3) and (4), the ceiling of ((m-1)/m)R_{local} and $R_{backbone}$ is respectively expressed as following,

$$\frac{M-1}{M}\sqrt{\frac{8}{\pi}}\left(\frac{w_1}{\Delta}\right)\sqrt{\frac{N}{M}}\ ,\ \sqrt{\frac{8}{\pi}}\left(\frac{w_2}{\Delta}\right)\sqrt{M}$$

If m=M, then we can derive equation (11) from above analysis,

$$\frac{M-1}{M}\sqrt{\frac{8}{\pi}}\left(\frac{w_1}{\Delta}\right)\sqrt{\frac{N}{M}} = \sqrt{\frac{8}{\pi}}\left(\frac{w_2}{\Delta}\right)\sqrt{M} \tag{6}$$

The solution of equation is expressed as equation (7).

$$M = \frac{w_1}{w_2}\sqrt{N} + 1 \tag{7}$$

If there are large number of N in network, equation (7) is simplified as (8).

$$M = \frac{w_1}{w_2}\sqrt{N} \tag{8}$$

Therefore, if the optimal size of cluster is denoted by δ, then,

$$\delta = \frac{N}{M} = \frac{w_2}{w_1}\sqrt{N} \tag{9}$$

B. The Process of EARP

EARP is triggered if and only if the system boots or cluster heads can not cover all member nodes. The specific process is described as following.

Step 1: Calculate R_i and S_i of node n_i If $S_i \geqslant 2$, this node is involved in running for cluster head, and go to *step* 2. Else, it quit running for cluster head.

Step 2: Seek all neighbors of the node n_i which is involved in running for cluster head. If N_i is the degree of n_i, then:

$$N_i = \left|N\left(n_i\right)\right| = \sum\{dist\left(n_i, n_i^{'}\right) < range\} \tag{10}$$

Where, $n_i^{'}$ is the neighbor which is not in any clusters.

Step 3: Calculate the deviation between node degree and optimal node degree,

$$\Delta_i = \mid N_i - \delta \mid \tag{11}$$

Step 4: Calculate the average distance of all neighbors of n_i, which is denoted by $\overline{D_i}$,

$$\overline{D_i} = \frac{\sum_{n_j^{'} \in N(n_i)} dist(n_i, n_j^{'})}{N_i} \tag{12}$$

Step 5: Calculate the movement speed of node, and obtain the mobility according to equation (1), which is denoted by M_i.

$$M_i = \frac{\sum_{j=1}^{m}\left|M_i(n_j^{'})\right|}{m} \tag{13}$$

Where $n_j^{'}$ is the neighbor of n_i, and the amount of $n_j^{'}$ is m. As the location of node is known in advance, $M_i(n_j^{'})$ can be calculated using equation (19).

$$M_i(n_j^{'}) = 10\lg\frac{dist_{new}(n_i, n_j^{'})}{dist_{old}(n_i, n_j^{'})} \tag{14}$$

Where $dist_{old}(n_i, n_i^{'})$ is the distance between n_i and $n_i^{'}$ at the previous measuring time; $dist_{new}(n_i, n_i^{'})$ is the distance between n_i and $n_i^{'}$ at current time.

Step 6: Calculate the total weight. the weight W_i can be calculate as following:

$$W_i = w_1 M_i + w_2 \overline{D_i} + w_3 \Delta_i \quad (15)$$

Where w_1, w_2, w_3 are weight factors, they are different in different condition, but $w_1+w_2+w_3=1$. More import parameters, the greater its weight factor.

Step 7: each node compares its own weight with its neighbors' with 1-*hops*. If its own weight is the minimum among all neighbors, it will declare itself as cluster head; if there are two nodes with the same weight, and then the one with smaller ID number will declare itself as cluster head.

Step 8: Node n_i send a *CH* message declaring itself is the cluster head to its all neighbors. If a node has received several such messages, it will select the node with minimum ID as its cluster head, and send a JOIN message back to source node.

Step 9: Repeat all steps from *step* 3 to 8 until all nodes find its cluster head.

C. Cluster Maintenance

In an ideal cluster, the distance between nodes is very close and they move slowly, so the cluster is relatively stable and the ruling set is less updated. In fact, some nodes may constantly move in MANET, old cluster may be destroyed in a short time; it is inevitable the old cluster and ruling set are changed and updated frequently [7]. EARP routing maintains clusters using the following mechanism.

Case 1: Change the attachment of cluster. Each member node must track whether the position has been changed between itself and cluster head, when the signal strength of n_i decreased to a certain extent, it will send a message to cluster head so that the cluster head changes its old attachment. According to current location of n_i and its moving direction, old cluster will transfer n_i to another appropriate cluster head, and n_i will join the new cluster. The both cluster head must update the member list at last.

Case 2: Update ruling set. If a member node n_i has left a cluster and there are no available cluster head within the transmission radius of n_i, it shows the old clusters are unable to cover whole network. At this time, EAOW algorithm must be activated immediately to find a new ruling set again.

The overhead in case 1 is less, but it takes place frequently; whereas the whole network is involved in case 2, so the overhead is much more. After the cluster structures are constructed using EAOW algorithm, the member node in cluster must monitor the strength of signal received from cluster head. When the signal is too weak to communicate stably, node will send a message it will leave this cluster to the cluster head, and then it try to join in a neighbor cluster, cluster heads will update the member list.

IV. SIMULATION AND ANALYSIS

To evaluate EARP and compare it to existing routing protocol, simulations are performed. A Mobility Framework for NS 2 [8], a discrete event simulator written in C++, is used as a tool. the scene is 1000m × 1500m, the number of nodes are respectively 10, 60, 110, 160, 210, 26, 310 and 360 in scene, and they are randomly distributed ; The pause time is 0s. The best size of cluster is approximately $\sqrt{N}$ in EARP protocol. Data source is CBR, packet size is 512 bytes, and it sends a packet in each second. Each experiment spent 600s; the initial energy each node has is 400J, and out power and received power are not change.

Fig. 1 shows data delivery rate is rapidly increased while the number of nodes is less than 160, but it is slowly decreased while the number of nodes is more than 160. That is because some paths will become longer if there are more nodes in network, more and more packets are dropped in process of transmission, then the two data delivery rates are decreased. With the increase of number of nodes, the network topologies become more and more complex. The minimum ID clustering algorithm used by CBRP must re-cluster frequently, which greatly consumes more maintenance overhead. Whereas EAOW algorithm used by EARP had considered the mobility, each cluster can keep relatively stable; it does not need re-cluster frequently. Hence, its packet delivery rate is more than that of CBRP.

Fig. 2 shows the relationship between network lifetime and the number of nodes. With the increase of number of nodes, the network lifetime both incline slowly. That is because more nodes, more complicated topology and more data forwarded, which make load increase rapidly and residual energy is run out in a short time. Thus, network lifetime is declining. The network lifetime of EARP is apparently longer than that of CBRP in case of the same number of nodes. In CBRP routing protocol, a cluster head, the node with the minimum ID, will consume much more energy than member nodes. In general, the cluster head will soon run out of its energy. Whereas the cluster head will become a member node when its energy is less than the specific threshold in EARP, and a new cluster must be elected. This kind of mechanism can effectively prolong network lifetime.

Fig. 3 shows the relationship between routing overhead and the number of nodes. From fig. 4 we find more nodes, more routing overhead. As there are more nodes, both protocols must maintain more routing information in cache and great amount of control message should be forwarded. But the two lines are very close, showing EARP only increases a little routing overhead.

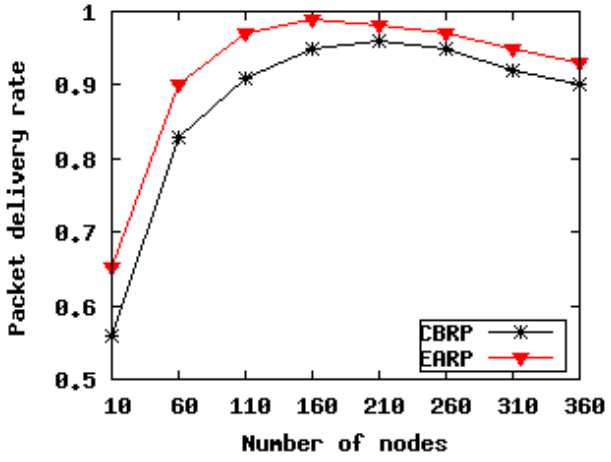


Fig. 1 Packet delivery rate and number of nodes

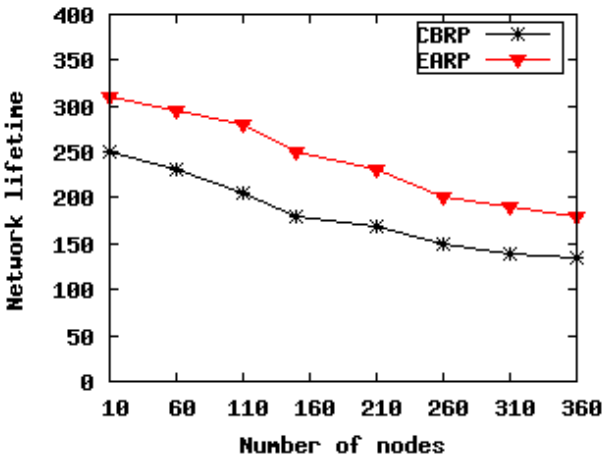


Fig. 2 Network lifetime and number of nodes

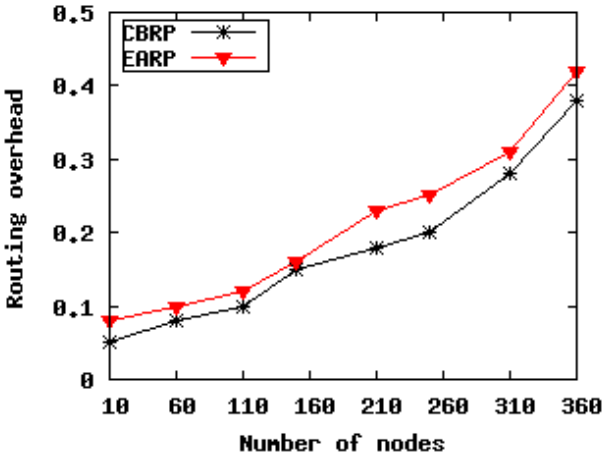


Fig. 3 Routing overhead and number of nodes

V. CONCLUSION

This paper proposes a clustering routing protocol EARP. The main idea of EARP is to restrict cluster head candidate according to its residual energy and the throughput of whole network. EARP can improve the performance of network and enhance the ability of load balancing. The results of simulation shows EARP can prolong network lifetime and enhances the robustness of network. In the future work, we might focus on how to include QoS support into EARP, so that it can provide a certain quality of service.

ACKNOWLEDGMENT

This paper is supported by National Natural Science Foundation of China (No: 60672137, 60773211, 60970064), Open Fund of the State Key Laboratory of Software Development Environment (No: SKLSDE-2009KF-2-02), New Century Excellent Talents in university (No: NCET-08-0806), Specialized Research Fund for the Doctoral Program of Higher Education of China (No：20060497105).

REFERENCES

[1] Lee Keun-Ho; Suh Heyi-Sook; Han Sang-Bum. An authentication protocol based on CBRP in MANET. 6th International Conference on Advanced Communication Technology: Broadband Convergence Network Infrastructure, v 1, p 407-412, 2004.

[2] Hosseini Seno; Seyed Amin; Budiarto R. A routing layer-based hierarchical service advertisement and discovery for MANETs. Ad Hoc Networks, v 9, n 3, p 355-367, May 2011.

[3] Huang, Tsung-Chuan; Liao, Chun-Kai; Dow, Chyi-Ren. Zone-based hierarchical routing in two-tier backbone ad hoc networks. Proceedings - IEEE International Conference on Networks, ICON, v 2, p 650-654, 2004.

[4] Li, Dao-Quan; Xue, Wei-Hua; Cao, Qi-Guang. Research on ad hoc network clustering algorithm based on signal strength. Proceedings - 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems, ICIS 2010, v 1, p 719-722, 2010.

[5] Tarng, Jenn-Hwan; Chuang, Bing-Wen; Wen, Yi-Luen. A radio-link adaptive routing protocol for mobile Ad Hoc Network. IEEE Vehicular Technology Conference, v 2, p 678-682, 2006.

[6] Karunakaran, S.; Thangaraj, P.. An adaptive weighted cluster based routing (AWCBRP) protocol for mobile Ad-hoc networks. WSEAS Transactions on Communications, v 7, n 4, p 248-257, April 2008.

[7] Meghanathan, Natarajan. A simulation study on the impact of mobility models on the network connectivity, hop count and lifetime of routes for ad hoc networks. Informatica (Ljubljana), v 34, n 2, p 207-221, June 2010.

[8] Chen, Lixin. NS2 based performance measurement of mobile MANET routing protocols. Journal of Computational Information Systems, v 3, n 1, p 109-115, February 2007.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Formal Support for Cyber Physical System Specification Using Aspect-Oriented Approach

Lichen Zhang
Faculty of Computer Science and Technology
Guangdong University of Technology
Guangzhou 510090, China

***Abstract*—Cyber-physical systems pose considerable technical challenges, ranging from the distributed programming paradigms to networking protocols with timeliness as a structuring concern, including systems theory that combines physical concerns and computational concerns Cyber Physical systems is deemed correct in case it abides to the functional correctness as well as a set of quantitative characteristics Formal specification techniques for such systems have to be able to describe all these concerns. Unfortunately, a single specification technique that is well suited for all these concerns s is yet not available. Instead one finds various specialized techniques that are very good at describing individual concerns of cyber-physical system. This observation has led to research into the combination and semantic integration of specification techniques. This paper proposes an approach for specifying real time cyber physical systems based on aspect-oriented formal specification, which exploits the diversity and power of existing formal specification languages. We provide an aspect-oriented specification approach based on the combination of Object-Z and Time-CSP. We increase separation variable in the basic elements and crosscutting concerns to achieve formalization of aspect-oriented characteristics. This aspect oriented formal specification method simplifies the requirement analysis process of real time cyber physical systems. Two case studies illustrate the specification process of aspect-oriented formal specification for real time cyber physical systems.**

Keywords- Aspect-oriented; Real Time; Cyber Physical Systems; Formal Method; Object-Z; Timed-CSP

I. INTRODUCTION

The specification of requirements on the observable behavior of real time cyber physical systems is an important step in the engineering of these systems. Requirements specifications help avoiding inconsistencies in the requirements, they are the basis for deriving correct system designs, they are essential in establishing the system's correctness by serving as a basis for testing and formal verification, and they are very important in the documentation of system requirements. The complexity of software projects in the area of real time cyber physical systems calls for automated tool support. This can only be provided if the underlying engineering methods have a formal foundation. Formal Description Techniques FDTs like Object-Z[1]and Timed-CSP [2] [3] have been successfully applied to the specification of ''traditional'' communication protocols, services and network applications. The terms Cyber-Physical Systems (CPS)[4] have come to describe the research and technological effort that will tightly conjoin real-world physical objects and computing systems. The integration of physical processes and computing is not new. Embedded systems have been in place since a long time to denote systems that combine physical processes with computing. The revolution will come from extensively networking embedded computing devices, in a blend that involves sensing, actuation, computation, networking, pervasiveness and physical processes. Such systems pose considerable technical challenges, ranging from the distributed programming paradigms to networking protocols with timeliness as a structuring concern, including systems theory that combines "physical concerns" (control systems and signal processing) and "computational concerns" (complexity, schedulability, computability, scalability, etc.). Applications of Cyber-Physical Systems include, among others, critical infrastructure control (electric power, water resources, gas and fuel distribution, transportation, etc.), process control and manufacturing, highly dependable medical devices and systems, traffic control and safety, advanced automotive systems, energy conservation and environmental control. Cyber Physical systems is deemed correct in case it abides to the functional correctness as well as a set of quantitative characteristics, frequently called *QoS* characteristics. We conjecture that the Cyber Physical systems systems that are currently being developed are largely software driven, which is why our methods are relevant to the engineering of these systems.

Aspect-oriented programming (AOP) [5] [6]is a new software development technique, which is based on the separation of concerns. Systems could be separated into different crosscutting concerns and designed independently by using AOP techniques.

In this paper, we provide some ideas for the aspect – oriented formal specification of real time cyber physical systems and two well known case studies to validate aspect-oriented formal specification.

II. ASPECT-ORIENTED FORMAL SPECIFICATION

Aspect-oriented approaches use a separation of concern strategy, in which a set of simpler models, each built for a specific aspect of the system, are defined and analyzed. Each aspect model can be constructed and

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.45

evolved relatively independently from other aspect models. This has three implications:

- An aspect model can focus on only one type of property, without burden of complexity from other aspects. Hence an aspect model is potentially much simpler and smaller than a traditional mixed system model. This is expected to dramatically reduce the complexity of understanding, change, and analysis.
- Different levels of detail or abstraction can be used in the construction of different aspect models. This allows us to leverage of existing understanding of certain aspect of the system to reduce the complexity of modeling and analysis. For example, if the timing property of a component/subsystem is well understood, we can build an extremely simple timing model for the component.
- Existing formal notations normally are suitable for describing one or a few types of system properties. By adopting the aspect concept, we can select the most suitable notation to describe a given aspect. Likewise, we can select the most suitable analysis techniques to analyze a given property.

Although the use of mathematical logic is a unifying theme across the discipline of formal methods, there is no single best "formal method". Each application domain requires different modeling methods and different proof approaches. Furthermore, even within a particular application domain, different phases of the life-cycle may be best served by different tools and techniques. In many ways, Object-Z and Timed CSP complement each other in their capabilities. Object-Z has strong data and algorithm modeling capabilities. The Z mathematical toolkit [7]is extended with object oriented structuring techniques. Timed CSP has strong process control modeling capabilities. The multi-threading and synchronization primitives of CSP are extended with timing primitives. Moreover, both formalisms are already strongly influenced by the other in their areas of weakness. Object-Z[8][9] supports a number of primitives which have been inspired by CSP notions such as external choice and synchronisation. CSP practitioners tend to make use of notation inspired by the Z mathematical toolkit in the specification of processes with internal state. The approach taken in the TCOZ[10] notation is to identify operation schemas (both syntactically and semantically) with (terminating) CSP processes that perform only state update events; to identify (active) classes with non-terminating CSP processes; and to allow arbitrary (channelbased)communications interfaces between objects.

Aspect-oriented specification is made by extending TCOZ notation with aspect notations. The schema for aspect specification in has the general form as shown in Fig.1, Fig.2 and Fig.3.

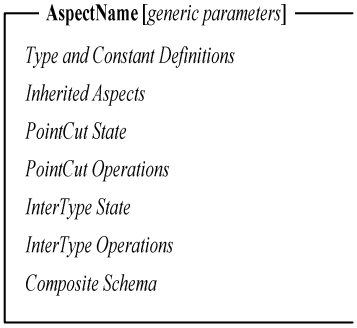


Figure 1. Aspects of Model Structure

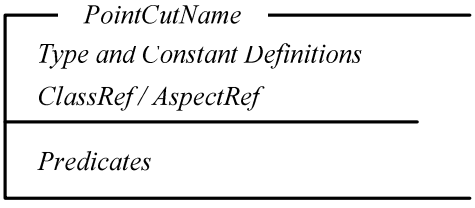


Figure 2. PointCut Operation Schema of Structure

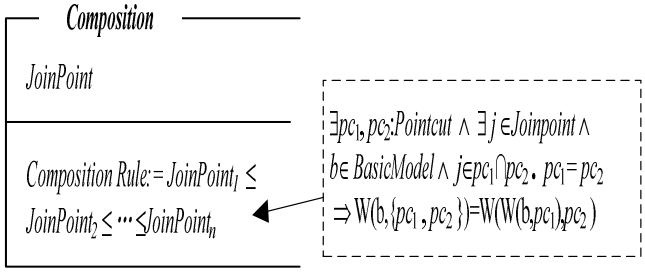


Figure 3. Composition Schema of Structure

III. CASE STUDY: ASPECT-ORIENTED FORMAL SPECIFICATION OF A TRAIN CONTROL SYSTEM

The Problem that must be addressed in operating a railway are numerous in quantity, complex in nature, and highly inter-related . For example , collision and derailment, rear-end, head-on and side-on collisions are very dangers and may occur between trains. Trains collide at level crossing. Derailment is caused by excess speed, wrong switch position and so on. The purpose of train control is to carry the passengers and goods to their destination, while preventing them from encountering these dangers. Because of the timeliness constraints, safety and availability of train systems, the design principles and implementation techniques adopted must ensure to a reasonable extent avoidance of design errors both in hardware and software. Thus a formal technique relevant to design should be

applied for train systems development. The purpose of our exercise is to apply aspect -oriented formal method s to develop a controller for train systems that tasks as input:

A description of the track configuratin

A sequence of description of the moves of each of those trains.

The controller should take care of trains running over the track. It should control the safety of the configuration, ie. no two trains may enter the critical section. When one critical section is occupied, some others, which share some part of section with this one , should be locked. The controlled can control the status, speed, position of trains. In order to keep the description focused , we concentrate onsome particular points in train control systems rather than the detailed descriptions of all development process. The specification is made by integrating Object-Z and CSP[11][12][13].

A train controller limits the speed of the train, decides when it is time to switch points and secure crossings, and makes sure that the train does not enter them too early as shown in Fig.4.

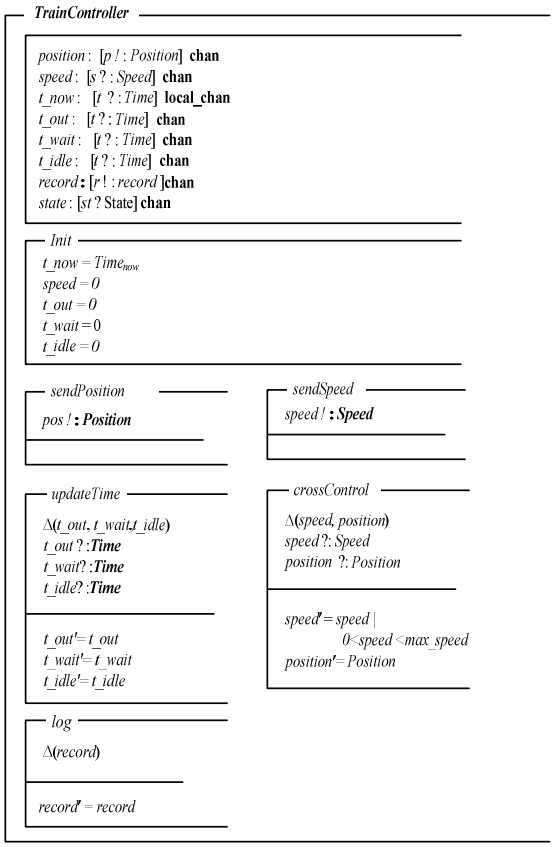


Figure 4. Control Model of Train Dispatching System

In automatic railway Crossing system., The sensors detect the presence of the train near rail crossing and barrier shuts down when train is approaching to the railway crossing. Once train crosses the rail crossing barriers opens by itself as shown in Fig.5 and Fig.6.

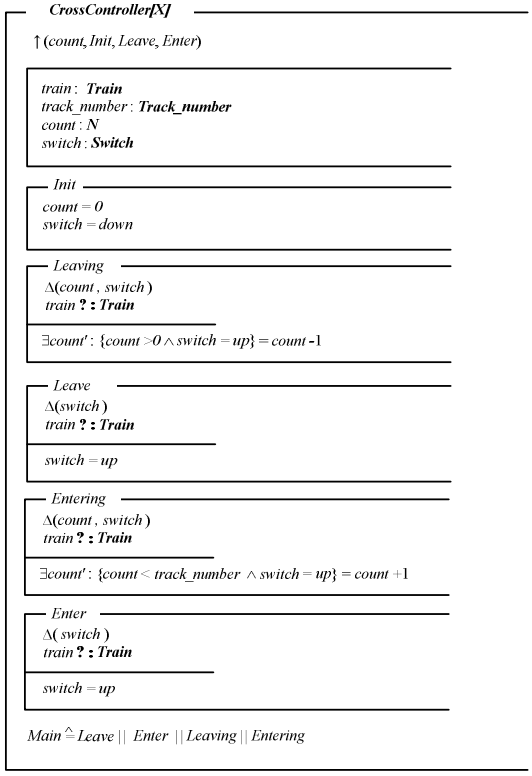


Figure 5 Trains through the intersection mode

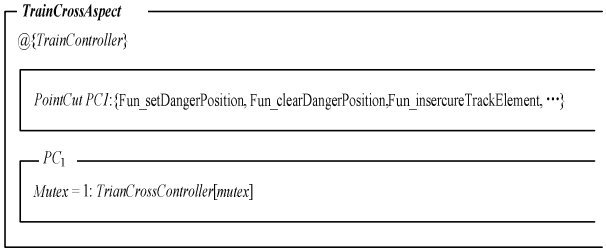


Figure 6. Crossroads aspects model

To extend state-based and behavioural techniques with real-time aspects, different approaches exist. One approach is unifying a state based language and Timed CSP , an extension of CSP with a time-out operator. In this paper ,we take a different approach. We integrates CSP and Object-Z with aspect-oriented technique to specify real-time systems as shown in Fig.7 .

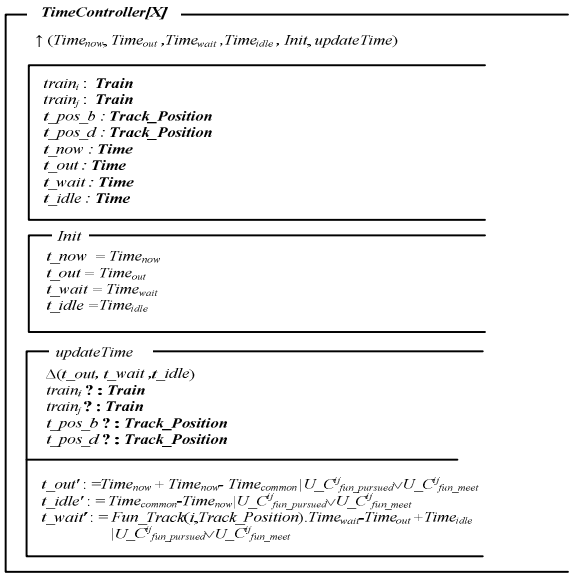


Figure 7 Time Aspect Formalization Method

Woven model is shown as Fig.8

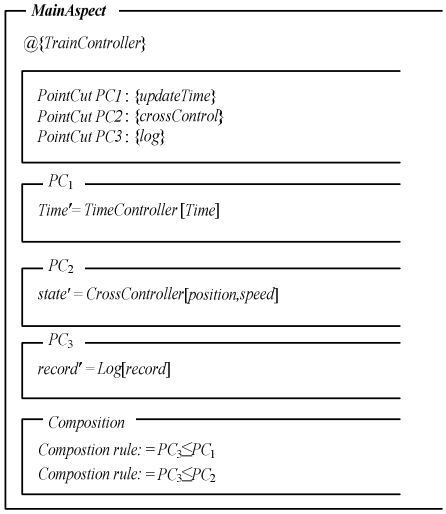


Figure 8 Woven Aspects of Diagram.

IV. CASE STUDY: ASPECT-ORIENTED FORMAL SPECIFICATION OF A LIFT SYSTEM

A lift consists of four parts [14]: a door for allowing access to and from the lift, a shaft for transporting the lift, an internal queue for determining the lift itinerary, and a controller for coordinating the behaviour of the other components. The door cycle is initiated by receipt of an *open* signal from the lift controller and completed by sending a *close* signal as shown in Fig.9[14]. Floors may be divided into two classes,those from which it is possible to travel upward and those from which is possible to travel downward. The *TopFloor* is a floor from which only downward travel is possible , A *MiddleFloor* is a floor from which both upward and downward travel is possible. A *MiddleFloor* is a floor from which both upward and downward travel is possible. The different floor model is shown as Fig.10[14]. The following timing properties must be captured in the model: lift travel time between two consequent floors is a constant, however there is a constant time delay for acceleration and braking; without interrupts, the lift door should be kept at the `open' state for a fixed time period before closing. The maximum time to pass from one floor to another is *t* for each floor travelled plus a delay of *delay* caused by initial acceleration and final braking of the lift. The shaft model is captured in Fig.11[14].

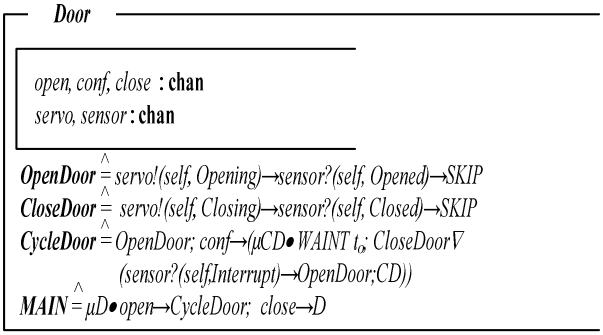


Figure 9 Life Door Mode

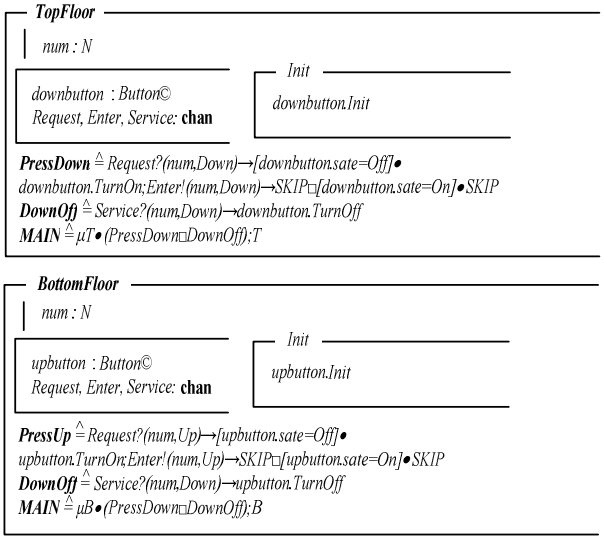


Figure 10 Different Floor Mode

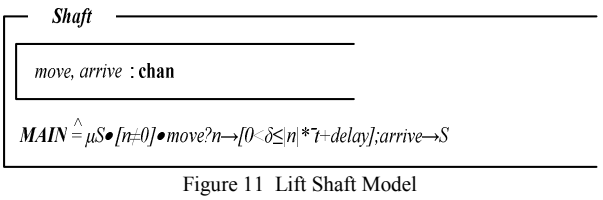


Figure 11 Lift Shaft Model

The lift controller keeps record of the current floor and movement direction and provides the interface between the lift environment and the other lift components as shown in Fig.12[14].

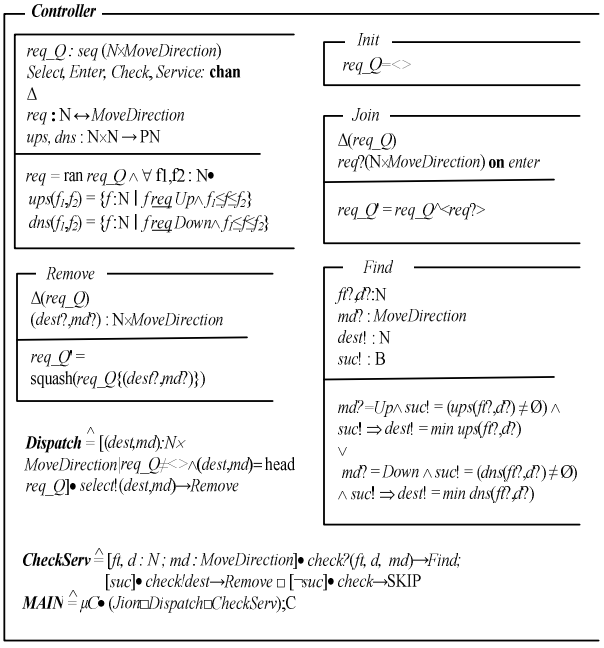


Figure 12 Lift System Controller Model

A lift can not go upward and downward simultaneously , Lift Move Aspect of Control Model is expressed as Figure 13 and Figure 14.

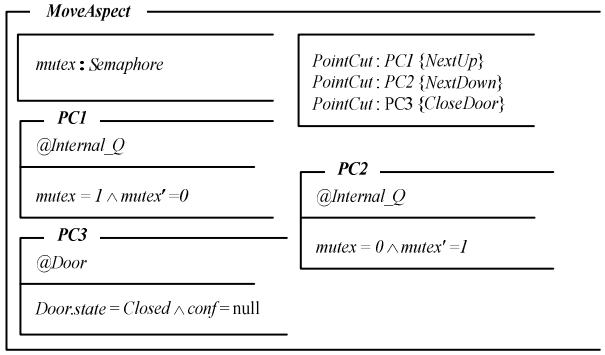


Figure 13 Lift Move Aspect of Control Model

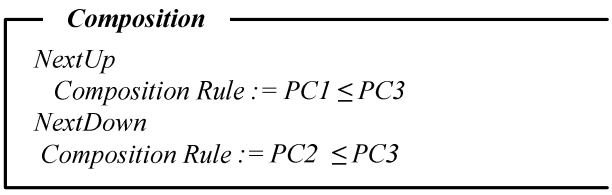


Figure 14 Lift Move Aspect of Composition Model

V. CONCLUSION

In this paper we proposed to use aspect-oriented formal specification for real time cyber physical systems, we provided some ideas for aspect-oriented formal specification of real time cyber physical systems. Two case studies was used to illustrate the specification process of aspect-oriented formal specification for real time cyber physical systems.

The further work is devoted to Integrated aspect-oriented formal specification with UML further.

ACKNOWLEDGMENT

This work is supported by the Major Program of National Natural Science Foundation of China under Grant No.90818008.

REFERENCES

[1] Graeme Smith. The Object-Z Specification Language[M]. Software Verification Research Centre University of Queensland. 2000

[2] Reed G M, Roseoe A W. A timed model for communicating sequential processes[C]//Pro ICALP'86. Lecture Notes in Computer Science. Berlin:Springer,1986.

[3] Davies J, Schneider S. A Brief History of Timed CSP[J]. Theoretical Computer Science, 1995, 138(1):243-271.

[4] Edward A.Lee. Cyber Physical Systems Design Challenges. Object Oriented Real-Time Distributed Computing (ISORC), 2008.11th IEEE International Symposium. January 23,2008.

[5] Kiczales G,et al.Aspect-Oriented Programming. In:proc.of the 11th European Conf. on Object-Oriented Programming, June 1997.

[6] Aspect-Oriented Software Development. http://aosd.net/.

[7] Jonathan Bowen. Formal Specification and Documentation using Z[M]. 2003

[8] Graeme Smith. The Object-Z Specification Language[M]. Software Verification Research Centre University of Queensland. 2000

[9] J.S. Dong, L. Zucconi, and R. Duke.Specifying parallel and distributed systems in Object-Z. In G. Agha and S. Russo, editors, The 2nd International Workshop on Software Engineering for Parallel and Distributed Systems, pages 140–149, Boston, Massachusetts, 1997. IEEE Computer Society Press.

[10] B. P. Mahony and J.S. Dong. Blending Object-Z and Timed CSP: An introduction to TCOZ. In The 20th International Conference on Software Engineering (ICSE'98). IEEE Computer Society Press, April 1998.

[11] Jochen Hoenicke. Specification of Radio Based Railway Crossings with the Combination of CSP, OZ, and DC. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.4394

[12] Johannes Faber, Swen Jacobs, Viorica Sofronie-Stokkermans. Verifying CSP-OZ-DC Specifications with Complex Data Types and Timing Parameters. Integrated Formal Methods 2007, July 3rd

[13] Aleš JANOTA. Using Z Specification for Rail Way Interlocking Safety. Periodica Polytechnica Ser. Transp. ENG.VOL.28,NO.1–2,PP.39–53(2000)

[14] Brendan Mahony and Jin Song Dong. Timed Communicating Object-Z[J]. IEEE TSE 26(2):150-177,Feb 2000.

The Research of Web Services Management Based on WSDM

Lou Yuan-sheng, Ji Xiao-kui, Yue Lu-lu
College of Computer & Information
Hohai University
Nan Jing, P.R.China, 210098
Wise.lou@163.com,xiaogui4595@163.com

Gao Jian-bin
China Navy Army serial No: 91669.
Hai kou, P.R.China, 571100
gjb92093@hhu.edu.cn

***Abstract*—As the Web Services technology is widely used in the enterprise's core business, how to manage these resources is becoming to extremely important problem. In this paper, after analysising the characteristics of resouce management in distributed environment and the main content of WSDM, a novel Web Service management scheme is introduced, and then the main components in the scheme is introduced, as well as the relationship between them. Also we give out the implementation of the main compoents, and design a typical case used to verify the feasibility of the scheme.**

Keywords—Web Service, WSDM Resource, Web Service Management;

I. INTRODUCTION AND RELATED WORKS

As more and more Web Services are used into applications and being integrated into business' applications, how to monitor and manage these IT resources becomes to a extremely important problem [1]. This is the typical Web services management issues. It has two parts: how to manage Web services itself and how to use Web service to manage the heterogeneous and distributed systems [2].

Web services management is an important part of Web service technology, and has been given attention both in academic and industry adequately. We can divide those studies into three levels: infrastructure level, application level and business level [3]. In this paper we focus on application level. In this field, Bart et.al proposed WSML (Web Service Management Layer), and use AOP technology to achieve its core functions, include dynamic service selection and service composition, client service Management, and support the selection, integration and composition rules [4]. Vladimir proposed WSOL (Web Service Offerings language), focus management of information standard, such as: service classification, function and accessibility constraints, pricing, penalties and other important management information [5]. The application of research in Web services management-level research has some significance, but not generic.

WSDM (Web Services Distributed Management) is a management technology developed by the OASIS Web Services Distributed Management Technical Committee, which includes two parts: Management Using Web services (MUWS)[6][7] and Management of Web services (MOWS) [8]. MUWS defines how the resources to provide outside management interface for implementation and management by Web services technology. MOWS based on MUWS, uses Web service technology to manage theirselves. In WSDM, Web service endpoint provide managed resource's manageability interface, make consumers through the Web service endpoint to obtain the necessary management services [9]. Web service uses Web services technology as the basis for the distributed management framework, and use the SOAP protocol for exchanging information, so that WSDM standard is better both in cross-platform and interoperability[10][11].

In this paper, based on WSDM standards, combined with its main technical characteristics, give consideration to the distribution characteristic of distributed resource management, we proposed a new Web service management scheme.

II. THE FRAMEWORK OF WEB SERVICES MANAGEMENT BASED ON WSDM

Agent is an entity with control functions, it receives information, and then according to their own knowledge, rules and control logic to process the information and then forwards the information to the receiver. Agent is commonality, it means that the Agent communication language independent Agent itself. This characteristic makes it ideal for solving the problem of heterogeneous resource. Therefore, Agent is ideal for dynamic and scalable environment [12].

Agent use the public language to communicate, and their semantics independent of the specific Agent. So the Agent can act as intermediate processing units. And it represents an entity with specific interface, and provides public interface, which is the key to solve the heterogeneous resources. In this paper, we assume that the Agent can receive all the messages from different Web services and can process these messages correctly.

We propose a three-tier Web service resource management scheme after analysing the WSDM standard's features and considering the Agent's commonality. In this scheme, we propose a reasonable WSDM resource model; it's the basis of the scheme. The figure1 shows the framework, the three layers are: Manager layer, WSDM layer and Agent layer.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.98

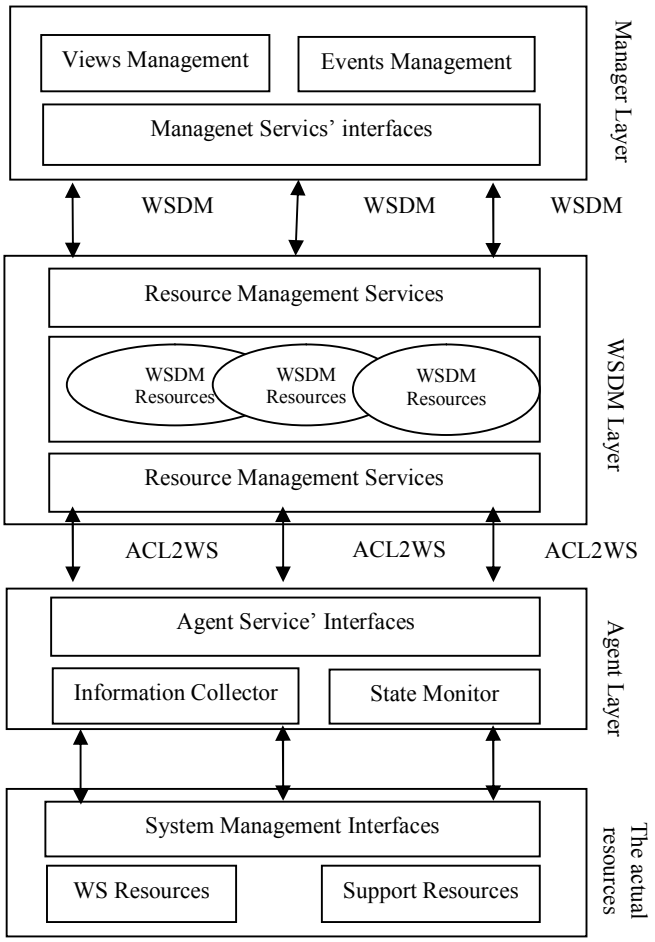


Figure 1. The framework of web servics management based on WSDM

A. *Manager Layer*

Manager layer is the manageability consumer. It manages the manageable resources by invoking the management services' interfaces which are providing by the WSDM layer. This layer has the following components.

- **Configuration Component:** configure the parameters of the target system mainly, such as the setting the load parameter and buffer size management.
- **View Component:** view the metric of the managed resource, such as the CPU utilization, bandwidth, memory utilization, number of the requests and the number of the requests is processed.
- **Event Component:** subscribe and collect the management events.When a management event occurs, the manageability consumer receives the event through this component, and according to the event content to perform the appropriate management action.
- **Execute Component:** execute the management activities of manageability consumers.

B. *WSDM Layer*

This layer is located between the manager layer and the Agent layer. In this layer, the information of the actual resources is collected by the agent mapped to the stateful value of WSDM resource object's attribute, and maintain consistency between the two, and the management services interfaces mapped to the agent service interface. The WSDM resources associated with managed objects reside in this layer. Those managed objects contain information about how to communicate with WSDM resources, and manageability consumer maintains the managed resource through the resource management services. Main components in this layer are as fellows.

- **Resource Management Component:** It is container of Web service resources. It is responsible for managing the whole life cycle of the resources, searching resources and saving resources, and so on. In the resources' storage mechanism, we design different storage strategy according to different types of resource state. Resource state value can be stored on the hard driver or in memory. It mainly based on the updating frequency of the state value. Such as the type of the underlying platform attribute is immutable, it is suitable to store on hard driver. But the CPU utilization, the memory utilization is changeable; it is suitable to be stored in memory. Only when exception occurred, the information should be saved into log message.
- **Function Query Component:** manageability attribute mapped to specific manageability function. The manageability consumer's request mapped to the specific manageability function through the component. The mapping shows in the table I.

TABLE I. ATTRIBUTE-FUNCTION MAPPING TABLE

Resource Attributes	Manageability Functions
ResourceID	IdentityCapability
Description	DescriptionCapability
Manageability	ManageabilityCapability
CorrelatableProperties	CorrelatableCapability
Metric	MetricCapability
OperationalStatus	OperationalStatusCapability
State	StateCapability

Resource attribute includes: ResourceID, Description, Manageability, CorrrelatableProperties, Metric, OperationalStatus and State. The name of the manageabilty function includes: IdentityCapability, DescriptionCapability, ManageabilityCapability, CorrelatableCapability, MetricCapability, OperationalStatusCapability and StateCapality. For different managed object, the corresponding management may be different; these functions may or may not exist. This table can be redesign according to the characteristics of the managed object.

This component first receives the resourceID and resource attribute name, which transmit from the resource management component. And then find the corresponding management functions through the mapping table. In the end the component return the function name to the function implementation components.

Function Implementation Component: the implementations of the management functions. Each management function holds the corresponding status value of management attribute. The management function is implemented by changing the status value. In order to ensure consistency between the managed resource object and the actual managed resource, we take the synchronization strategy. The method is that if the manageability consumer change the status value of the managed resource object, while need to notify the agent which is corresponding to the actual managed resource, and then the agent invoke the system management interface to finish the management action. So the consistency has been maintained. In the other way, if the actual resource's state has been changed, the agent will send the management event to the function implemented component, then the component parse the event and modify the corresponding status values, and then send the SOAP message to the resource management component.

C. *Agent Layer*

Agents located in this layer. It is mainly responsible for monitoring and collecting information about the managed resources. And also it executes managed action by invoking system management interfaces. At the same time the agent interfaces provide basic management services for the WSDM layer. The developers can choose the technology they wanted to develop the agent. And the developers can design different agents according to the features of the managed resources. In this layer, the information collected by agent is encapsulated as web service announcements. And then these announcements has been transmit to resource management service in WSDM layer, after parsing the content, the corresponding attribute values will be full into the appropriate management attribute fields.

III. PROTOTYPE SYSTEM AND AUTHENTICATION

In this paper, we design and develop an prototype system in order to validating the scheme which is introduced in part 2. This part is used to describe the function and its implementation about the system. In the prototype system, we take Web service container for example, and then demonstrate the modeling process about this kind resource modle. In the experimental part, we demonstrate the management process.

A. *WSDM Resource modeling description*

WSDM resource is the basis of the scheme. And the management approach has been expansion on the WSDM resource model. Here we take Web service container for example to demonstrate the WSDM resource model, the description of this model is shown in the table II:

TABLE II. WSDM RESOURCE MODEL

```
<rs:schema
targetNamespace="http://experiment.com/wsdmResource/Container.xsd">
<rs:element name="container">
        <rs:complexType>
        <rs:sequence>
        <rs:element ref="muws-pl-xs:ResourceID"/>
                        <rs:element ref="muws-pl-xs:Caption"/>
                        <rs:element ref="muws-pl-xs:Description"/>
                        <rs:element ref="muws-pl-xs:Version"/>
                        <rs:element ref="muws-p2-xs: OperationalStatus"/>
                        <rs:element ref="muws-p2-xs: ObjectPool">
                        <rs:complexType>
                        <rs:sequence>
                                        <rs:element name="InPoolSize"/>
                                        <rs:element name="OutPoolSize"/>
                                        <rs:sequence>
                        </rs:complexType>
                        </rs:element>
                        <rs:element ref="muws-p2-xs: ThreadPoolSize">
                        <rs:complexType>
                        <rs:sequence>
                                        <rs:element name="ActiveCount"/>
                                        <rs:element
                                name="ComplentTaskCount"/>
                                        <rs:element
                                name="CurrentTreadCount"/>
                                        <rs:sequence>
                        </rs:complexType>
                        </rs:element>
                        …
                        …
        </rs:sequence>
        </rs:complexType>
</rs:element>
</rs:schema>
```

After establish the resource model, we can use the WSDL2JAVA tool which provide by MUSE to generate the corresponding java class.

B. *The Main Component Implementation*

Here we use a simple resource management container to simulate the main functions of resource management component. Figure II shows the structure of the container.

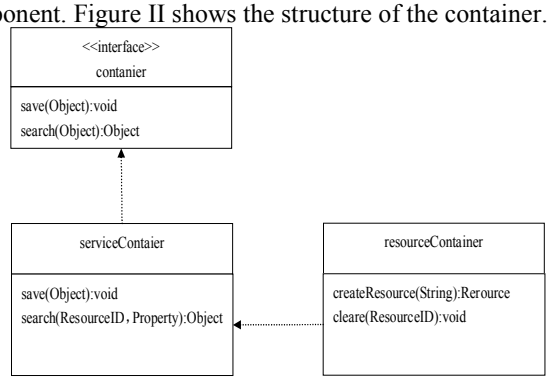


Figure 2. The Structure of the Container

ResourceContainer is a resource management container which is responsible for the creation and destruction of WSDM resources object. The method createResource (String) is used to create a WSDM resource object, the input parameter is the type of the resource object. Such as

createResource ("WebServer"), it means create a WSDM resource object whose type is Web Server. WebServerResource.clear (ResourceID) is used to destroy a managed object. The input parameter ResourceID is a javaBean, which identifies a unique resource object. ResourceID has four different attributes: identifier, name, version and semanticsindetifier. If the method clear (ResourceID) is invoked, the corresponding object in the container will be destroyed, and the corresponding data in the database will be deleted.

The class ServiceContainer is mainly used to save and search resources. The method save (Object) is used to storage the resource to database according to the parameter Object. The method search (ResourceID, Property) is used to find out the resource according to ResourceID and Property. The input parameter ResourceID has been demonstrated in above. And the input parameter Property is used to define the manageability attribute, such as Description, Manageability, CorrelatableProperties, Metric, OperationalStatus and State. There is one point should be noted: the above manageability attributes has been created as a javaBean in our system, and they are all implemented the same interface Property. The method getState (ResourceID) is used to get the state in life cycle of the resource object. The method recoverStatus (ResouceID) is used to recover the status of the resouce object. We can finish the main function about life cycle management through the methods be mentioned above.

We use a simple java class to simulate the main function of the function query component. The figure III shows the structure of the class:

Search
search (ResourceID):ResourceType search (ResourceType):List search (ResourceID,ManageType):ManageAbility

Figure 3. The Search class' structure

In this class we define some different search methods. They have different input parameters and different functions. The searching process is: these methods will return the manageablity according to the input parameters through querying the mapping table. If the result is null, it will return an error flag.

In the same way, we use a simple java class to simulate the main function of the function implementation component. The figure IV shows the structure of the class:

ManagementImpl
get（Object）：int update（object）：void send（AID）：boolean receive（Object）：boolean

Figure 4. The ManagementImpl class' structure

The method get (Object) uses to get the attribute value corresponding to the object. The method update (Object) uses to update the attribute value corresponding to the object. The method send(AID) uses to sent management event to agent, and receive the event content.

Other main functions of the system are similar to the above function. So we don't see them here.

C. *Typical Case*

In our current prototype system, we just have a simple service management. And the target managed system is a Web service system which provides querying service. The specific scheme is that: using the pototype system to monitor the target system's operational status, and adjust the number of the servers providing service according to the target system's load condition, when its overload, new server will be added in order to meet the customers' need and ensure the quality of the service.

Specific experiment is: using four PC as a client to run Jmeter tool, Jmeter sends querying requests to the target system. At the same time, the management system's administrator can monitor the whole running condition of the target system through the management system. The administrator will view the running condition of the target system through calling different methods which define in view component, such as NumberOfRequest , MaxResponseTime, AvarageResponseTime, CPU Utilization, etc., and he can aslo customize the management event (in this case we only have one overloaded event.).

With the acceleration of frequency queries, the target system's load is increasing, when the system load exceeds the threshold we set, Agent component will trigger an overload event, and then sent it to the management system, the system's view component will display the overload condition. Then the system administrator will call the method startServer () to add a new Web Server which provide the same service. The system will return to normalstate after its load is restored to below the setting threshold.

The results show in the last four figures:

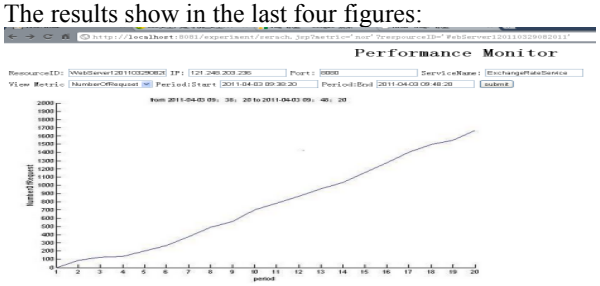


Figure 5. NumberofRequest's Quering result

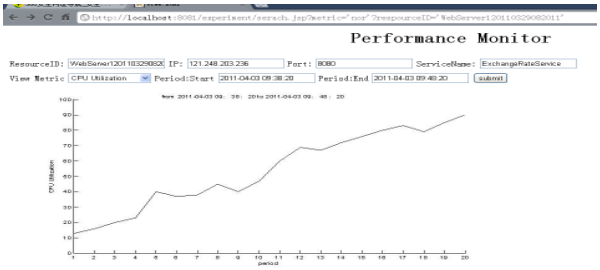


Figure 6. CPU Utilization's Quering result

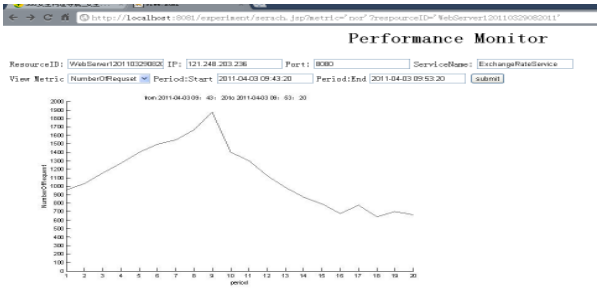


Figure 7. NumberofRequest's Quering result after adding server

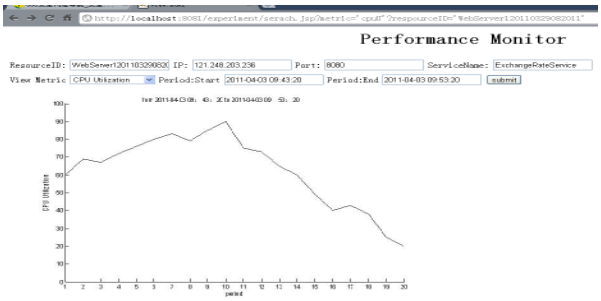


Figure 8. CPU Utilization's Quering result after adding server

From the comparison of these results above, we can see that: after adding new server, the target system's load recovers to normal balance. So the adding new server action is work. So it shows that our scheme is feasible.

IV. CONCLUSION

In this paper, we focus on the need for Web service management. We proposed a new Web service management scheme after analysis the WSDM standard and combining with the management features underling distributed environment. We also introduce the agent technology using into the scheme, so the agent can use the managed resources' manageability.

ACKNOWLEDGMENT

The work is supported by the National High-Tech Research and Development (863) Plan of China under Grant No.2007AA01Z178.

REFERENCES

[1] Mark Potts,Igor Sedukhin,Heather Kreger,Ellen Stokes.Web Service Manageability-Specification http://www.ibm.com/developerworks/library/specification/ws-manage/

[2] Micael P.Papazoglou,Willem-Jan van den Heuvel.Web service management: A survey.IEEE Computer Society,2005,9(6):58-64.

[3] Verheecke B.,Cibran M.A.Aspect-Oriented Programming for Dynamic Web Service Monitoring and Selection.Processdings of the European conference on Web Services,2004.

[4] Tosic V.,Ma W.,Pagurek B.,Esfandiari B.Web Services Offerings Infrastructure(WSOI)-A Management Infrastrusture for XML Web Services.IEEE ,IFIP,2004.

[5] Web Services Distributed Management:Management Using Web Services part1(MUWS1.0 part1).OASIS Standard,Mar,2005.

[6] Web Services Distributed Management:Management Using Web Services part2(MUWS1.0 part2).OASIS Standard,Mar,2005.

[7] Web Services Distributed Management:Management Of Web Services(MOWS1.0).OASIS Standard,Mar,2005.

[8] Karl Czajkowski.Donald F Ferguson.Ian Foster.WS-Resource Framework v1.0.2004. http://www.globus.org/wsrf/specs/ws-wsrf.pdf

[9] Xuan Thang, Ryszard Knowalczyk,"Enable Agent-based Management ot Web Services with WS2JADE",QSIC 2005:407-412.

[10] Moding Statusful Resources Using Web Services.http://www-106.ibm.com/developerworks/library/ws-resource/ws-modelinggresources.pdf.

[11] Management Using Web Service – A proposal Architecture and Roadmap.IBM,CA,HP.

[12] M.Luck, P. McBurney and C. Preist. Agent technology:Enabling next generation computing(a roadmap for agent based computing).AgentLink Report,2003.

A Distributed Artificial Immune Network for Optimizing Tracer Kinetic Models with MATLAB Distributed Computing Engine

Rui LU [1]

1. School of IoT Engineering
Jiangnan University
Wuxi 214062, Jiangsu Province, China
e-mail: lurui.sytu@gmail.com

Li LIU [1,2*]

1. School of IoT Engineering
Jiangnan University
2. Wuxi Fourth People's Hospital
Wuxi 214062, Jiangsu Province, China
e-mail: liuli_sytu@hotmail.com

Yuting CHEN [1]

1. School of IoT Engineering
Jiangnan University
Wuxi 214062, Jiangsu Province, China
e-mail: chenyuting871226@163.com

***Abstract*—artificial immune networks (AIN) as one of the new intelligent soft computing methods have been widely used in many application fields. The AIN shows good ability of global optimization, especially in parameter optimization of pharmacokinetic models. The AIN search global optimum based on the principles of clone selection and immune network. However, as one of the heuristic-based optimal algorithms, the evolution of memory cells in the AIN is time consuming compared with gradient-based optimal algorithms. In this paper, a distributed AIN with distributed clone selection evolutionary strategy is proposed to improve the efficiency of the AIN. Then a distributed artificial immune network is implemented with MATLAB Distributed Computing Engine. One of the advantages of MDCE is that it is easy to run optimal algorithms programmed in MATLAB platform. In the experiments, parameters of the [^{18}F] Fluoro-2-deoxy2D-glucose (FDG) tracer kinetic model are optimized with the distributed AIN. Experimental efficiency of the algorithm is discussed.**

Keywords-artificial immune network; tracer kinetic model; distributed computing; MDCE

I. INTRODUCTION

The artificial immune system [1] is a bio-inspired soft computing method. Artificial immune network (AIN) as a branch of artificial immune system -produces the interaction mechanism between network cells. The AIN algorithms have been successfully applied in multi-modal function optimization [2] and dynamic environment optimization [3]. In our previous work, a PKAIN algorithm has been proposed to optimize parameters of both linear and nonlinear pharmacokinetic compartmental models [14]. Compared with other artificial immune algorithms, the PKAIN is based on clonal selection evolution [1] and the simplex updates of the memory cells to implement optimization. However, the clonal selection evolution is one of the most time-consuming steps in each iterative updates of the PKAIN generations. In order to obtain global optimizations using the PKAIN algorithm, much more memory cells should be generated in the network. As a result, it costs much more time compared with those gradient-based optimal algorithms. In this paper, a distributed PKAIN using a distributed clonal selection evolutionary strategy is proposed to save working time for optimizing parameters of the compartment model function. The tracer kinetics model is to be done as an example of the distributed PKAIN to evaluate the efficiency of the distributed strategy analysis. The MATLAB Distributed Computing Engine (MDCE) is setting as an experimental platform. One of advantages of the MDCE is that it integrates both simple distributed function interfaces and the MATLAB programming functions. Experimental results show that the distributed PKAIN with MDCE can improve the efficiency of the artificial immune network. It keeps the accuracy and reduces time consuming when using for parameters optimization.

II. TRACER KINETIC MODELING

Tracer kinetic modeling with positron emission tomography (PET) requires measurements of the time activity curves in both plasma and tissue to estimate physiological parameters [4]. Dynamic PET images can be evaluated visually or quantified by tracer kinetic modeling, which uses a tracer plasma time-activity curve (PTAC) as an input function in order to characterize the target tissue time-activity curve (TTAC) for a predefined region of interest (ROI). This characterization is achieved using physiological parameters [5]. Compartmental model is the most commonly used model to describe the uptake and clearance of radioactive tracers in tissue [6]. In general, the body is considered to be a system, and the distribution kinetics of the tracer can be described as many compartments, which refer to organs or tissues where the rates of absorption and transportation are similar. So, an appropriate model to fit PTAC and TTAC

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.2

are most important. Firstly, a mathematical tracer kinetic model is selected based on the principle of pharmacokinetics. Then parameters of the model are optimized to fit the TTAC, PTAC curve. Currently fluorine generation of deoxidizing glucose (FDG) kinetic model [7] is relatively commonly used, as shown in Fig.1. The three compartments [15] four parameters model which is originally proposed by Sokoloff et al is used to describe metabolic process of FDG [4] . At first the FDG get into the tissue cells through the cell membrane, and then they are transformed into a FDG-6-P by sugar kinesis in cells, eventually captured by cells. Conversely, the reverse process is also feasible in the model. There are four parameters (called rate constants) $k_1 \sim k_4$ in the compartment model and they represent the tracer transport rate between compartments. The rate constant k_1 is measured in ml (g min)$^{-1}$, and k_2 , k_3 and k_4 are measured in min^{-1}. C_B , C_E and C_M represent FDG concentration in the plasma, FDG concentration in the tissue, and FDG-6-phosphate (FDG-6-P) concentration in the tissue, respectively. In addition, in tracer dynamics research this model has another parameter f , it represents the image in the PET that the influenced coefficients by the radioactivity in the plasma of the surrounding tissue imaging factor.

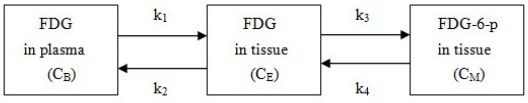


Figure 1. The FDG tracer kinetic model

The actual tissue activity measured by PET can be expressed as

$$C_T(t) = C_i(t) + f \cdot C_B(t) \quad (1)$$

Where $C_i(t) = C_E(t) + C_M(t)$, $C_T(t)$ is the function of tissue tracer time-activity curve (TTAC). Through calculate and optimize this model, we can get the following equation:

$$C_T(t) = k_1/(\alpha_2 - \alpha_1).[(k_3 + k_4 - \alpha_1).e^{-\alpha_1 t} + (\alpha_2 - k_3 - k_4).e^{-\alpha_2 t})] \otimes C_B(t) + f.C_B(t) \quad (2)$$

The $\otimes$ represents convolution, and

$$\alpha_1 = (k_2 + k_3 + k_4 - \sqrt{(k_2 + k_3 + k_4)^2 - 4k_2k_4})/2 \quad (3)$$

$$\alpha_2 = (k_2 + k_3 + k_4 + \sqrt{(k_2 + k_3 + k_4)^2 - 4k_2k_4})/2 \quad (4)$$

PTAC curve function $C_B(t)$ and TTAC curve function $C_T(t)$ are indicated as input function and output function respectively.

III. PKAIN: An Artificial Immune Network for Pharmacokinetics

In our previous work, an artificial immune network for parameter optimization of pharmacokinetics (PKAIN) is proposed and tested in several models [8]. The evolution of the PKAIN artificial immune network is described as follows [14]. In the first step, the artificial immune network is initialized by randomly generating network cells in solution space. Given a tracer kinetic model, parameter solution spaces such as low-bound and upper-bound of output parameters are defined. To solve FDG kinetic model which has been mentioned in Section II, parameters of $k_1 \sim k_4$ and f are required to be optimized simultaneously. One of the solutions of these parameters is encoded into a memory cell of the artificial immune network. For each cell m of $M^{(n)}$, clone selection with concurrent simplex mutation [9] step is done to generate new M^* population. Then m cell is replaced by the cell who is calculated with the highest affinity in the M^*. In the network suppression process, similar cells with lower affinity are deleted to maintain a relatively smaller network scale. The new generation $M^{(n+1)}$ of memory cells is generated. If the stop criterion is met, the evolution of memory cells stops. Finally, the memory cell with the highest affinity is decoded into an optimal parameter solution. Details of the PKAIN algorithm can be referred in the literature [10].

IV. Distributed Artificial Immune Network Based on Distributed Clone Selection with Concurrent Simplex Mutation

The main aim of the distributed PKAIN is to propose an efficient method to accelerate PKAIN algorithm executing speed. Recently, a parallelized artificial immune network for fuzzy clustering (PAINFCM) is designed and applied to fuzzy clustering [11]. In the PAINFCM algorithm [12], there were two models presented to improve efficiency of AIN for fuzzy clustering. One was the parallelized affinity calculation of the mutated antibodies. Based on the master-slave model, the AIN evolved in the master processor, and distributed affinity calculation of antibodies to slaves' processors. The other one is coarse-grained version of PAINFCM algorithm which was proposed to parallelize clone expansion.

For a traditional AIN, the clonal selection process of an artificial immune network is a computational implementation of the clonal selection principle [2] to solve optimization problems, emphasizing multimodal and combinatorial optimization [13]. The process of clonal selection is composed of clone expansion, affinity mutation and selection steps. Firstly, memory cells are sorted in their affinity values decadently. In the second step, the clone number of each memory cell is inversely proportional to its order, that is, memory cells with higher affinity will have more offspring. After clone expansion, antibodies of population undergo affinity maturation.

The Experiments show that the artificial immune network optimization of data analysis, with the increase in

the size of data processing algorithms to run slowly [1]. By the time complexity of algorithm analysis found, the algorithm of memory cell clones large population affinity calculation determined the rate of change in population variability situation is the main reason for the speed of algorithm execution. Therefore, this line of clone selection evolution, we can use independent methods of distributed computing will affect the time course of distribution to the various computing units, and each computing unit takes part of clonal selection tasks in the AIN, so that the distribution of the process of cloning can be completed faster.

For the PKAIN, it has a new clonal evolution called clonal selection with concurrent simplex mutation. There is a partition-based concurrent simplex method to be designed. That is after affinity mutation of traditional clonal selection process, antibodies are considered as a natural partition group to do simplex mutation. The number of cells of X is denoted as Nc , $Nc > L+1$. After executing concurrent simplex to X , there are $Nc - L$ number of new cells that have been updated. These new cells together compose the new generation of cells X . Then, the cell with the highest affinity is selected to substitute for the network cell m .

For the distributed PKAIN strategy, the only constraint of mutation is that mutation rate α is proportional to normalized affinity value. To solve this problem, normalized affinity should be distributed to processors. So, a new distributed PKAIN method with distributed clonal selection with concurrent simplex mutation is proposed. The procedure of the PKAIN with distributed clonal selection with concurrent simplex mutation is described by the following pseudocodes.

Initialize artificial immune network $M^{(0)}$.
While stopping criterion is not met do
Affinity calculation of each cell m in $M^{(n)}$, where n is the iteration number

Distribute p = 1…np $M^{(n)}(p)$ to processors

For each $m \in M^{(n)}(p)$ C= *clone_expension(m,*N_C*)*;

C^a = *affinity_mutation(*C);

C^s= *simplex_mutation*(C^a);

m^* = *selection(m,* C^s*)*;
END-Distribute and receive *M*(p)*
$M = M^{(n)} \cup M^*$;
$M^{(n+1)}$= *network Suppression-Update(M)*;
END Do
Decode the m with the highest affinity into the optimal solution

V. EXPERIMENTS AND RESULTS

A. Data acquisition

Dynamic FDG studies were performed on four mice subjects. Each one have been recorded their sex and weigh. The experiment injected these mice with ^{18}F-FDG which performed at Siemens Invent Micro-the PET nuclear medicine for molecular imaging equipment in a period of time. For the purpose of this study, the data acquisition soft called Invent Acquisition workplace was implemented to dispose images from PET. The regions of interest (ROIs) selected for modeling with the proposed method were based on visual identification of their corresponding TAC.

B. MATLAB Distributed Computing Engine

MATLAB Distributed Computing Environment is a software platform in this study: one Intel Xeon server platform, four Lenovo Intel (R) Core (TM) 2.0 computers and network composition. Development software is the MATLAB R2009a, the MATLAB Distributed Computing Server 4.3, and the MATLAB Distributed Computing Toolbox 4.3. The MDCE is running with MATLAB distributed server and its toolbox program interface. It starts after the software installation services of distributed computing, programming for the distributed environment. The MDCE and distributed computing toolbox can be used to run on cluster computers for distributed and parallel MATLAB applications. Distributed algorithm tasks, including separate tasks, have no communication between them. They integrate the technical standards-based MPI functions in the MATLAB environment to support development of distributed applications. They can also create a distribution array of applications, support FOR loops and global array. This application does not require message passing, distributed computing or parallel through the implementation of the algorithm can improve the efficiency and save system development time. The client-job manager-workers architecture of the distributed computing configuration is described as follows.

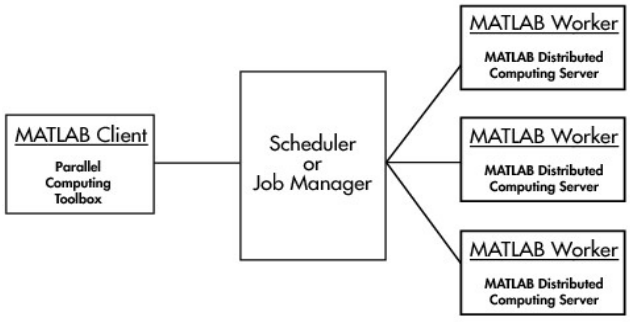


Figure 2. The client-job manager-workers architecture of the MDCE

According to the Fig.2, the MDCE enables us to coordinate and execute independent MATLAB operations simultaneously on a cluster of computers (various computing units), speeding up the execution of large MATLAB jobs. A job is some large operation that you need to perform in your MATLAB session. The MATLAB session in which the job and its tasks are defined is called the client session. Often, this is on the machine where you program MATLAB. The client uses MDCE server to

perform the definition of jobs and tasks. The job manager is the part of the server software that coordinates the execution of jobs and the evaluation of their tasks. The job manager distributes the tasks for evaluation to the server's individual MATLAB sessions called workers.

In the experiment, we use a master-slave distributed model. It is a model of coarse-grained parallel computing model. Master-slave model sets to host management of distributed PKAIN network encoding and cell population regeneration steps. As in the Fig.2, the client is the host management through the job-manager which divides the distribution of tasks to each worker. Then the slave workers take the responsibility for the clone selection and affinity calculation process. The job-manager is the MDCE scheduling program and deployed in distributed computing server. The distributed PKAIN clone selection evolution is the process of independent of sub-populations. Therefore, the end of the distribution of each sub-population of the affinity calculation results are brought together, MDCE provides a computing interface (PAFOR), can be achieved through distributed PKAIN clone selection strategy.

C. *Speed ratio analysis*

In order to evaluate the performance of distributed PKAIN algorithm, using Amdahl law namely absolute speed speedup ratio (index) is analyzed. Acceleration as formula calculation Eq.(5), where $T(1)$ means a processor serial algorithm in the execution time, $T(p)$ means p processors serial algorithm in the execution time.

$$SP(p) = \frac{T(1)}{T(p)} \tag{5}$$

For the cell clone selection constant impacts the speedup of the distributed PKAIN; the number of cloned cells of each memory cell is set as 50,100,200,400,800 respectively, for comparative analysis. Initialize artificial immune network parameters with the initial population size N = 40, mutation rate beta = 5, each iteration randomly generate new cell number d = 0.4, the number of iterations gen = 100, network suppression parameter ts = 0.1.

TABLE I. SPEEDUP OF DISTRIBUTED PKAIN WITH DIFFERENT THE NUMBER OF CLONED CELLS

Nc	CPU(processors)				
	P=2	***P=4***	***P=6***	***P=8***	***P=10***
50	1.08	2.01	4.46	5.46	6.14
100	1.91	3.45	4.36	5.01	5.90
200	1.77	3.36	4.29	5.80	6.64
400	1.68	3.34	5.62	6.92	7.00
800	2.24	3.13	5.48	6.83	7.67

Notes: Nc is the number of cloned cells, and p is the number of processors.

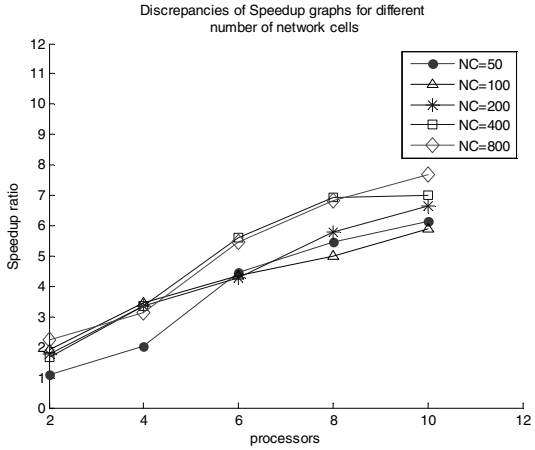


Figure 3. Discrepancies of Speedup graphs for different number of network clone cells

According to the above table and image, memory cells population become to influence the speedup ratio as an important factor with its numbers increased. As can be seen from Table I, this table gives to quantify the cloned population number compared with the increase of population and processor computing unit, the running time speedup goes up, and ultimately the number of species in 800 is that its running time is running single-core the 1/7.67 times. At this time population has reached a relatively high speed of operation (Fig.3). However, the distributed process as only the distribution of clone selection does not affect the accuracy of parameter optimization algorithms, so it can be shown to achieve the purposes that improved parameter optimization of tracer kinetics model to avoid the time-consuming.

VI. CONCLUSIONS

In this paper, the PKAIN method is to be accelerated to solve tracer kinetics modeling in a distributed way. The method to acquire distributed solutions is described in details and be implemented to optimization parameters for the compartment model of FDG tracer. While the accuracy results of the distributed algorithm kept, it can obtain higher speed. As the clone selection don't show any change of the algorithm efficiency. The distributed master-slave unit shows that distribution computing can save much more time in solving tracer kinetics problems. In addition, this algorithm also suitable for problems of other tracer kinetic models.

ACKNOWLEDGMENT

This research is supported by the Fundamental Research Funds for the Central Universities No.JUSRP10928 and Science and Technology Support of Wuxi City No. CSE01014.

REFERENCES

[1] de Castro L N. , Timmis J. Artificial immune systems as a novel soft computing paradigm[J]. Soft Computing, 2003, 7: 526–544.

[2] Ada G L, Nossal G J V. The Clonal Selection Theory [J]. Scientific American, 1987, 257:50-57.

[3] de Franca, F.O., Von Zuben, F.J., de Castro, L.N., An artificial immune network for multimodal function optimization on dynamic environments. Proc. of the GECCO conf, ACM Press, Washington DC, pp. 289-296, 2005.

[4] Dagan Feng, Xianjin Li, Sung-Cheng Huang. A New Double Modeling Approach for Dynamic Cardiac PET Studies Using Noise and Spillover Contaminated LV. IEEE Transactions on Biomedical Engineering,1996,43(3):319-326

[5] Chun, J S, Hahn, S.Y.. A study on comparison of optimization performances between immune algorithm and other heuristic algorithms[J]. IEEE Transactions on Magnetics, 1998,34: 2972-2975.

[6] Anderson, D. H. (1983). Compartmental modeling and tracer kinetics. Berlin ; New York, Springer-Verlag.

[7] Zui Y F,Bei J. Kinetic model parameter estimates of liver FDG metabolism[J]. Journal of Zhejiang University of Science and Technology,2007,47(12):4-10.

[8] Wang G J. Pharmacokinetics. chemical industry Press,2005:88-120.(in Chinese).

[9] Jerne N K. Towards a network theory of the immune system[J]. Annals of Immunology, 1974, 125C: 373-389.

[10] Ada G L, Nossal G J V. The Clonal Selection Theory [J]. Scientific American, 1987, 257:50-57.

[11] Jerne N K. Towards a network theory of the immune system [J]. Annals of Immunology, 1974, 125C: 373-389.

[12] Li Liu, Wenbo Xu.A parallelized artificial immune network for fuzzy clustering. International Journal of Computer Mathematics, 2009,87(6):1401- 1414.

[13] de Castro L N, Von Zuben F J. Learning and Optimization Using the Clonal Selection Principle. IEEE Trans. On Evol. Comp.,2002, 6(3):239-251.

[14] Liu L, Zhou SD, Lu HW, Xie F, Xu WB. Parameter optimization of pharmacokinetics based on artificial immune network[J]. Applied Mathematics and Mechanics, 2008,4(59): 549-558.

[15] Bertoldo A , V icini P, Sambuceti G, et al. Evaluation of compartmental and spectral analysis models of [18F]FDG kinetics for heart and brain studies w ith PET [J]. B iom ed ical Engineering, IEEE T ransactions on, 1998, 45 (12) :1429-1448.

Integration of 1D and 3D Simulations of Engine Cooling System: After Keyed- Off

Pang, S.C.
Department of Mechanical Engineering
University of Malaya
Kuala Lumpur, Malaysia
psuhchyn@yahoo.com

Masjuki, H. H.
Department of Mechanical Engineering
University of Malaya
Kuala Lumpur, Malaysia
masjuki@um.edu.my

Hazrat, M.A.
Department of Mechanical Engineering
University of Malaya
Kuala Lumpur, Malaysia
alihazrat20@yahoo.com

Kalam, M.A.
Department of Mechanical Engineering
University of Malaya
Kuala Lumpur, Malaysia
kalam@um.edu.my

***Abstract*: Engine cooling system is crucial to maintain working temperature of vehicles' engine. The two main heat transfer fluids in typical cooling system are coolant and air. Under hood air flow when approaching radiator is highly non- uniform, thus 3D computational fluid dynamics is required to estimate air mass flow rates. While coolant circuit consists of numerous components, 1D thermal- fluid simulation is utilized to study temperature behavior of the coolant. Automotive designers should design a robust engine cooling system which is working well in both normal and severe driving conditions. When vehicles are keyed- off suddenly after some distance of driving, the coolant temperature tend to increase dramatically. This is because soaking heat at engine could not be transferred away timely, as water pump and cooling fan have stop working after keyed- off. In this study, transient coolant temperatures are observed closely with variety patterns of soaking heat. It is suggested to prolong operation of water pump and fan for a short while after keyed- off to reduce transient coolant temperature.**

Keywords- Basic Engine Cooling System, Transient, Thermal Soak, Keyed- off.

1. Introduction

Engine cooling system is very crucial for a vehicle as it ensures engine always running at its optimum operating temperature. In water cooling system, coolant flows in water jacket and transfers heat away from cylinder head and body. At heat exchanger/ radiator, hot coolant will transfer the heat to cooling air. A complete numerical study of engine cooling system consists of 2 portions, air side and coolant side.

For air side, computational fluid dynamics (CFD) is required to explore geometry effect of under hood components towards cooling air flow. Basically, there are two main energy sources which drive the flow of cooling air; one is ram air for high speed driving vehicle, another one is cooling fan for low speed driving vehicle. The cooling air path affects air supply to radiator thus its cooling capacity. The cooling air path consists of grille, bumper, condenser, radiator, fan, shroud and engine body. Air which passes through more flow resistances will have a lower momentum when approaching radiator frontal surface. This is also true for air passes through more heat source, where they will have higher thermal energy approaching radiator frontal area. Hot air recirculation from engine body and exhaust components, could affect radiator thermal performance adversely.

From the coolant side, one dimensional thermo-fluid simulation includes every cooling component inside coolant circuit modeling. It mimics many electronics components inside an electronics circuit. This allows us to study and to understand how the overall system is working. For instance, transient coolant peak temperature is simulated and studied in this paper. In order to design a robust engine cooling system, system behavior under extreme driving condition is always a main concern. One of the examples is keyed off the engine after maximum speed driving or hill climb driving. Sudden keyed- off of engine, will cause coolant temperature to increase significantly as water pump and cooling fan stop working. When temperature of coolant fluctuates with time, this is so called un- steady or transient temperature behavior. The desirable coolant temperature is below 95˚C as high or boiling coolant temperature may shorten the life span of an engine and other components in the coolant circuit.

Zyl [1] developed a CFD simulation using StarCD, for the cooling air flow. The numerical model was validated with experiment result at wind tunnel. Ecer [2] calculated the velocity over the radiator and air- to boil temperature by CFD using PASSAGE. Franchetta [3] investigated a computational procedure and implemented it in VECTIS that delivers accurate transient predictions with substantially

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.53

reduced computational time. Lawrence [4] improved under hood air flow by CFD. While many researchers utilized CFD to model under hood air flow, Kumar [5] combined CFD and Flow Network Modeling to achieve the same objective. For thermo- fluid simulation, Ning [6] utilized AMESim to perform system analysis for engine cooling system. Eichlseder [7] simulated both cooling air and coolant circuit using KULI. For coupling of CFD and thermo- fluid simulation, Rok [8] successfully integrated StarCD and Flowmaster. Kim [9] coupled Fluent, StarCD and KULI for automotive air conditioning analysis. In this research, air mass flow through radiator is fed from CFD into 1D thermo- fluid simulation of coolant circuit. Finally, transient coolant temperature after keyed- off could be analyzed in 1D thermo- fluid model.

Abbreviations and Acronyms

C: Porous viscous resistance [kg/m^3 s]
D: Porous inertial resistance [kg/m^4]
F: Corrections factor
V: Velocity [m/s]
ΔP: Pressure drop [Pa]
L: Length [m]
Re: Reynolds number
D_H: Hydraulic diameter, coolant and air side [m]
A: Reference area, for coolant and air side [m^2]
m: Mass flow rates, for coolant and air side [kg/s]
μ: Dynamic viscosity [kg/ m s]
U: Overall heat transfer coefficient [W/ m^2 K]
x: Radiator thickness [m]
K: Thermal conductivity for coolant and air [W/ m K]
C_p: Specific heat, for coolant and air [J/ kg K]
ε: Radiator efficiency [%]

2. Numerical Modeling

a) 3D CFD Model

CFD is a powerful simulation tool to model fluid flow and fluid conjugate heat transfer. In this paper, Star-CCM+ was utilized to model under hood air flow under different driving conditions. The main steps involved in a CFD model establishment were geometry import, surface meshing, volume meshing and physics modeling. Firstly, a full- scale half- car- body geometry drawing was imported into Star- CCM+ [11]. Secondly, mesh size definition, feature curve definition, multiple regions definition and boundaries definition at surface meshing stage. Error- free and quality- promising surface mesh is important before proceeding to volume meshing. Thirdly, polyhedral and prism- layer mesh type were selected in volume meshing. After volume meshing, cell with premium quality should be assured (minimum face validity<0.8, negative volume cells and etc.). A good quality volume mesh will lead to a better accuracy and a converged solution later. Lastly, physics modeling allows us to define initial conditions and boundaries condition, to create interfaces between regions, to define regions type and to select physics continuum. In physic continuum selection, we decide viscous scheme (laminar or turbulent), flow type (segregated or coupled), turbulence model, space, time and equation- of- state.

Table 1 summarized multiple regions which are established and their significant modeling features. Porous media method was used to define porosity of heat exchangers. Viscous resistance and inertial resistance were derived from characteristic chart of pressure drop over length and velocity, as in equation (1) [11], with linear regression. Radiator cores were modeled as dual stream heat exchanger where two types of fluids exchange heat with each other. Condenser was modeled as a single stream heat exchanger where only one type of fluid is participating. Fan was modeled as a rotating reference frame, as actual fan blade drawing was available. As fan rotates, it can provide momentum to under hood air flow, especially during low speed driving. Heat exchangers' temperature (condenser and radiator) and solid which provides constant heat flux (engine body and exhaust pipe) will affect under hood thermal environment.

$$\Delta P/ L = CV + DV^2 \qquad (1)$$

TABLE 1. CFD PHYSICS MODELING SUMMARY

Regions	Region Type	Physics Continuum	Modeling Feature
Free Stream	Fluid	Air	Ram Air Inlet
Condenser	Porous	Air	Single Stream Heat Exchanger
Radiator Core (Air)	Porous	Air	Dual Stream Heat Exchanger
Radiator Core (Coolant)	Porous	Water	Dual Stream Heat Exchanger
Radiator Upper Tank	Fluid	Water	-
Radiator Lower Tank	Fluid	Water	-
Engine Body	Solid	Solids	Constant Heat Flux
Fan	Fluid	Air	Moving Reference frame

b) 1D Thermo- fluid Model

Flowmaster [10], 1D thermo- fluid software, was used for coolant circuit modeling. A simplified coolant circuit consists of engine mass, engine heat source, thermostat, radiator, reservoir expansion tank, water pump, pump speed control, pipes and air flow source were built. The model was used to study individual component's effect onto whole engine cooling system. The software allows us to include numerous components into a circuit and to link them together (similar to an electronic circuit with different electronic components). Then, it is required to set parameters of each and every component in details. For instance, for radiator, the parameters are coolant flow area, coolant hydraulic diameter, airside flow area, airside hydraulic diameter, curve of coolant side pressure drop vs. flow rates, curve of airside pressure drop vs. flow rates and etc. For radiator, surface of heat transfer vs. air flow vs. coolant flow was available at a determined inlet temperature

difference. The heat transfer surface was normalized and converted to Nusselt surface (Nu vs. Re (air) vs. Re (coolant)), by applying Equations 2-10 [10]. Equation 2 and 3 normalized coolant flow and air flow to their respective Reynolds number. Equation 4 and 5 related heat transferred with fluid flow and fluid temperatures. As heat transferred was known and fluid temperatures were computed with equations (4) and (5), equation (6) can be used to calculate their overall heat transfer coefficient. Then, with equation (7), overall heat transfer coefficient was normalized to Nusselt number. Equation (8) and (9) were used to calculate effectiveness of heat exchanger.

From the CFD under hood air flow model, we can obtain radiator air flow rates at different driving speed, at different fan rotation speed or both. By coupling of CFD air flow input and industry data of engine heat and coolant flow rates at different driving speed and conditions, transient coolant temperature (especially after keyed- off) was studied in 1D thermo-fluid simulation. As it was a transient simulation, heat flow, coolant flow and air flow varied with time horizon. It can be observed how these factors are interacting with each other and affecting coolant temperature (time behavior). When the air flow rates and coolant flow rates are at their minimum, it is defined as keyed- off time. Keyed- off time is fixed at 4500s. However, different patterns of soaking heat flow at engine were tested to observe their corresponding time behavior coolant temperature, Table 2.

$$Re_{coolant} = \left(\frac{\dot{m}D_H}{\mu A}\right)_{coolant} \quad (2)$$

$$Re_{air} = \left(\frac{\dot{m}D_H}{\mu A}\right)_{air} \quad (3)$$

$$q = \dot{m}_c C_{p,c}\left(T_{c,in} - T_{c,out}\right) \quad (4)$$

$$q = \dot{m}_a C_{p,a}\left(T_{a,out} - T_{a,in}\right) \quad (5)$$

$$U = \frac{q}{FA * LMTD}, \qquad where$$

$$LMTD = \frac{[(T_{c,in} - T_{a,out}) - (T_{c,out} - T_{a,in})]}{ln\frac{(T_{c,in} - T_{a,out})}{(T_{c,out} - T_{a,in})}} \quad (6)$$

$$Nu = \frac{Ux}{\kappa_{air}} \quad (7)$$

$$q_{max} = (\dot{m}C_p)_{min}\left(T_{c,in} - T_{a,in}\right) \quad (8)$$

$$\varepsilon = \frac{q}{q_{max}} \quad (9)$$

TABLE 2. DIFFERENT SOAKING HEAT PATTERN

Test Scenarios	Soaking Heat Pattern
Case 1	Maximum heat persisted until 4200s.
Case 2	Maximum heat persisted until 3900s.
Case 3	Maximum heat persisted until 3700s with heat tails.

3. Results and Discussions

From CFD model, we could visualize how the air flows through bumper, front end opening, grille, front- end- structure, condenser, radiator, fan, shroud and engine body. As condenser and radiator were modeled as porous medium with a defined flow resistance, air could flow through these heat exchangers. In fact, radiator's air flow rates were dependent on the layout of under hood components, the ram air flow rates and also the fan rotation speed. At high speed driving, ram air was the main momentum source for radiator air flow. At low speed driving, fan was the main momentum source for radiator air flow. Table 3 summarized radiator air flow under different driving conditions.

CFD model also showed that the bumper, front end opening were affecting ram air which entering under hood. When air entered under hood, it was blocked by the front- end mechanical structure. Other heat exchanger before radiator (in this case the condenser) was also affecting the air flow rates and air temperature which entering radiator. And, this would definitely affect the thermal performance of the radiator. Shroud (component behind radiator) helps fan to pull more air through radiator. Hot air recirculation from hot engine surface and hot exhaust surface could affect radiator's performance. Figure 1(a) showed air velocity contour approaching the surface of condenser and radiator (driving at 110km/hr). Air velocity was lowest at the center of condenser (cross shape), as it was blocked by bumper and front- end- structure. Figure 1(b) was side view of under hood air flow, it visualized how ram air entering under hood via the front end opening. It also displayed how air escaped from fan shroud before approaching engine body.

TABLE 3. CFD RADIATOR AIR FLOW UNDER DIFFERENT DRIVING CONDITIONS

Ram Air [km/hr]	Fan [rpm]	Radiator Air Flow [m/s]	Radiator Air Flow [kg/s]
3.6	2000	0.8496	0.2806
50	50	0.8867	0.2928
50	2000	1.4231	0.4700
110	50	2.3931	0.7904
160	50	3.6850	1.2170

In 1D thermo- fluid simulation, simulation time horizon was 7000s. Normal driving period occurred during time 0- 2000s. Hill climbing period occurred during 2000- 3500s. Slow down period occurred during time 3500- 4500s. Keyed- off periods occurred after time 4500s, Figure 2(a). As it was mentioned, industry data of engine heat and coolant flow under different engine speed were available. However, soaking heat data after keyed- off was not available at this moment. So, different soaking heat was presumed to observe their corresponding transient coolant temperature behavior. In three simulation scenarios, common keyed-off time was 4500s. For Case 1, the maximum heat persisted until 4200s. For Case 2, the maximum heat persisted until 3900s. While for Case 3, the maximum heat persisted until 3700s with a little amount of soaking heat which was reduced linearly during 4500- 7000s (heat tails). Figure 2(a), 2(c) and 2(e) showed engine heat, coolant flow rates and air flow rates varied over time. While figure 2(b), 2(d) and 2(e) showed respective coolant temperature at radiator upper hose and radiator lower hose.

For Case 1, the soaking heat was the largest in quantity. In figure 2(b), the coolant temperature started to

increase from 0s before thermostat opened. Thermostat opened at 1100s and coolant temperature remained stable until 3500s. At 3500s, coolant flow and air flow started to slow down and coolant temperature increased to 140˚C (huge amount of heat cannot be dissipated on time). For Case 2, with intermediate soaking heat, coolant peak temperature was 115 ˚C. While for Case 3, with mild soaking heat, the coolant peak temperature was approximate 100 ˚C. As there was a "heat tail" in Case 3, coolant temperature remained above 80 ˚C until 6000s. Figure 3(a) showed the radiator effectiveness over time for Case 3. Air temperature rise and inlet temperature difference was also plotted, as effectiveness is a ratio of them (Eq. 6, 9 and 10). Lastly, figure 3(b) was the thermostat opening ratio over time for case 3.

From observation of all 3 simulation cases, we have learned that time lagging between soaking heat and fluids flow (pump speed and fan speed) would decide the coolant temperature pattern. The larger time lagged between them (heat lagged behind air flow and coolant flow), the higher the coolant peak temperature was. It has been observed that, we could reduce the time lagging between soaking heat and fluids flow by keeping fan and pump running for a couple more minutes after keyed- off. This objective could be met by installing electrical fan and electrical water pump. Meanwhile, we are brain- storming and testing other innovative and novel alternatives to minimize coolant peak temperature in 1D thermo- fluid simulation.

4. Conclusions

Cooling air flow was one of the crucial factors which affecting the radiator thermal performance. CFD allowed us to obtain dynamics air flow rates when approaching radiator. Improvement of air flow after re-design of bumper, front- end- opening and under hood layout could be visualized in CFD. Then, by inputting CFD air flow data into 1D thermo –fluid simulation, it allowed us to study transient coolant temperature after keyed- off. Huge soaking heat after keyed- off could damage engine when coolants temperature was above its boiling point.

Re- design of coolant circuit in 1D thermo- fluid simulation was a good solution at beginning stage of engine cooling system design. In this research, it has been observed that, we could reduce the time lagging between soaking heat and fluids flow by keeping fan and pump running for a couple more minutes after vehicle keyed- off. Finally, integration simulation of dynamic 3D CFD and simplified 1D thermo- fluid models would yield us benefits from both 3D and 1D simulation.

ACKNOWLEDGMENT

I would like to express my appreciation to Proton Holdings Berhad, which is a Malaysia's leading automotive manufacturer. The precious company supported us in this research financially and technically.

REFERENCES

[1] J. M. V. Zyl (2006), "Numerical Modeling and Experiment Investigation of The Flow and Thermal Processes in A Motor Car Vehicle Under Hood", Master Thesis at Stellenbosch University.

[2] A. Ecer and C. Toksoy (1995), "Air Flow and Heat Transfer Analysis of an Automotive Engine Radiator to Calculate Air- to- Boil Temperature", Journal of SAE International, Paper Number: 951015.

[3] M. Franchetta and T.G. Bancroft (2006), "Fast Transient Simulation of Vehicle Underhood in Heat Soak", Journal of SAE International, 2006-01-1606.

[4] V. Lawrence (2001), "Underhood Airflow Simulation of A Passenger Car Using Computational Fluid Dynamics", Journal of SAE International, 2001-01-3800.

[5] V. Kumar, S. Kapoor and G. Arora (2009), "A Combined CFD and Flow Network Modeling Approach for Vehicle Underhood Air Flow and Thermal Analysis", Journal of SAE International, 2009-01-1150.

[6] G. Ning (2009), "Simulation of Engine Cooling System Based on AMESim", 2nd International Conference on Information and Computing Science, Manchester, United Kingdom.

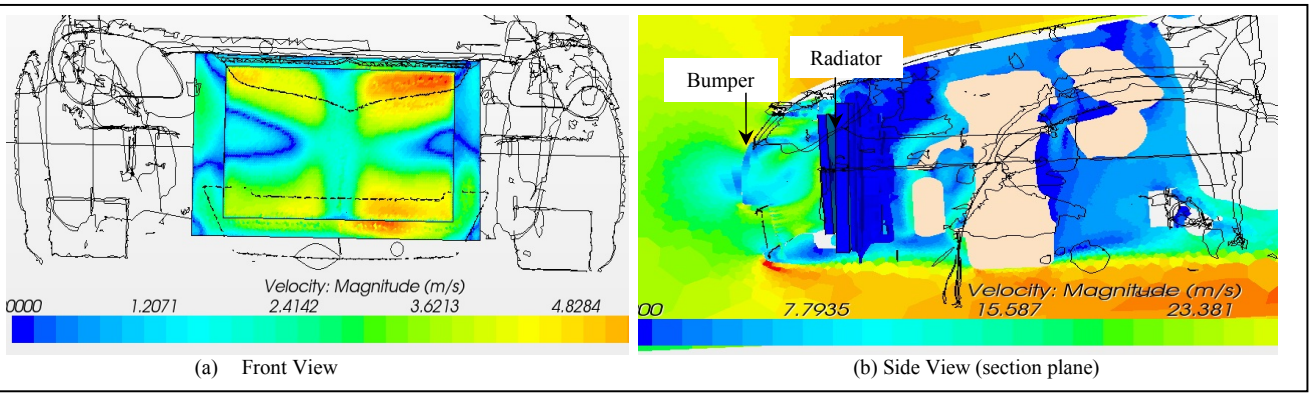
[7] W. Eichlseder and G. Raab (1997), "Use of Simulation Tools With Integrated Coolant Flow Analysis for the Cooling System Design, Journal of SAE International, Paper Number: 971815.

[8] S. H. Rok, H. Pastohr (2007), "A Real Time Numerical Analysis of Vehicle Cool- down Performance", Hyundai Motor Company.

[9] H. J. Kim and C.J. Kim (2007), "A Numerical Analysis for The Cooling Module Related to Automobile Air- conditioning System", Journal of Applied Thermal Engineering, 28, 1896-1905.

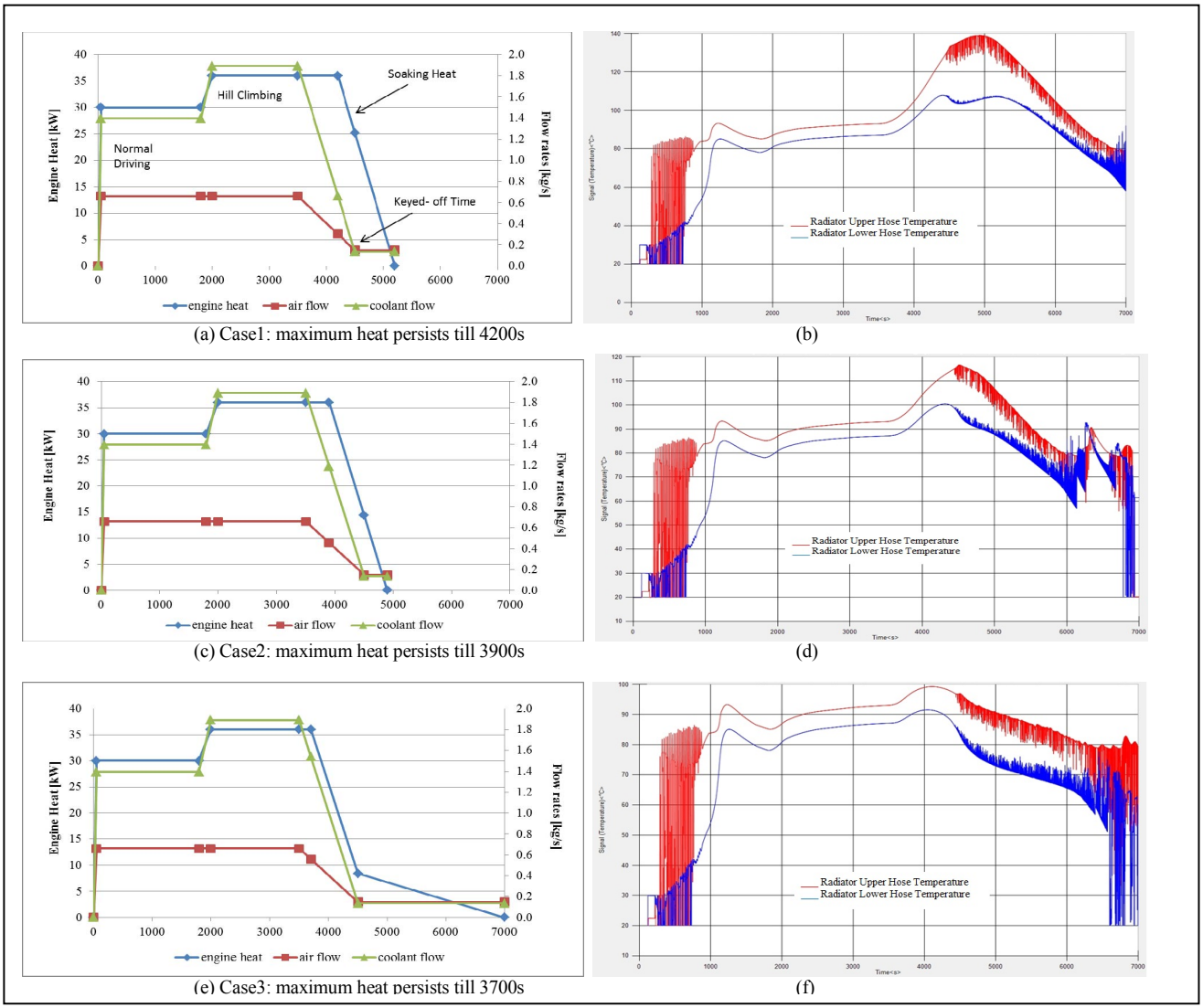
[10] Flowmaster V7 Manual (2010), Flowmaster Ltd.

[11] Star CCM+ User Guideline (2010), CD- adapco.



(a) Front View (b) Side View (section plane)

FIGURE 1. F UNDER HOOD AIR FLOW (FRONT AND SIDE VIEW).

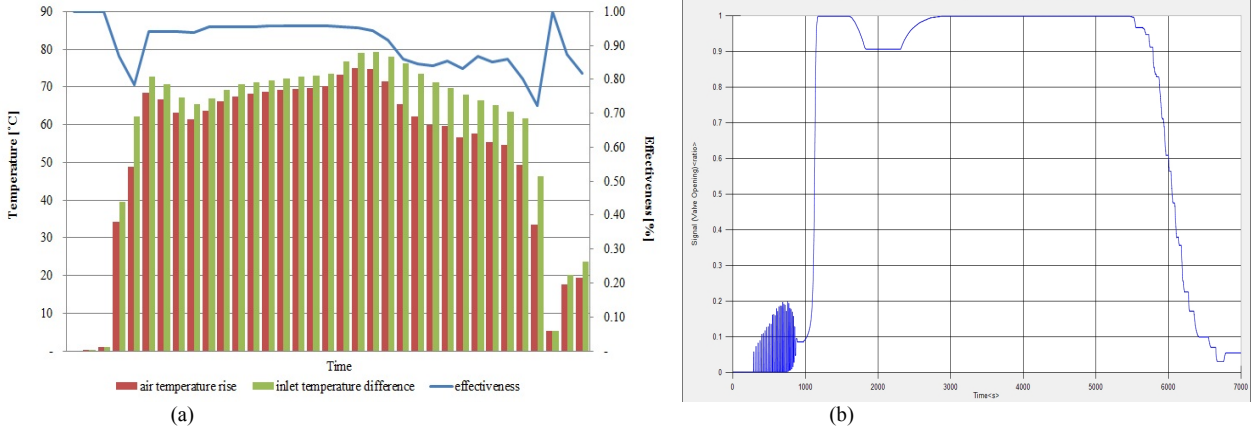


(a) Case1: maximum heat persists till 4200s (b)

(c) Case2: maximum heat persists till 3900s (d)

(e) Case3: maximum heat persists till 3700s (f)

FIGURE 2. TRANSIENT COOLANT TEMPERATURE WITH DIFFERENT HEAT SOAK PATTERNS



(a) (b)

FIGURE 3. HEAT EXCHANGER EFFECTIVENESS AND THERMOSTAT OPENING (FOR CASE 3).

Moment Problem of G-frames in Hilbert Spaces

Xiangyang Wang, Zhibiao Shu*
College of Mathematics and Computer Science
Fuzhou University
Fuzhou, Fujian350108, China
wxyang023@163.com, szb@fzu.edu.cn

***Abstract*—In this paper, we discuss the moment problem of g-frames in Hilbert spaces. Firstly, we give the definition of moment problem which is based on operators, and discuss the necessary and sufficient conditions of the existence of its solution. Secondly, we discuss its best approximative solution by orthogonal projection when it has no exact solution.**

***Keywords:* Moment problem; g-Riesz basis; Orthogonal projection; Best approximative solution**

I. INTRODUCTION

The moment problem was introduced first by Stieltjes in 1894 [1]. It can be defined as follows: Given a sequence $s_0, s_1 \ldots$ of real numbers. Find necessary and sufficient conditions for the existence of a measure μ on $[0,\infty)$ so that

$$s_n = \int_0^\infty x^n d\mu(x), for \quad n = 0,1,\ldots$$

Subsequently the moment problem was studied by many scholars and there were several variants of it such as Hamburger moment problem and Hausdorff moment problem [2, 3]. The moment problem has as its prototype the following classical trigonometric moment in $L^2[\pi,-\pi]$:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\phi(t)e^{-\mathrm{int}}dt = c_n \quad (n = 0,\pm1,\pm2,\cdots).$$

Young shows in its logical extension to an abstract Hilbert space, the aforement trigonometric moment problem takes the form

$$\langle f, f_n\rangle = c_n \quad (n = 1,2,3,\cdots) \qquad (1)$$

where $f_1, f_2, f_3, \cdots$ belong to H, $\{c_i\}_{i\in N}$ is an element of $l^2(N)$ .The moment problem has been applied to many fields such as irregular sampling [4]. Therefore how to find the solution of (1) is a practical and theoretical issue which is worthy of being studied. Young [5] proved that if $\{f_i\}_{i\in N}$ is complete, (1) has one solution at most and the solution is unique. O.Christensen [6] showed that (1) has a unique solution if and only if $\{f_i\}_{i\in N}$ is a Riesz basis of the Hilbert space H . Moreover he discussed the connection between the solution of (1) and the Riesz-Fischer sequence [5, 7]. If (1) has no exact solution, there is a best approximative solution $f = \sum_{j\in N} c_j S^{-1} f_j$ when $\{f_i\}_{i\in N}$ is a frame of the Hilbert space H . In 2005, Sun introduced a g-frame in a complex Hilbert space and discussed some properties of it [8, 9]. G-frames have some properties similar to those of frames, but not all the properties are similar.

In this paper, we consider the moment problem in Hilbert space for bounded linear operators, and discuss the existence of the solution of moment problem. In a g-frame we have a best approximative solution when the exact solution does not exist.

II. PRELIMINARIES

Throughout this paper, U and V are complex Hilbert spaces with inner product given by $\langle\cdot,\cdot\rangle$, and $\{V_j\}_{j\in J}$ is a sequence of closed subspaces of V , where J is a subset of integers Z . $L(U,V_j)$ is the collection of bounded linear operators from U into V_j . The operator Λ_j is in $L(U,V_j)$ for any $j\in J$ .The space $l^2(\{V_j\})$ is defined by

$$l^2(\{V_j\}) = \left\{\{f_j\}_{j\in J} : f_j \in V_j, j\in J \ \text{ and } \ \sum_{j\in J}\| f_j \|^2 < +\infty\right\},$$

with inner product given by

$$\langle\{f_j\}_{j\in J},\{g_j\}_{j\in J}\rangle = \sum_{j\in J}\langle f_j, g_j\rangle \quad \{f_j\}_{j\in J},\{g_j\}_{j\in J} \in l^2(\{V_j\}_{j\in J}).$$

Therefore $l^2(\{V_j\})$ is a complex Hilbert space.

Definition 2.1 [8] A sequence $\{\Lambda_j \in L(U,V_j)\}_{j\in J}$ is called a generalized frame or simply a g-frame for U with respect to $\{V_j\}_{j\in J}$, if there are two positive constants A and B such that

$$A\| f \|^2 \le \sum_{j\in J}\|\Lambda_j f\|^2 \le B\| f \|^2, \quad \forall f \in U.$$

The constants A and B are called the lower and upper frame bounds, respectively.

Definition 2.2 [10] A sequence $\{\Lambda_j \in L(U,V_j)\}_{j\in J}$ is called a g-Riesz-Fisher sequence, if there is a positive constant A such that

$$A\sum_{j\in I}\| g_j \|^2 \le \| \sum_{j\in I}\Lambda_j^* g_j \|^2, g_j \in V_j$$

for any finite subset I of J .

Definition 2.3 Let $\{\Lambda_j\}_{j\in J}$ be a sequence of $L(U,V_j)$, and $\{g_j\}_{j\in J}$ belongs to $l^2(\{V_j\}_{j\in J})$. It is natural to ask whether we can find a $f\in U$ such that

$$\Lambda_j f = g_j, j\in J \qquad (2)$$

This problem is called a moment problem.

*: Corresponding author.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.61

Lemma 2.4 [11] A sequence $\{\Lambda_j\}_{j\in J}$ is a g-Bessel sequence for U with respect to $\{V_j\}_{j\in J}$ with bound B if and only if

$$Q:\{g_j\}_{j\in J}\to\sum_{j\in J}\Lambda_j^* g_j$$

is a well defined bounded linear operator from $l^2(\{V_j\}_{j\in J})$ into U and $\|Q\|\le\sqrt{B}$. Furthermore, if Q^* is the adjoint operator of Q, then we have

$$Q^* f=\{\Lambda_j f\}_{j\in J},\forall f\in U.$$

Lemma 2.5 [11] A sequence $\{\Lambda_j\in L(U,V_j)\}_{j\in J}$ is a g-frame for U with respect to $\{V_j\}_{j\in J}$ if and only if

$$Q:\{g_j\}_{j\in J}\to\sum_{j\in J}\Lambda_j^* g_j$$

is a well defined bounded linear operator from $l^2(\{V_j\}_{j\in J})$ onto U, where the g-frame bounds are $\|Q^+\|^{-2}$, $\|Q\|^2$ and Q^+ is the pseudo-inverse of Q.

Lemma 2.6 [10] Let $\{\Lambda_j\}_{j\in J}$ be a sequence belonging to $L(U,V_j)$, and its analysis operator T is defined by:

$$T:\{f\mid f\in U,\sum_{j\in J}\|\Lambda_j f\|^2<+\infty\}\to l^2(\{V_j\}_{j\in J}),$$

$$Tf=\{\Lambda_j f\}_{j\in J}.$$

If $\{\Lambda_j\}_{j\in J}$ is g-complete, then $\{\Lambda_j\}_{j\in J}$ is a g-Riesz-Fischer sequence for U with respect to $\{V_j\}_{j\in J}$ if and only if T is surjective.

Lemma 2.7 [12] Let M be a nonempty closed subspace of a Hilbert space H. To each vector $u_0\in H$, there corresponds a vector $v_0\in M$ such that $d(u_0,M)=\|u_0-v_0\|$. Moreover, we have $u_0-v_0\in M^\perp$ and v_0 is the only vector satisfying this equality.

Lemma 2.8 [12] Let M be a nonempty closed subspace of a Hilbert space H. If the linear operator $\prod:H\to H$ is an orthogonal projection such that $R(\prod)=M$, then we have $\|u-\prod u\|=\inf_{v\in M}\|u-v\|$ for all $u\in H$.

III. Main Results and Their Proofs

Firstly, if (2) has a solution, the following theorem will give the necessary and sufficient conditions of the uniqueness of the solution.

Theorem 3.1 If (2) has one solution, then the solution is unique if and only if the sequence $\{\Lambda_j\in L(U,V_j)\}_{j\in J}$ is g-complete (i.e. $\{f\in U:\Lambda_j f=0,j\in J\}=\{0\}$).

Proof. For the sufficiency, if there exist $f,g\in U$ satisfying $\Lambda_j f=\Lambda_j g=g_j$ for $j\in J$, we obtain $\Lambda_j(f-g)=0$. Since $\{\Lambda_j\in L(U,V_j)\}_{j\in J}$ is g-complete, we have $f=g$.

For the necessity, if $\{\Lambda_j\in L(U,V_j)\}_{j\in J}$ is not g-complete, then there is a nonzero $f\in U$ such that $\Lambda_j f=0$ for $j\in J$. Let g be the solution of (2), then we obtain that $\Lambda_j g+\Lambda_j f=\Lambda_j(g+f)=g_j$, this implies that $g+f$ is also a solution of (2). And this is contradictory to the uniqueness of the solution. Therefore $\{\Lambda_j\in L(U,V_j)\}_{j\in J}$ is g-complete.

We give an unsolvable example in the following.

Example 3.2 Suppose that $\{e_j\}_{j=1}^3$ is an orthogonal basis for U.

For $j=1,2,3$, we let

$$V_1=\overline{span}\{e_1,e_2\},V_2=\overline{span}\{e_2,e_3\},V_3=\overline{span}\{e_2,e_3\}$$

and $\Lambda_j:U\to V_j$ is defined by:

$$\Lambda_1 f=\langle f,e_1\rangle e_1+\langle f,e_2\rangle e_2,$$
$$\Lambda_2 f=\langle f,e_2\rangle e_2+\langle f,e_3\rangle e_3,$$
$$\Lambda_3 f=\langle f,e_2\rangle e_3+\langle f,e_3\rangle e_2.$$

It is easy to check that Λ_j is a bounded linear operator for $j=1,2,3$ and $\{\Lambda_j\}_{j=1}^3$ is a g-frame for U with respect to $\{V_j\}_{j=1}^3$.

Given a sequence:

$$g_1=e_1+2e_2,g_2=2e_2+4e_3,g_3=2e_2+4e_3.$$

Assume $f=ae_1+be_2+ce_3$, then we have the following equations:

$$\begin{cases}ae_1+be_2=e_1+2e_2\\ be_2+ce_3=2e_2+4e_3\\ be_3+ce_2=2e_2+4e_3\end{cases}$$

It is esay to calculate that $b=2$ and $b=4$ which is a contradiction. Therefore (2) has no solution.

Next we discuss that under what conditions (2) is solvable.

Theorem 3.3 Let $\{\Lambda_j\}_{j\in J}$ be a sequence belonging to $L(U,V_j)$ and $\{g_j\}_{j\in J}$ is an element of $l^2(\{V_j\}_{j\in J})$. In order that the equation system $\Lambda_j f=g_j,j\in J$ shall admit at least one solution $f\in U$ for which $\|f\|\le M$, it is necessary and sufficient that

$$|\sum_{j\in J_1}\langle f_j,g_j\rangle|\le M\|\sum_{j\in J_1}\Lambda_j^* f_j\|$$

for any finite sequence $\{f_j\}_{j\in J_1}$ of $l^2(\{V_j\}_{j\in J})$ and any finite subset J_1 of J.

Proof. The necessity is trivial: if $\{f_j\}_{j\in J_1}$ is an arbitrary finite sequence of $l^2(\{V_j\}_{j\in J})$, then

$$|\sum_{j\in J_1}\langle f_j,g_j\rangle|=|\sum_{j\in J_1}\langle f_j,\Lambda_j f\rangle|=|\sum_{j\in J_1}\langle\Lambda_j^* f_j,f\rangle|$$
$$\le\|f\|\cdot\|\sum_{j\in J_1}\Lambda_j^* f_j\|\le M\|\sum_{j\in J_1}\Lambda_j^* f_j\|$$

For the sufficency, let

$W=\{\sum_{j\in J_1}\Lambda_j^* g_j \text{ for any finite } J_1\subset J \text{ and } g_j\in V_j, j\in J\}$

μ is defined on W by setting $\mu(\sum_{j\in J_1}\Lambda_j^* f_j)=\sum_{j\in J_1}\langle f_j, g_j\rangle$ for any finite sequence $\{f_j\}_{j\in J_1}$ of $l^2(\{V_j\}_{j\in J})$. Consider that if $\sum_{j\in J_1}\Lambda_j^* f_j=\sum_{j\in J_1}\Lambda_j^* h_j$, we have

$$\begin{aligned}&\|\mu(\sum_{j\in J_1}\Lambda_j^* f_j)-\mu\sum_{j\in J_1}\Lambda_j^* h_j)\|\\ =&\|\mu(\sum_{j\in J_1}\Lambda_j^* f_j-\sum_{j\in J_1}\Lambda_j^* h_j)\|=\|\mu(\sum_{j\in J_1}\Lambda_j^*(f_j-h_j))\|\\ =&\sum_{j\in J_1}\langle f_j-h_j, g_j\rangle\|\le M\|\sum_{j\in J_1}\Lambda_j^*(f_j-h_j)\|=0.\end{aligned}$$

Hence μ is defined unambiguously and it is easy to check μ is a bounded linear functional on W, with norm not exceeding M. Now we extend μ to the closure of W by continuity. Given the orthogonal projection $P:U\to\overline{W}$ and let $\eta=\mu P$. It follows that η is a bounded linear functional on U with bound M.

By Riesz representation theorem, there is a vector $f\in U$ such that $\eta(g)=\langle g, f\rangle$ for any $g\in U$, and $\|f\|=\|\eta\|\le M$. Consider that

$$\eta(\sum_{j\in J_1}\Lambda_j^* f_j)=\langle\sum_{j\in J_1}\Lambda_j^* f_j, f\rangle=\sum_{j\in J_1}\langle f_j,\Lambda_j f\rangle$$

and

$$\eta(\sum_{j\in J_1}\Lambda_j^* f_j)=\mu(\sum_{j\in J_1}\Lambda_j^* f_j)=\sum_{j\in J_1}\langle f_j, g_j\rangle,$$

we have $\sum_{j\in J_1}\langle f_j, g_j\rangle=\sum_{j\in J_1}\langle f_j,\Lambda_j f\rangle$. For a fixed but arbitrary $j\in J$, let $J_1=\{j\}$, we have $\langle f_j, g_j\rangle=\langle f_j,\Lambda_j f\rangle$. Since $\{f_j\}_{j\in J_1}$ is an arbitrary sequence of $l^2(\{V_j\}_{j\in J})$ and both f_j and g_j belong to V_j, we obtain $g_j=\Lambda_j f$.

According to the 2.6, if the operator T is surjective (2) have solutions. Next we consider the moment problem when the sequence $\{\Lambda_j\in L(U,V_j)\}_{j\in J}$ is a g-frame. The author of ariticle [10] showed that when $\{\Lambda_j\}_{j\in J}$ is a g-Riesz-Fischer sequence, the analysis operator Q^* of $\{\Lambda_j\}_{j\in J}$ is surjective. Under this condition, (2) has a unique solution. Hence for a g-frame which is not a g-Riesz-Fischer sequence, there indeed exist sequences $\{g_j\}_{j\in J}\in l^2(\{V_j\}_{j\in J})$ such that (2) has no solution.

The natural question is whether we can find a f in U such that $\sum_{j\in J}\|g_j-\Lambda_j f\|^2=\min_{g\in U}\{\sum_{j\in J}\|g_j-\Lambda_j g\|^2\}$.

Before answering the above question, we introduce the orthogonal projection from $l^2(\{V_j\}_{j\in J})$ onto R_{Q^*}.

Theorem 3.4 Let a sequence $\{\Lambda_j\in L(U,V_j)\}_{j\in J}$ be a g-frame for U with respect to $\{V_j\}_{j\in J}$ with bounds A and B. If P is the orthogonal projection from $l^2(\{V_j\}_{j\in J})$ onto R_{Q^*}, then P is defined by $P(\{g_j\}_{j\in J}):=\{\Lambda_j\sum_{k\in J}\widetilde{\Lambda}_k^* g_k\}_{j\in J}$ for any $\{g_j\}_{j\in J}$ belongs to $l^2(\{V_j\}_{j\in J})$.

Proof. Let $\widetilde{T}$ be the operator defined by:

$$l^2(\{V_j\}_{j\in J})\to R_{Q^*},\ \widetilde{T}(\{g_j\}_{j\in J}):=\{\Lambda_j\sum_{k\in J}\widetilde{\Lambda}_k^* g_k\}_{j\in J}.$$

Firstly, we prove that $\widetilde{T}$ is a bounded linear operator.

For all $\{g_j\}_{j\in J},\{h_j\}_{j\in J}\in l^2(\{V_j\}_{j\in J})$, consider that

$$\begin{aligned}&\widetilde{T}(\{g_j\}_{j\in J}+\{h_j\}_{j\in J})\\ =&\widetilde{T}(\{g_j+h_j\}_{j\in J})=\{\Lambda_j\sum_{k\in J}\widetilde{\Lambda}_k^*(g_k+h_k)\}_{j\in J}\\ =&\{\Lambda_j(\sum_{k\in J}\widetilde{\Lambda}_k^* g_k+\sum_{k\in J}\widetilde{\Lambda}_k^* h_k)\}_{j\in J}=\{\Lambda_j\sum_{k\in J}\widetilde{\Lambda}_k^* g_k\}_{j\in J}+\{\Lambda_j\sum_{k\in J}\widetilde{\Lambda}_k^* h_k\}_{j\in J}\\ =&\widetilde{T}(\{g_j\}_{j\in J})+\widetilde{T}(\{h_j\}_{j\in J})\end{aligned}$$

and

$$\begin{aligned}&\|\widetilde{T}(\{g_j\}_{j\in J})\|^2\\ =&\|\{\Lambda_j\sum_{k\in J}\widetilde{\Lambda}_k^* g_k\}_{j\in J}\|^2=\|\{\Lambda_j\sum_{k\in J}S^{-1}\Lambda_k^* g_k\}_{j\in J}\|^2\\ =&\sum_{j\in J}\|\Lambda_j\sum_{k\in J}S^{-1}\Lambda_k^* g_k\|^2\le B\|\sum_{k\in J}S^{-1}\Lambda_k^* g_k\|^2\\ \le&\frac{B}{A}\|\{g_k\}_{k\in J}\|^2.\end{aligned}$$

Hence, $\widetilde{T}$ is a bounded linear operator.

Secondly, let $g=\sum_{j\in J}\widetilde{\Lambda}_j^* g_j$, since

$$\begin{aligned}\widetilde{T}^2(\{g_j\}_{j\in J})=\widetilde{T}(\{\Lambda_j g\}_{j\in J})&=\{\Lambda_j\sum_{k\in J}\widetilde{\Lambda}_k^*\Lambda_k g\}_{j\in J}\\ &=\{\Lambda_j\sum_{k\in J}S^{-1}\Lambda_k^*\Lambda_k g\}_{j\in J}=\{\Lambda_j S^{-1}\sum_{k\in J}\Lambda_k^*\Lambda_k g\}_{j\in J}\\ &=\{\Lambda_j g\}_{j\in J}=\widetilde{T}(\{g_j\}_{j\in J}),\end{aligned}$$

we obtain $\widetilde{T}^2=\widetilde{T}$. Hence we have that $\widetilde{T}$ is a projection from $l^2(\{V_j\}_{j\in J})$ onto R_{Q^*}.

Finally, for all $\{g_j\}_{j\in J},\{f_j\}_{j\in J}\in l^2(\{V_j\}_{j\in J})$, we obtain

$$\begin{aligned}\langle\widetilde{T}(\{g_j\}_{j\in J}),\{f_j\}_{j\in J}\rangle&=\langle\{\Lambda_j\sum_{k\in J}\widetilde{\Lambda}_k^* g_k\}_{j\in J},\{f_j\}_{j\in J}\rangle\\ &=\sum_{j\in J}\langle\Lambda_j\sum_{k\in J}\widetilde{\Lambda}_k^* g_k, f_j\rangle\\ &=\sum_{j\in J}\langle\sum_{k\in J}\widetilde{\Lambda}_k^* g_k,\Lambda_j^* f_j\rangle\end{aligned}$$

$$= \langle \sum_{k\in J} \widetilde{\Lambda}_k^* g_k, \sum_{j\in J} \Lambda_j^* f_j \rangle = \langle \sum_{k\in J} S^{-1}\Lambda_k^* g_k, \sum_{j\in J} \Lambda_j^* f_j \rangle$$
$$= \langle S^{-1}\sum_{k\in J} \Lambda_k^* g_k, \sum_{j\in J} \Lambda_j^* f_j \rangle = \langle \sum_{k\in J} \Lambda_k^* g_k, \sum_{j\in J} S^{-1}\Lambda_j^* f_j \rangle$$
$$= \langle \sum_{k\in J} \Lambda_k^* g_k, \sum_{j\in J} \widetilde{\Lambda}_j^* f_j \rangle = \sum_{k\in J} \langle g_k, \Lambda_k \sum_{j\in J} \widetilde{\Lambda}_j^* f_j \rangle$$
$$= \langle \{g_k\}_{k\in J}, \{\Lambda_k \sum_{j\in J} \widetilde{\Lambda}_j^* f_j\}_{k\in J} \rangle$$
$$= \langle \{g_j\}_{j\in J}, \widetilde{T}(\{f_j\}_{j\in J}) \rangle.$$

It shows that $\widetilde{T}^* = \widetilde{T}$.

Hence, $\widetilde{T}$ is an orthogonal projection from $l^2(\{V_j\}_{j\in J})$ onto R_{Q^*}. Since orthogonal projection is unique, we have $P = \widetilde{T}$.

Remark 3.5 Let a sequence $\{\Lambda_j \in L(U,V_j)\}_{j\in J}$ be a g-frame for U with respect to $\{V_j\}_{j\in J}$. For a sequence $\{f_j\}_{j\in J} \subset U$, let $V = C$ and $\Lambda_j(\cdot) = \langle \cdot, f_j \rangle$, then $\{f_j\}_{j\in J}$ is a frame for U. Given a sequence $\{g_j\}_{j\in J} \in l^2(\{V_j\}_{j\in J})$ and by 3.4 we obtain that

$$P(\{g_j\}_{j\in J}) = \{\Lambda_j \sum_{k\in J} \widetilde{\Lambda}_k^* g_k\}_{j\in J}$$
$$= \{\Lambda_j \sum_{k\in J} S^{-1}\Lambda_k^* g_k\}_{j\in J}$$
$$= \{\Lambda_j S^{-1} \sum_{k\in J} \Lambda_k^* g_k\}_{j\in J}$$
$$= \{\langle \sum_{k\in J} S^{-1} f_k g_k, f_j \rangle\}_{j\in J}$$

which is 5.3.6 of article [6].

The following corollary can be derived by 3.4 directly.

Corollary 3.6 Let a sequence $\{\Lambda_j \in L(U,V_j)\}_{j\in J}$ be a g-frame for U with respect to $\{V_j\}_{j\in J}$. Then for any $f \in U$, we have

$$\| \{\Lambda_j f\}_{j\in J} \| \le \| \{g_j\}_{j\in J} \|$$

for all $\{g_j\}_{j\in J} \in l^2(\{V_j\}_{j\in J})$ satisfying $f = \sum_{j\in J} \widetilde{\Lambda}_j^* g_j$.

Corollary 3.7 [14] Let $\{(W_i, v_i)\}_{i\in I}$ be a fusion frame for H with fusion frame operator S_W, let $T_{S_W^{-1}W}$ denote the analysis operator for the fusion frame given by $\{(S_W^{-1}W_i, v_i)\}_{i\in I}$ and let P be the orthogonal projection of $(\sum_{i\in I} \oplus S_W^{-1}W_i)_{l^2}$ onto $Rng(T_{S_W^{-1}W})$. Then for every $f \in H$, we have $\| P\{v_i S_W^{-1}\pi_{W_i}(f)\}_{i\in I} \|_2 \le \| \{v_i g_i\}_{i\in I} \|_2$ for all $\{g_i\}_{i\in I} \in (\sum_{i\in I} \oplus S_W^{-1}W_i)_{l^2}$ satisfying $f = \sum_{i\in I} v_i^2 g_i$.

Proof. Since a fusion frame is a special g-frame, the properties of g-frame are also true to fusion frame. Let $\Lambda_i = v_i \pi_{S_W^{-1}W_i}, i \in I$, then for all $\{g_i\}_{i\in I} \in (\sum_{i\in I} \oplus S_W^{-1}W_i)_{l^2}$ we have $P(\{g_i\}_{i\in I}) = \{\Lambda_i \sum_{j\in J} \widetilde{\Lambda}_j^* g_j\}_{i\in I}$ by 3.4. Consider that

$$P(\{v_i g_i\}_{i\in I}) = \{v_i \pi_{S_W^{-1}W_i} \sum_{j\in I} (v_j \pi_{S_W^{-1}W_j} S_{S_W^{-1}W}^{-1})^* (v_j g_j)\}_{i\in I}$$
$$= \{v_i \pi_{S_W^{-1}W_i} \sum_{j\in I} v_j S_{S_W^{-1}W}^{-1} \pi_{S_W^{-1}W_j} (v_j g_j)\}_{i\in I}$$
$$= \{v_i \pi_{S_W^{-1}W_i} S_{S_W^{-1}W}^{-1} \sum_{j\in I} v_j^2 g_j\}_{i\in I}$$

and

$$P(\{v_i S_W^{-1}\pi_{W_i}(f)\}_{i\in I}) = \{v_i \pi_{S_W^{-1}W_i} \sum_{j\in I} (v_j \pi_{S_W^{-1}W_j} S_{S_W^{-1}W}^{-1})^* v_j S_W^{-1}\pi_{W_j}(f)\}_{i\in I}$$
$$= \{v_i \pi_{S_W^{-1}W_i} \sum_{j\in I} v_j S_{S_W^{-1}W}^{-1} \pi_{S_W^{-1}W_j} v_j S_W^{-1}\pi_{W_j}(f)\}_{i\in I}$$
$$= \{v_i \pi_{S_W^{-1}W_i} S_{S_W^{-1}W}^{-1} \sum_{j\in I} v_j^2 \pi_{S_W^{-1}W_j} S_W^{-1}\pi_{W_j}(f)\}_{i\in I}$$
$$= \{v_i \pi_{S_W^{-1}W_i} S_{S_W^{-1}W}^{-1} \sum_{j\in I} v_j^2 S_W^{-1}\pi_{W_j}(f)\}_{i\in I}$$
$$= \{v_i \pi_{S_W^{-1}W_i} S_{S_W^{-1}W}^{-1} S_W^{-1} \sum_{j\in I} v_j^2 \pi_{W_j}(f)\}_{i\in I}$$
$$= \{v_i \pi_{S_W^{-1}W_i} S_{S_W^{-1}W}^{-1} f\}_{i\in I}$$

we obtain when $f = \sum_{i\in I} v_i^2 g_i$, $\{v_i S_W^{-1}\pi_{W_i}(f)\}_{i\in I}$ is just the orthogonal projection of $\{v_i g_i\}_{i\in I}$ onto $Rng(T_{S_W^{-1}W})$. By 3.6 it is easy to know that $\| P\{v_i S_W^{-1}\pi_{W_i}(f)\}_{i\in I} \|_2 \le \| \{v_i g_i\}_{i\in I} \|_2$.

The following we will discuss the best approximative solution of (2).

Theorem 3.8 Let $\{\Lambda_j \in L(U,V_j)\}_{j\in J}$ be a g-frame for U with respect to $\{V_j\}_{j\in J}$ and let $\{g_j\}_{j\in J} \in l^2(\{V_j\}_{j\in J})$. Then there exists a unique element in U, such that $\| \{g_j\}_{j\in J} - \{\Lambda_j f\}_{j\in J} \|^2 = \min_{\{h_j\}_{j\in J} \in R_{Q^*}} \{\| \{g_j\}_{j\in J} - \{h_j\}_{j\in J} \|^2\}$. And this element is $f = \sum_{j\in J} \widetilde{\Lambda}_j^* g_j$ which is called the best approximative solution.

Proof. By 2.7, there exists a unique sequence $\{f_j\} \in R_{Q^*}$ such that

$$\| \{g_j\}_{j\in J} - \{f_j\}_{\in J} \|^2 = \min_{\{h_j\}_{j\in J} \in R_{Q^*}} \{\| \{g_j\}_{j\in J} - \{h_j\}_{\in J} \|^2\}$$

Further by 2.8, we obtain that $\{f_j\}_{j\in J}$ is just the orthogonal projection of the sequence $\{g_j\}_{j\in J} \in l^2(\{V_j\}_{j\in J})$ onto R_{Q^*}. Consider that $Q^* f = \{\Lambda_j f\}_{j\in J}$ for any f of U and by 3.4, we have $\{f_j\}_{j\in J} = \{\Lambda_j \sum_{k\in J} \widetilde{\Lambda}_k^* g_k\}_{j\in J}$ and $f = \sum_{j\in J} \widetilde{\Lambda}_j^* g_j$. In addition, the obtained f is unique by the uniqueness of the orthogonal projection.

ACKNOWLEDGMENT

This work is partially supported by NSFC under Grant No. 11071270, Natural Science Foundation of Fujian Province under Grant No. 2010J01328 and the Technology Innovation Platform Project of Fujian Province under Grant No.2009J1007. The authors are very thankful to Professor Zhu for carefully reading, very helpful suggestion and useful comments.

REFERENCES

[1] T. J. Stieltjes, "Recherches sur les fractions continues," Ann.Fac.Sci.Toulouse, 1894, 8, J1-122, 9, A1-47; Oeuveres 2, 402-566.

[2] G. D. Lin, "On the moment problem ," Statist.Prob.Lett. , 1997, 35:85-90.

[3] H. Gzyl, A. Tagliani, "Hausdorff moment problem and fractional moments," Appl.Math.Comput.2010, 216: 3319-3328.

[4] O. Christensen, "Moment problems and stability results for frames with applications to irregular sampling and Gabor frames," Appl.Comput.Harmon.Anal. , 1996, 3: 82-86.

[5] R. M. Young, "An introduction to non-harmonic Fourier series," New York: Academic Press, 1980.

[6] O. Christensen, "An introduction to frames and Riesz bases," Boston: Birkhäuser, 2002.

[7] D. F. Li, M. Z. Xue, "Basis and frames in Banach spaces (in Chinese)," Beijing: Science Press, 2007.

[8] W. C. Sun, "G-frames and g-Riesz bases," Math.Anal.Appl, 2006, 322 (1): 437 - 452.

[9] W. C. Sun, "Stability of g-frames," Math.Anal.Appl, 2007, 326 (2): 858-868.

[10] X. J. Huang, Y. C. Zhu, "G-Riesz-Fischer sequence in Hilbert spaces (in Chinese)," Journal of Fuzhou University (Natural Science Edition), 2008, 36 (6): 779-783.

[11] Y. C. Zhu, "Characterizations of g-frames and g-Riesz bases in Hilbert spaces," Acta Mathematica Sinica, English Series, 2008, 24 (10): 1727-1736.

[12] C. Heil, "A basis theory primer," Boston: Birkhäuser, 2011.

[13] E. Suhubi, "Functional analysis," Kluwer Academic Publishers, 2003.

[14] P. G. Casazza, G. Kutyniok and S.D. Li, "Fusion frames and distributed processing," Appl.Comput. Harmon. Anal., 2008, 25:114-132.

A Highly Practicable Distributed Architecture of Network Stream Media

XIAOFENG HU
China's ministry of culture

FANGZHU WANG
BEIJING SUPER SKY company

YANPING CHENG
China's ministry of culture

JINLIANG ZHANG
China's ministry of culture

Abstract: The rapid development of computer network has greatly popularized online education. Currently, most online education systems are using a learning platform of three-split screen courseware, in which stream media take up a majority of resources in bandwidth and storage space. Due to more and more students receiving online education, the load of stream media service is getting heavier, so that the traditional centralized architecture and image architecture cannot satisfy the access requirements of a large number of nationwide students. Therefore, a highly practicable stream media application architecture with quick response and high concurrency, reliability, stability, extendability. This paper first discusses the development of online education as well as the problem, then deeply explores the key technology to realize distributed architecture of stream media according to characteristics of online education, with the introduction of P2P and CDN architecture, along with stream media subdivision technology.

Key words: online education; stream media; P2P; CDN

I. INTRODUCTION

There are 3 courses in the development of online education: correspondence education, radio-TV education and education based on computer networks and multimedia[1]. Many regions are converting from the original remote education to online education, as a result, a new pattern of coexistence of various media and online education is formed. Vigorous development of online education system is of great realistic significance for promoting education popularization and establishing lifelong education in China, to realize great-leap-forward development of education.

The bandwidth and storage resource of servers of an education website are limited. Under sudden flow visit, the resource of both bandwidth and servers could become the bottleneck for user access leading to decreased access quality. Centralized content placement of traditional Internet inevitably results in disorderly flow of the entire network, only to reduce the integrity of resources, thus unable to guarantee effective service. The hit rate of online courses is rising continually, which challenges the network load. It's conspicuous that stream media servers are overloaded and there is heavy pressure on network bandwidth, etc. Stream media consumes much bandwidth and storage space, so only upgrading bandwidth, storage and servers cannot support high concurrent visits, and that costs too much. With the rapid development of online education, the reliability, consistence, security and high usability of network system is getting more and more important. Therefore, it's absolutely necessary to increase extendability, service capability and usability of websites. A better architecture is needed to solve all the problems. Presently, P2P and CDN are widely used in stream media application, but their combination, especially in online education, is to be advanced. This paper puts forward a highly practicable architecture of stream media combined with P2P and CDN, with high response, concurrency, reliability, stability and extendability.

II. STREAM MEDIA CDN TECHNOLOGY

A. Overview of CDN

CDN(Content Delivery Network) is to select a nearest edge server throughout the nation for user request. When users access websites joining in CDN service, DNS (domain name resolution) request will be disposed of by redirecting DNS, which provides the address of the nearest edge server for user through a predefined strategy such as content type, location and network load situation. Meanwhile, the website can self-adapt its network situation. When one or more nodes become invalid, it can reschedule a usable edge server node for users. Furthermore, it should adapt to dynamic changes of user requests, that is to reallocate edge servers according to new user request content.

As far as current video coding technique is concerned, the bandwidth to play video courseware is 200kbps order of magnitude, taking up 90% bandwidth and storage space of the website. Disk access of stream media consumes much resource too. CDN architecture must be adopted to make users watch fluent courseware video with good learning experience. The fundamental principle of stream media CDN is to copy and deliver frequently hit media content chosen by users from stream media central server to edge servers nearest to end users, and to provide video for end users by edge server nearest to request terminal. The idea of CDN design is that the needed bandwidth and time of service on initial server and edge servers are much less than that of non-CDN environment where end users visit central stream media server all together. CDN not only enlarges the stream media range visited by users, quickens the response, effectively lowering the pressure on central device and backbone network, but also well relieves the jam in backbone network caused by big traffic and unbalanced distribution of servers, thus it is a very effective network mode of VOD(video on demand)[3].

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.4

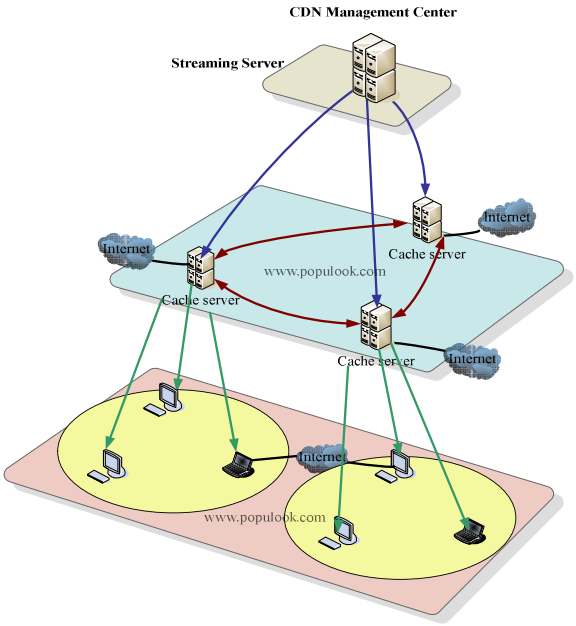


Figure 1 Stream Media CDN Mode

B. *Intelligent DNS Schedule*

The basic function of CDN is nearby visit. Request Routing based on DNS is widely used in lots of CDN application on Internet, with flexible and convenient technical realization and low cost. DNS servers process domain names of requested sites or contents, then users search initial names in local DNS server. If local DNS cache fails, it will send name search to DNS root servers, which returns the address of authoritative DNS server. Authoritative DNS is controlled by CDN Service Provider. It will return the address of an edge server near the user according to current network situation like load monitoring information and routing map. Finally, user can get the content from the designated edge server. This procedure is transparent for users, that is to say, users can visit the education site by inputting the original domain name without any settings in client-side after adding cache.

After user provides the domain name to be visited for courseware player, the IP of real serving host machine is not directly resolved due to the adjustment of CDN for resolution procedure. The resolution result is usually the corresponding CNAME. To obtain the real IP address, it is needed to resolve the CNAME again. In that process, global load balance DNS is used, for example, to resolve corresponding IP according to location, load pressure and the Network Operator, to realize nearby visit. Then the IP address of CDN edge node is obtained. The player sends stream media visit requests to edge nodes after acquiring the real IP. The visiting procedure is completed with nearby access. Client can obtain date returned from edge nodes and display them, thus the data request is accomplished.

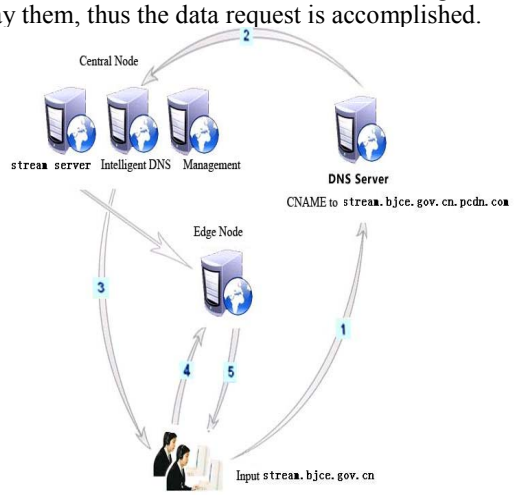


Figure 2 Data Request Procedure

Courseware content is delivered to each CDN nodes to respond to user access request. CDN nodes are located on the edge of each network, so users can reach edge nodes to obtain content through less intermediate transmission. Thus, CDN mode can effectively reduce the influence on final access quality of intermediate network transmission.

Either the complicated network structure of Internet, or the bottleneck of connecting different Network Operator is alleviated by CDN architecture. Even when part of network device or link broke down, visiting CDN nodes would hardly be affected.

Hence, CDN architecture can ameliorate bad influence by time delay of network, quicken the response of websites, and improve access quality linked from different network operators, providing good service for users of various locations.

In reliability, CDN uses a structure with redundancy in nodes. If some nodes are wrong, the access can be directed to other normal nodes, so that the service would not be interrupted.

C. *PULL and PUSH Delivery of Stream Media*

Under multi-server, system automatically synchronizes media file on each server by PUSH (for hot courseware) or PULL (for unpopular courseware) request. Edge nodes use PUSH to request and store the complete data of popular courseware before its release. As for unpopular ones, edge nodes will not keep the data. While being visited, it PULLs the data from central nodes to serves users. Edge nodes can obsolete cached data by LRU algorithm according to the size of storage space.

The PULL way, to respond to user request on time, must be able to download data from other nodes while playing, and satisfy the need to drag to watch other chapters. The traditional way is to make users ask central nodes for data while edge nodes download from central nodes. When the video is downloaded completely, more user request for stream media service can be satisfied. This method has some disadvantages:

1) It disobeys the nearby visit principle;

2) When edge nodes have not downloaded complete video, all users would visit central nodes, increasing its bandwidth and load pressure;

3) It enlarges the storage space of edge nodes.

This paper uses stream media slicing technique, and each slice can be played independently. Edge nodes PULL a slice of data according to user request. Thus, all users can visit the nearest edge nodes with quick response.

D. Stream Media Slicing

If asking each edge node to synchronously store these huge courseware, the storage will cost too much, and it takes much time to hand out them. To lessen storage space and distributing time, edge nodes only store part of data. Hot courseware is usually pre-stored, and unpopular ones are PULLed from other nodes when needed.

Slicing Stream Media storage uses stream based content storage and delivery of mixed stream media service network instead of file based storages and centralized media delivery of traditional CDN. Each edge server contains service and storage function of stream media.

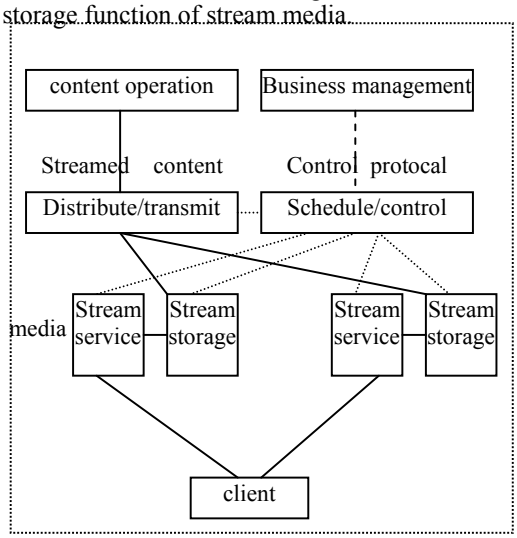


Figure 3 Slicing Stream Media Technique

This paper slices stream media data with the unit of second. One slice is one second long data. The central node first resolves video file to obtain its compressing format of audio and video, metadata, timestamp of each record, then slices the video by second. Client first acquires the metadata to play, and while requesting for data in certain second, only needs to obtain the slice of the second.

Usually, for all kinds of reasons, slices on edge nodes which users belongs to are incomplete, then that node will request other nodes for slice data. In this way, the need for storage space and resource is reduced a lot while the reliability and stability of stream media service is enhanced.

When to be extended, it is needed to synchronize video files on new nodes, so they can be put into use soon. Thus, the good extendability is achieved.

E. P2P Accelerating Distribution

Peer-to-peer network(P2P) is a newly emerging network technology. It breaks through the traditional client/server model and establishes a communication mechanism as end-to-end direct exchange of data or service without trunk equipment. In P2P network, each terminal node can be both client and server to provide service to other requesting nodes, that is to say, every node has equal status. P2P, with the idea of service evenly distributed, allocates the traditional server task to every terminal node, using terminal nodes to accomplish limited storage and computation. The more nodes in P2P, the richer is service resource, and the better of service quality[4].

Organizing edge nodes with P2P and its category service and multi-points transmission capability can realize mutual backup and content exchange between edge nodes, improve content distribution ability of CDN, and reduce bandwidth and load pressure of central nodes while accelerating distribution.

During P2P sharing, edge nodes select the same network operator and light-loaded edge nodes in priority.

P2P not only accelerates distributing, but also costs less bandwidth.

III. CONCLUSION

Through the above discussion, the new constructing method of stream media network with P2P and CDN architecture provides a highly practicable supporting platform of stream media network system with better network service for various applications. Its advantage in remote education prominently lies in reducing load of each stream media server and effectively relieving pressure on output bandwidth. Thus, the network really becomes a highly effective platform supporting all kinds of teaching activities, and a convenient fast channel for remote education, providing practical service for teachers and students. Looking forward to the future, the aim of continuous extending of 3G and internet and convergence of three networks is to provide more contents and added services for internet users. As for online education, with the increase of users and abundance of teaching resources, the room for stream media CDN technique will be larger, so stream media CDN has a bright future in online education.

[1] Jiang Qing, Yan Qiuyi. Problems and Strategies in Modern Remote Education. Education Informatization, 2006(1): 12-13.

[2] Yuan Yasheng, Jia Zhuosheng. Application of CDN in Remote Education[J]. Journal of Huazhong University of Science and Technology, 2003.

[3] Gang Peng.CDN:Content Distribution Network,USA:Department of Computer Science,State University of New York,2003:137-142

[4] Zhang Zongyong. Research of IPTV Service System Based on P2P[D]. Xi An: Northwestern Polytechnical University, 2006.

[5] Hua K A,Tantaoui MA,Tavanapong W,Video delivery technologies for large-scale deployment of multimedia applications,Proceedings of the IEEE,Volume 92,Issue 9,Sep 2004:1439-1451.

*Parallel Implementation of Fractal Image Compression in Web Service Environment**

Yan Fang**
College of Computer and Information
Fujian Agriculture and Forestry University
Fuzhou, Fujian, China
e-mail: fangyan0399257@163.com

Hang Cheng，Meiqing Wang
College of Mathematics and Computer Science
Fuzhou University
Fuzhou, Fujian, China
e-mail: sandao163@163.com

***Abstract*—Fractal image compression algorithm has some problems, such as high computational complexity and long time consumption. In this paper the parallel implementation of fractal image compression algorithm in Web service environment is discussed. The experiments show this algorithm can greatly reduce the compression time and achieve a high speedup at the same quality of decompressed image. The proposed algorithm can achieve the speedup of about 500 times with a little loss of lenna image's PSNR and two node computers against fractal full search algorithm.**

Keywords- fractal； Web Service； image compression； parallel；

I. INTRODUCTION

Fractal image compression is a relatively recent image compression method [1]. It can achieve a high compression ratio. The decompression algorithm is very simple which only compute the fixed point of a fractal transform operator equation. Fractal image compression algorithm is suitable for the situation of one encoding and many decoding. However, the computational complexity is very high which is rising dramatically with the size of image[2]. At present, some researchers have put forward many methods in order to accelerate fractal coding [3-6], such as the classification method and the neighbor-region searching method. But these methods may reduce the quality of decompressing image or the compression ratio. In fractal compress algorithm, each range block needs to search many domain blocks, so the main computing work in fractal image compression lies in the procedure of matching search. The process of matching domain blocks is independent, so we can use parallel algorithms in order to achieve a high speedup on fractal image compression.

There are two kinds of architecture on conventional distributed technology. One is based on remote procedure call and the other is based on message. The technologies often used to build distributed computing systems are Microsoft's COM/DCOM, Sun's EJB/RMI and OMG's COBRA, which have some problem such as how to pass through the firewall, the differences of protocols and the data format conversion. However Web Service is based on the internet standard protocols, which uses the HTTP protocol and XML data to communicate [7]. It can easily pass through the firewall, across the different platform and language. So this paper proposes the distributed parallel computing system called WDCS which is based on Web Service, and uses this system to implement the fractal image compression parallel algorithm. The experiment results show this algorithm can greatly reduce the compression time and achieve a high speedup at the same quality of decompressed image.

II. FRACTAL IMAGE COMPRESSION

Fractal image compression is based on the iterated function system. The image data is transformed into affine transformation coefficients based on fractal theory. The amount of coefficient is less than that of the image data, so the image can be compressed. Iterated Function System theory was first applied to image coding by Barnsley. In his method, an image is considered as an Iterated Function System. In 1989, Arnaud Jacquin proposed an automatic method of fractal image compression in his master thesis [8]. In this method, an Iterated Function System is based on a small part of an image instead of on an entire image. It became the main method of the fractal image compression at present. The process of fractal image compression is as follows:

1) **Partitioning image.** An image is partitioned into non-overlapping small blocks of equal size, called range block, as shown in Fig.1(a). The original image is also partitioned into overlapping blocks with double size of the range block, called domain block, as shown in Fig.1(b). Each domain block is shrunk as the same size of the range block by averaging the intensities of pairwise disjoint groups of neighboring pixel intensities. The shrunk domain block is denoted as D , which is also known as a codebook block.

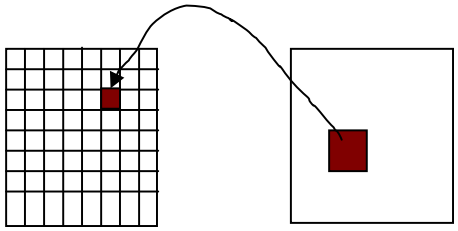
2) **Searching.** For each range block, an optimal approximation $R \approx \alpha \mathrm{D} + \beta \mathrm{I}$ can be obtained by searching through all or a subset of the domain blocks, based on the fixed-point theory of the Iterated Function System (IFS). α is the scaling factor. β is the offset factor. I is a unit vector.

3) **Saving the coding data.** When the domain block satisfies $E(R, D_k) = \min E(R, D_i)$, it is the solution of the optimal approximation $R \approx \alpha \mathrm{D}_{\mathrm{k}} + \beta \mathrm{I}$. The values of α and β which are known as the compression codes for R , can be obtained by using the Least Square methods.

The decoding process is relatively simple. Any image with the same size of the original image can be used as the initial image. According to the compression codes, we can get the decompression image through finite iteration calculation.

*This work was supported by Fujian Agriculture and Forestry University, China under Grant No.2010022.
**Corresponding author.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.66



（a）range block （b）domain block

Fig.1 the sketch map of block match

III. WEB SERVICE TECHNOLOGY

A. *Web Service introduction*

Web Service is known as XML Web Service. Web Service is URL addressable, which is publicly a function set on the internet. It is based on the HTTP protocol, XML data and SOAP protocol, so it can shield the different operating system and different programming language, avoiding the data format conversion from CORBA or DCOM into other protocols. So it can be used as a new platform to establish the distributed application.

B. *Web Service architecture*

Web Service is an implementation of service-oriented architecture (SOA), which combines component-oriented method and Web technology [9], as shown in Fig.2.

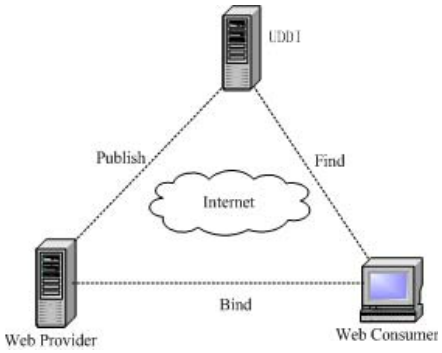


Fig.2 Web Service Architecture

Web Service architecture is based on Web Service Provider, Web Service Consumer and Web Service Broker. The three roles collaborate to support the "publish, find and bind" paradigm.

1) **Web Service Broker** It is implemented by network registered nodes which mainly contain UDDI (Universal Description, Discovery, and Integration). It allows for the lookup of service provider interfaces to interested service consumers.
2) **Web Service provider** It designs and implements web services, which are described through WSDL and registered in the UDDI registration center.
3) **Web Service consumer** It initiates the enquiry of the service in the registry, binds to the service over a transport, and executes the service function. The service consumer executes the service according to the interface contract.

IV. DISTRIBUTED PARALLEL COMPUTING SYSTEM BASED ON WEB SERVICE (WDCS)

The distributed parallel computing system based on Web Service is composed of monitor, client and Web Service computing nodes, as shown in Fig.3.

1) **Monitor** Web Service computing nodes can be added, deleted and scheduled on the monitor, and the client can search these computing nodes easily from the monitor.
2) **Client** It is the core of WDCS. Firstly, it searches those needed Web Service computing nodes from the monitor. Secondly, it will establish a connection if the nodes are found. Thirdly, it partitions the parallel computing tasks into many smaller task blocks which are then put into a task pool, sending task blocks located in the task pool to these connected computing nodes. Finally, it receives results from the computing nodes.
3) **Web Service computing node** It is responsible for computing task blocks and sending result to client.

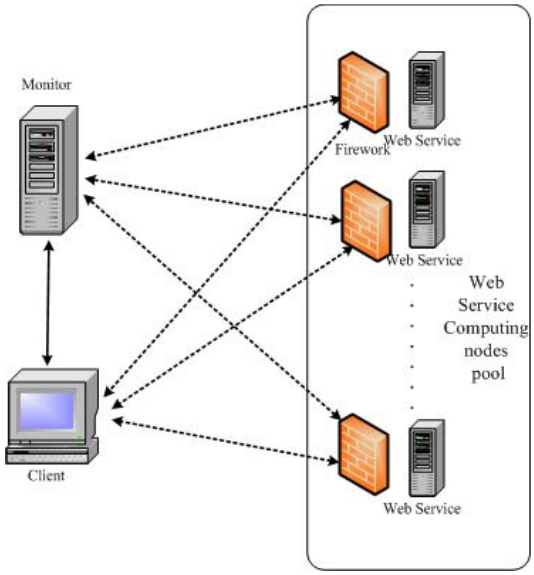


Fig.3 WSDC System

V. FRACTAL IMAGE COMPRESSION PARALLEL ALGORITHM BASED ON WEB SERVICE

The conventional fractal image compression algorithm described in Section II is a single serial algorithm. The matching procedures for range blocks are independent，so we can use distributed parallel algorithm on it.

The following algorithm is the parallel implement of the fractal image compression process in Web Service environment.

1) The original image is partitioned into non-overlapping small range blocks of equal size, such as the size of 4×4, 8×8 or 16×16.
2) According to a certain step length, all the range blocks are divided into multiple task blocks. The

value of step length can be determined by the number of the image's lines or Web Service computing nodes.

The structure framework of the task block is as following.

```
public struct Taskblock
{
    BlockID; //the ID of task block
    BlockData; // the data of task block
    BlockResult; // Save the computational results which Web Service computing nodes send back
    BlockFinish; // Used to indicate whether the task is finished. If finished, it is marked as true. On the contrary, it is marked as false and the task is allocated to another computing node again.
}
```

Based on the structure of task blocks and try-catch exception handling statement, we can ensure the WDCS system's reliability and fault tolerance.

3) Client can search its required Web Service computing nodes from the monitor, and then establish connection with them by binding.
4) Client creates a thread for each Web Service computing node which has been established connection with client. This thread is responsible for assigning task blocks to corresponding Web Service computing node and receiving the computational result.

This paper uses the thread pool (ThreadPool) technology of C #. NET. Automatic optimization management of ThreadPool can solve the problem that researchers need to spend plenty of time and vigor managing execution process of multiple threads and the competition for resources occupation. So ThreadPool can improve the parallel operation ability of WDCS system in this paper.

5) All computational results are sorted according to the BlockID of task blocks, and then written into compressed file together.

The following figure 4 show the flow diagram of fractal image compression parallel algorithm based on Web Service.

VI. EXPERIMENTS AND THE RESULTS

In order to simplify the experimental environment, the client is connected to Web Service computing node directly by hard-code, without the monitor. If the monitor is an independent computer, the compression time will be increased, because communication between monitor and clients will take some time. This experiment is based on a local network consisting of five PCs each of Intel Pentium4 2.93GHz CPU with a 1GB RAM. Their operating system is Microsoft Windows XP. One computer is used as the client installed Visual C#.NET 2005, the other four computers are used as Web Service computing nodes installed IIS and .NET Framework. There are different firewalls in these four computers, such as 360, skynet, windows and rising firewall. In this paper, we use fractal image compression algorithm with different search scope to verify the effectiveness of this paper proposed algorithm.

Test1: Using the 8-bit grey conference image of 256*256 pixels(as shown in Fig.5(a).) and the 8-bit grey lenna image of 512*512 pixels (as shown in Fig.6(a).). In this experiment, we use global search for domain block and the size of range block is 8*8. The compression ratio of conference image is 15.977. The decompressed image is shown in Fig.5 (b). The compression ratio of lenna image is 15.994. The decompressed image is shown in Fig.6 (b). The experimental data is shown in table1.

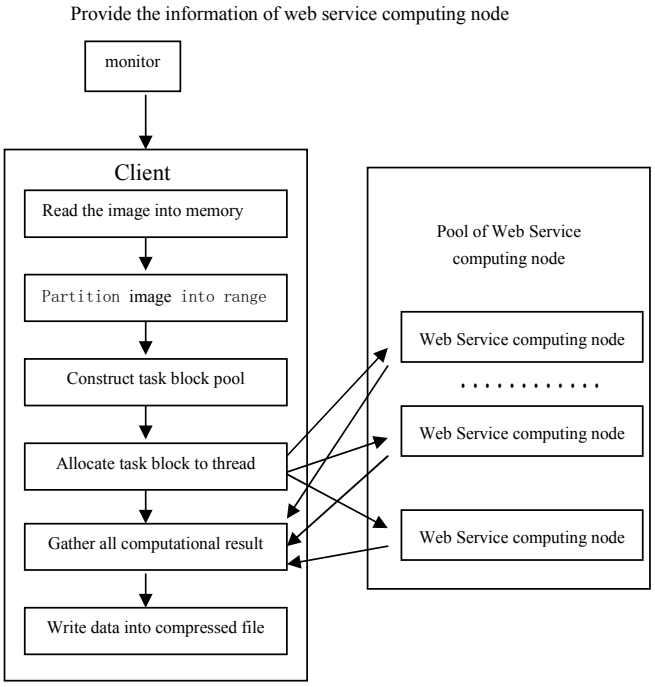


Fig.4 The flow diagram of fractal image compression parallel algorithm based on Web Service

TABLE 1. COMPARISON OF THE EXPERIMENT RESULT BETWEEN THE PROPOSED ALGORITHM AND THE CONVENTIONAL SERIAL ALGORITHM BASED ON THE FULL SEARCH

Number of Computing nodes	Conference image			Lenna image		
	PSNR	*Running time (second)*	*Speedup ratio*	*PSNR*	*Running time (second)*	*Speedup ratio*
0 (Single serial algorithm)	32.544	113	—	34.360	1934	—
2		23	4.91		377	5.13
4		12	9.42		192	10.07

Fig.5 (a) conference original image

(b) decompressed image

Fig.6 (a) lenna original image

(b) decompressed image

Fig.7 (a) lenna original image

(b) decompressed image

Fig.8 (a) movie original image

(b) decompressed image

From the table 1, we can find that：(1)at the same PSNR and compression ratio, the more number of Web Service computing nodes it has，the greater the speedup ratio will be and the shorter compression time will be. (2) In the same number of Web Service computing nodes, the speedup ratio of lenna image is greater than conference image. The more data image has, the better effect will be.

Test2: Using the 8-bit grey lenna image of 512*512 pixels (as shown in Fig.7(a).) and the 8-bit grey movie image of 720*576 pixels(as shown in Fig.6(a).). In this experiment, we use neighbor search for domain block and the size of range block is 8*8. The compression ratio of lenna image is 15.994. The decompressed image is shown in Fig.7 (b). The compression ratio of movie image is 15.996. The decompressed image is shown in Fig.8 (b). The experimental data is shown in table2.

TABLE 2. COMPARISON OF THE EXPERIMENT RESULT BETWEEN THE PROPOSED ALGORITHM AND THE CONVENTIONAL SERIAL ALGORITHM BASED ON THE NEIGHBOR SEARCH

Number of Computing nodes	**Lenna image**			**Movie image**		
	PSNR	*Running time (second)*	*Speedup ratio*	*PSNR*	*Running time (second)*	*Speedup ratio*
0 (Single serial algorithm)	33.235	23	—	31.764	37	—
2		4	5.75		6	6.17
4		3	7.67		4	9.25

From the table 2, we can find that the results of test2 is similar to test1. But when the number of Web Service computing nodes is 4, the speedup of lenna image is lower than the movie image. The two main reasons are as follows: (1) Local search for domain block can largely reduce compression time. The single serial algorithm only need a little time in itself, so the parallel algorithm reduce only 1 or 2 second compression time, the speedup ratio can be increased greatly; (2) the communication of WDCS needs some time, so the compression time can not been decreased unlimitedly.

From test1 and test2, at the same PSNR or compression ratio, the proposed algorithm in this paper can reduce the compression time and increase the speedup ratio greatly. If an existing fractal rapid serial image compression method is applied to this algorithm, the speedup ratio is much higher than the global search. In this paper, we use neighbor search of domain blocks which are located in the six steps scope of the range. The experimental data is shown in table3.

TABLE 3. COMPARISON OF THE EXPERIMENT RESULT BETWEEN THE PROPOSED ALGORITHM BASED ON THE NEIGHBOR SEARCH AND THE SERIAL ALGORITHM BASED ON THE FULL SEARCH

algorithm	**Lenna image**			**Movie image**		
	PSNR	*Running time (second)*	*Speedup ratio*	*PSNR*	*Running time (second)*	*Speedup ratio*
Single serial algorithm（global search）	34.360	1934	—	33.804	4809	—
Algorithm of this paper（neighbor search，2 computing nodes）	33.235	4	483.5	31.764	6	801.5

From table 3, we can find that t when we use neighbor search of domain block based on this parallel algorithm, the speedup ratio is 483.5 times for lenna image and 801.5 times for movie image.

VII. CONCLUSION

This paper proposes fractal image compression parallel algorithm based on Web Service technology. This algorithm can greatly reduce the compression time and achieve a high speedup at the same PSNR and compression ratio. Moreover the WDCS system of this paper has good expansibility owing to Web Service. It can not only use easily some existing computers providing Web Service with different locations and platforms to achieve ability of high-performance parallel processing computer. But these computing nodes do not need to reconfigure the firewall, which can be free from infection by opening some tcp ports. It is very suitable for large-scale image or video compression.

REFERENCES

[1] M.Barnsley, L.Hurd. Fractal Image Compression. A K Peters, Wellesley, 1993.

[2] M. Wang, S. Zheng, W. Zheng. Improving codebooks in fractal image compression. Mini-Micro Systems, 23 : 350 - 353,2002. (In Chinese)

[3] Lai C., Lam K., Siu W.. Improved searching scheme for fractal image coding. IEE Electronics Letters, 2002, 38 (25): 1653~1654.

[4] Zhang Chao, ZhouYi-ming, Zhang Zeng-ke. Fast fractal image encoding based on special image features. Tsinghua Science and Technology, 2007, 12(1):58-62.

[5] C K Lee, W K Lee. Fast fractal image coding based on local variances[J]. IEEE Thans Image Process 1998, 7(6): 888-891.

[6] A E Jackquin. Fractal Image Coding. IEEE Thans Image Process 1993, 81(10).

[7] Yue Kun, Wang Xiao-Ling, Zhou Ao-Ying. Underlying techniques for Web services: A survey. Journal of Software, 2004, 15(3): 428-442(in Chinese).

[8] Jacquin A E. A fractal theory of iterated markov operators with applications to digital image coding. Georgia: Georgia Institute of Technology, 1989.

[9] Du ZX, Huai JP. Research and implementation of an active distributed web service registry. Journal of Software, 2006,17(3):454-462.

A New Carry-free Adder Model for Ternary Optical Computer

Yanping Liu, Junjie Peng*, Yuanyuan Chen, Hui He
School of computer of engineering and science
Shanghai university
Shanghai, China
jjie.peng@shu.edu.cn

Haitao Su
Beijing Benz-DaimlerChrysler Automotive Co., Ltd.
Shanghai, China

***Abstract*—Based on parallelism of MSD addition, a new carry-free adder model for optical computer is proposed. Using MSD number system and four kinds of logic transforms, the model is built with input part and three optical calculators. With the model, addition is implemented with only logic operations without any carry. This can greatly improve the efficiency of addition operation. To guarantee the feasibility of the implementation of the model, the alignment issue between the data bits of operands on different optical calculators is discussed. And the method to solve the problem that the input state of light is dark while the output state is bright in the process of logic transform is put foward. The analysis results show that the model is correct and can much improve the efficiency of the adder compared with the traditional adder model.**

Keywords- optical computer; MSD algorithm; adder model; alignment of operand; redundancy

I. INTRODUCTION

Modefied signed-digit (MSD) addition was proposed in 1961 by Avizienis for the first time. It is a kind of parallel algorithm based on MSD digital system[1,2] which presents information in redundant digital form. The characteristic of MSD addition is that it is a carry-free method which means it can avoid carry propagation between the adjacent data bits in the process of addition operation. This is a big advantage over traditional addition methods and implementation. It can infer that in MSD addition, what affect the efficiency of addition operation most is not the number of bits of operands as other addition methods do but the MSD addition's logical steps.

Space parallelism is one of the absorbing characteristics of optical operation, which guarantees that addition operation in optical computer can use MSD schema to exploit the advantage of optical computing. While the development of ternary optical computer (TOC) has made great achievement, the applications of it is still confined to logical operations for some reasons, among which lack of adder is one of the main hinders. Of course, following the steps of MSD transforms, addition can also be implemented on the ternary optical computer. However, it at least needs to transform three times in sequence to execute a addition operation. What is more, these three transforms are complete in serial. That is, the inputs of the latter transforms are the feedback of the results of the former operations. This makes the addition operation be much inefficiency.

As addition operation is the basis of numerical computing applications, it is very necessary to develop an efficient adder for ternary optical computer to proceed additional operation. This paper is try to present an efficient adder model that can be used in TOC to promote its use not in logical operations but also arithmetic areas.

II. BASIS OF THE RESEARCH

Before discussing the new adder model proposed in this paper, it's necessary to give a brief introduction of two points both of which are the basis of the study. One point that need to be mentioned is ternary optical computer's operational unit[3] and the other one is MSD additional algorithm[2].

A. TOC's Operational Unit

The conception of TOC[4] was proposed by Professor JinYi at Shanghai University firstly in 2000. Since then, the group led by Prof. Jin himself have set much effort on the study of Ternary Optical Computer and much achievement[5,6,7,8] has been made. Among the numerous achivements, the successful development of the experiment system with thousands of bits is a big marvel. With the experiment system, all optical operations can be realized through reconfiguration[3] of the operational unit according to the truth table user provided. The structure of basic unit which is used to make up of logical operational unit is demonstrated in figure 1. Where Lcd array is used to proceed lights passing through and polaroid signed p1 and p2 is exploited to serve as selector of passing lights. The polarization of light passing through LCD is controlled by the control signal put on the LCD. That is, light might changes polarization after it passes through the LCD if only the control signal is suitable. There are two kinds of polaroid, one is horizontal type and the other vertical one. Horizontal Polaroid only allowes horizontal polarized lights pass through. For the same reason, vertical polaroid just enable vertical polarized lights pass through. According the theory of Decrease-Radix Design principle, a bit of operator for any kind logical operation can be implemented by several basic units presented in figure 1.

* Corresponding author：jjie.peng@shu.edu.cn

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.6

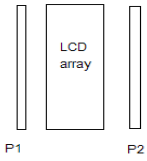


Figure 1. Structure of basic unit.

B. MSD Parallel Additional Algorithm

MSD is a kind of redudant digital number system[1,2]. Based this number system, MSD addition algorithm can implement parallel addition operation only with three carry-free logic operation. In this paper a new carry-free optical addition's model is presented based on MSD addition algorithm together with newly developed TOC.

Table 1 is the truth table of the four kinds of logical operation used in MSD addition which are operation T, W, T' and W'. Where -1 in table 1 represents vertical lights in the process of optical addition operation. 0 represents dark state and 1 horizontal light state respectively.

As more details of the discussion about MSD addition please refer literature [2] for reference.

TABLE I. FOUR KINDS OF TRUTH TABLE FOR MSD ADDITION

T a\b	-1	0	1	W a\b	-1	0	1	T' a\b	-1	0	1	W' a\b	-1	0	1
-1	-1	-1	0	-1	0	1	0	-1	-1	0	0	-1	0	-1	0
0	-1	0	1	0	1	0	-1	0	0	0	0	0	-1	0	1
1	0	1	1	1	0	-1	0	1	0	0	1	1	0	1	0

III. EASE MODEL OF ADDER FOR TOC

A. Carry-Free Adder's Model

The architecture of the adder model we proposed is demonstrated in figure 2. From the figure, it is not hard to figure out that the adder is physically consisted of four separate parts while logically in two parts. The left of the adder is the input part. It answers for light input and optical encoding. The right part of the adder is the addition operation. It consists of three group of sequentially arrayed logical operational units. The logical operational units of the adder is implemented with LCD, polaroid and light sensitive cell. The function of the encoding part is try to encode the natural light inputted to ternary output light. The encoded light will be used as the input of addition operators. The first logical operational units implement T operation and W operation at the same time, the second part proceeds T' and W' operations and the third realizes T operation. The function of additional embedded system is to create LCD's control signal and send them to LCD.

B. Determination of the Result of Logic Operation

According to theory of Decrease-Radix Design principle[3], one bit of operand needs a HH type operation unit and a VV type operation unit to complete T' operation. Signal H means horizontal type polaroid while V means vertical type polaroid. In HH type operation unit both p1 and p2 are horizontal polariods .

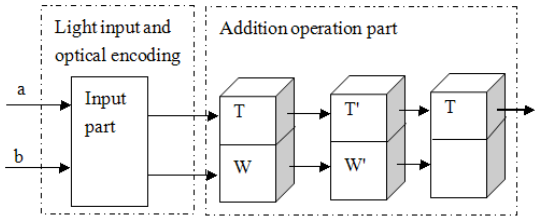


Figure 2. Adder's Model with Carrying Free Character.

From the output state of light sensitive cell which is expressed with 0 or 1, decoder can judge the result of logical operational units. Signal 0 and 1 mean dark state and bright state respectively. Based on the signal and the type of polaroid before crystal, the decoder can decode result of the light sensitive pixel. If output of one light sensitive spot is 1 and polaroid before the spot is V type, the result of the operational unit is vertical light. If the output of light sensitive spot is 1 and the polaroid before the spot is H type, the result of the spot is hoizontal light. And if the output of one light sensitive spot is 0, the result of the operational unit is no light.

The result of one bit operand for T' operation is determined by the results of one HH type operation unit and one VV unit. If the result of HH operational unit is horizontal light, one bit operand's result of T' operation is horizontal light. Else if the result of VV unit is vertical light, one bit operand's result of T' operation is vertical light. Otherwise this bit operand's result of T' operation is no light.

C. Adder's Working Process

Based on MSD additional algorithm and TOC, the working process for carry-free adder is as follows.

- Input operand and obtain the input light after it passes through the encoder. The input light can be expressed by a in the transform T and W of the truth table. In the process of addition, one of the most important factors is the control signal exerted on LCD. This signal is sent out from the embedded system. It is much related to the addend and the value of b in the truth table. With the direction of suitable signal, we can obtain the results of operation T and W.
- The input of the operation of T' and W' is the output light of operation T and W, they are expressed by a in the truth table. The control signal on LCD in the process of operation of T' and W' is much related to the value of b in the truth tatble. The second logic operation part is mainly used to implement the operation of T' and W'.
- The input of the third logic part is the result of the second logic operation part. That is, the result of T'

is expressed with value a in the truth table. The control signal on the LCD of the third logic part is much related to the result of W' operation and the operator b in the truth table. The third logic operation part is to implement the operation of T.

- The result of the third part is sent to the embedded system and then returned to user as the final result of addition.

IV. Alignment Issue between the Data Bits of Operands on Different Optical Calculators

It is necessary to solve the problem of alignment between the data bits of operands on different optical calculators, because after T operation operant result of location i need to be located at i+1.

The alignment of hardware for T, T' and T operation on three part presents in figure 3. Line a* means that the operant a has been shifted left for two bit with two zeros complemented on the right. This indicates that the number on the place of a1 is right most number before the shift. Line T gives the rule how the number of the result will be arranged after the operation T on the first operation part. Line t' is the rule how the number of the result will be arranged after the operation T'. The bottom line T is the arrangement of the number of the result after the operation T executed on the third logic part.

a*	a2	a1	0	0
T	t3	t2	t1	0
T'	t'4	t'3	t'2	t'1
T	t4	t3	t2	t1

Figure 3. Hardware's Alignment.

It can see from figure 3 that a1 is the lowest bit of the addend of operation T. The control signal is determined by the lowest bit of addend b. T2 is the output of of the result of the addition of the lowest bit. The result of the first bit of operation T is placed at t2. With this rule, the alignment of bits will be implemented following the formula $t_i = T(x_{i-1}, y_{i-1})$. To the operation T', the input is t2, the result of the operation T. The control signal is created by the sencond bit of operation W. And it's result is symbol t'3, the third bit of operation T'. With this rule the alignment is implemented $t'_i = T'(t_{i-1}, w_{i-1})$. Operation T on the third logic operation part is something like that on the first logic part. Its input is symbol t'3, the control signal on LCD is determined by the third bit of operation W' and it's result is t3. This can implement the alignment of $t_i = T(t'_i, w'_i)$.

V. The Realization of One Bit Addition

A. Related issue

When observing the truth table of T and W' operation, it's not hard to find that there exists a case that input a is no light while output has light. In T operation's truth table, a and b correspond to T operation's input and LCD's control signal respectly. Result is 1 when a is 0 and b is 1 which means the input is no light but the result is horizontal polarized light.

It's impossible that result is bright while its input is dark according to the physical characteristics of lighting device. The TOC's traditional solution is try to change no light state input into states with light through logical conversions before sending to operational units. For in carry-free adder, result of operating unit is put into the next operating unit directly as input, there has no opportunity to do logical conversion. Traditional method is useless and new solution for this issue is needed to be researched.

B. Redundant light method

When analysing logical operations of MSD addition, we have two conclusions, one is that the above issue only present in operation T and W' and the other is only when the input a and b are both 0, the result is 0. So redundant light route method is considered to solve this issue.

Two models of W'1 and redundant W'2 work together to accomplish the function of W' operation. W' operations with a being non-zero in the truth table is implemented by W'1 model. In these operations, the input and control signal of W'1 is the result of W operation and determined by the result of T operation. The other operations in the truth table, that is, W' operation with a being 0 is implemented by W'2 redundant model. In these operations, the input and control signal of W'2 is the result of T operation and determined by the result of W operation. The principle of W' operation presents in figer 4 where module 1 and module 2 implement W operation separately on the same time and module 3 and module 4 implement T operation separately in synchronization. So the inputs of one-bit W' operation is the results of two separate W operations and two T operations. That is, one part input is the result of two W operations and the other input is the result of two T operations. In the real application, the result of W operation is to implment the operations with a being non-zero in the truth table while the result of T operation is to fulfill the operation with a being zero in the table. To implement W' operation, two logic operation units are needed, one is HV type unit and the other is VH type unit. The results of two T operations and W operations are the inputs of both HV and VH type units. The result of the operation of module 1 and module 2 in figure 4 serves as the input of module W'1 while the result of the operation of module 3 and module 4 serves as the input of the redundant module W'2. Both W'1 and W'2 are to implement W' operation.

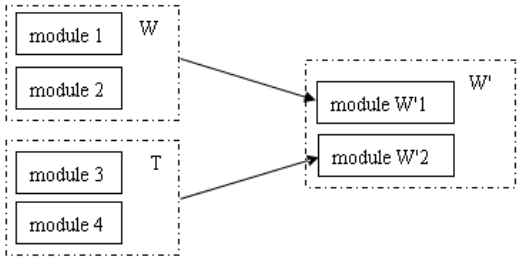


Figure 4. Physical structure of W' Operation.

To the T operation on the right most of figure 1, the inputs a and b are the results of T' operation and W' operation respectively. And it is implemented by T1 module and redundant module T2. Where T1 tries to implement the operations with a being non-zero in truth table while the result of T2 operation answers for the implementation of the operations with a being zero. In the process of operations, control signal in T1 and T2 is produced by the result of W' and T' operation respectively. The principle of T operation is as shown in figure 5 where module 1 and 2 in the upper dotted line box answer for T' operation while module 3 and 4 in the bottom dotted line box try to implement W' operation. The results of module 1 and 2 are the two inputs of module T1 and the results of module 3 and 4 are the two inputs of module T2. Both T1 and T2 are to implement T operation.

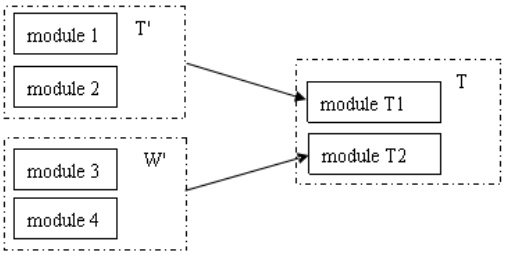


Figure 5. Physical Structure for T Operation.

In part one, a bit operand of T operation needs one VV and one HH logic operation units. A bit operand of W operation needs a VH and a HV logic operation units. In part two, a bit operant of T' operation needs a VV and a HH logic operation units. A bit operant of W' operation needs a VV and a HH logic operation units. As mentioned above, the case input is dark state while output is bright is impossible. This issue can be solved by redundancy of light path. Combine W' operation and redudant light path of T operation, the structure of one bit operand is as figure 6 shown.

VI. ANALYSIS OF THE MODEL

To implement MSD addition, traditional method needs to reconfiguration the logic operation units and feedback the temp results in the transforms at least three times[9]. This makes the addition operation is very inefficient. Using the carry-free adder model proposed in this paper, in the process of addition no data feedback of the temp results is needed and only one reconfiguration of the logic operation units is needed. This means that the extra time in the process of addition in the proposed method is much less than that of the tradition method. That is, the new method is more efficient over the traditional method.

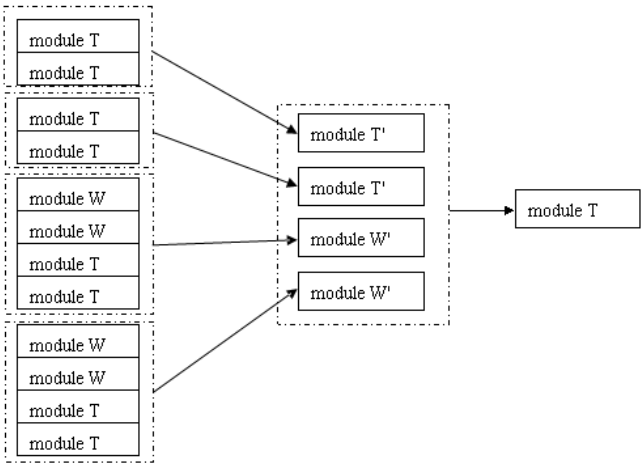


Figure 6. Physical Structure for Addition's One Bit Operand.

Table 2 presents a brief comparison between the implementation of the proposed optical adder and that of the traditional method. From the table, it is not easy to find that all the operations such as result image capture, decoding, data feedback, reconfiguration of logic operation units and so on need to be executed three times in traditional method, while only need to be executed once at most in the new proposed method. It can infer that the efficiency of the adder following the proposed method will be at least twice that of adder following the traditional method.

TABLE 2 COMPARISON OF THE PROPOSED METHOD WITH TRADITIONAL METHOD

Items \ Method	Traditional method	Carry-free adder
Result image capture times	3	1
Decoding times	3	1
Feedback times	3	0
reconfiguration times of logic operation units	3	1

VII. CONCLUSION

Carry-free adder's model is proposed in this paper. Based on MSD additional algorithm this model can process additional operation for the numbers with huge number of bits and it can fully exploit the parallel character of optical computer. Compared with traditional method it is more efficient for addition operation. For using the model implement addition, no decoding and feedback is needed. The analysis results show that this model is correct and is better than traditional method. The future work is to realize

the optical adder under the direction of model, so that the TOC can better be used in numerical computation area.

ACKNOWLEDGMENT

This work was supported by Shanghai Leading Academic Discipline Project (No. J50103), the Key Laboratory of Computer System and Architecture (Institute of Computing Technology, Chinese Academy of Sciences), the Ph.D. Programs Fund of Ministry of Education of China (No. 20093108110016) and Innovation Program of Shanghai Municipal Education Commission under grant (No.11YZ09).

REFERENCES

[1] Algirdas Avizienis. Signed-digit number Representations for Fast Parallel Arithmetic[J]. Electronic Computers, 1961-9,10;389−400.

[2] Richard P. Bocker, Barry L. Drake, Mark E. Lasher, and Thomas B. Henderson, Modified signed-digit Addition and Subtraction Using Optical Symbolic Substitution[J]. Applied Optics, 1986-8-1,25(15):2456−2457.

[3] Junyong Yan, Yi Jin, Kaizhong Zuo. Theory of Decrease-Radix Design Principle and its Application on Ternary Optical Computer[J]. Science in China, 2008-12, 38(12):2112-2122.

[4] Yi Jin, Huacan He and Yangtian Lv, Basic Principles of Ternary Optical Computer[J]. Science China(Series E), 2003-2, 33(2): 111-115.

[5] Xianchao Wang, Junjie Peng, Mei Li, Zhangyi Shen, Ouyang Shan. Carry-free Vector-matrix Multiplication on a Dynamically Reconfigurable Optical Platform[J]. Applied Optics, 2010-4-20, 49(12):2352-2362.

[6] Yi Jin, Huacan He and Lirong Ai, Carry in Principle of Ternary Optical Adder[J], Science China(Series E), 2004, 34(8):930-938.

[7] Chao Cai, Yi Jin, Jiulong Bao and Yutao Wang, Design of Balanced Ternary Half-adder in Ternary Optical Computer[A], Computer Engineering, 2007-9, 33(17):278-279.

[8] Weigang Huang, Yi Jin, Lirong Ai, Junyong Yan and Hao Sun, Design and Implementation of the 100-Bit Coder for Ternary Optical Computers[A], Computer Engineering and Science,2006, 28(4):139-142.

[9] Mei Li, Huacan He, Yi Jin, Xianchao Wang. An Optical Method for MSD Addition[J]. Photon Joumal, 2010-12, 39(6):1053-1057.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

A Workflow-based Cooperative Project Management System

Yi Chen, Kun Hou, Rui Wang
College of Computer and Information Engineering,
Beijing Technology and Business University,
Beijing, P.R.China
esteechenn@gmail.com

***Abstract*—In modern large enterprises, a complex project is usually composed of a series of processes and needs many persons with different roles cooperating with each other within limited time and investment. A workflow-based cooperative project management system (CPMS) that can support multi-user working is presented in this paper. The system is based on Browser/Server architecture that contains User Interface, Business Logic and Data Access three layers in which the core is an XPDL-based workflow engine. A series of cooperative mechanisms such as inclusive gateway, exclusive gateway and parallel gateway are designed to achieve the definition, configuration, and control of the workflow, the assignment, submission and approval of project tasks, the allocation and configuration of project resource. The version control techniques for workflow designs, project plans and project documents are also presented. CPMS has been used in an international enterprise that needs to manage thousands of projects each year. The application result demonstrates that the system can satisfy the needs of cooperative project management.**

Keywords—Workflow, Project management, Cooperative work, Visualization

I. INTRODUCTION

The goal of project management is to seek the unity of more tasks, fast process, low cost and high quality [1]. Time, cost, quality, and organization are essential elements for ensuring the successful implementation of a project. In modern enterprises, there hundreds even thousands of projects need to be managed. And a project may go through months or even years from start to the end, involve dozens or even hundreds of participants, and are divided into many tasks. Usually, a complex project needs many people distributed in different places to work cooperatively according to a certain workflow. Traditional handcraft management methods have not satisfied these requirements. Project management system based computer and network techniques is an effective way for managing large amount of projects.

Cooperative work is an important issue in project management. Reference [2] has mentioned three major dimensions of cooperative work in project management which include awareness, articulation, and distribution. Besides these three dimensions, we consider that there are another three aspects should be considered.

- Authority Control: As a project is completed with many individuals or departments cooperatively. It is inevitable that there will be overlaps among the tasks. Also different person plays a different role in a project and has different authority to do related task. So each participant of project should be assigned related authority in order to ensure the correctness of project progress.
- Tasks Cohesion: Each Individual involved in the project is asked to complete his certain task. In order to implement the smooth transition from one task to another correctly, a smooth interface between tasks is needed [3].
- Process Control: Process control is the core of project management. Every step of the project process should be ensured to proceed smoothly. For example, infinite loop, abnormal terminating in a project process should be avoided. Also a process can not be modified by the different people at the same time [4].

This paper presents a workflow-based cooperative project management system (CPMS) which can push the processed projects forward according to a certain workflow. CPMS can assign tasks of project to the manager and participant's work lists on time, let project participants in different departments or different regions work cooperatively, and assign various kinds of resources such as human resources, raw materials etc. The goal of CPMS is to make project management reach high efficiency, high resource utilization, and high-quality.

II. THE DESIGN OF CPMS

CPMS applied B/S (Browser/Server) three-tier architecture which is popular recently for the large software based on Internet. Each layer in CPMS has corresponding functions according to the requirement of cooperative enterprise project management. These functions are divided into many modules. Each module is relative independent and has different functions. Elements of each module work closely together to realize the function and achieve the requirement of "high cohesion and low coupling".

A. The Composition of CPMS

CPMS is based on .net framework under the Windows platform and B/S three-tier architecture. Personnel within enterprise can directly access the system via the Intranet, while the one outside the enterprise can access it via the Internet through VPN or encryption by authorized certification ways.

B. The Architecture of CPMS

The B/S three-tier architecture of CPMS is shown in Figure 1. It is composed of Data Access Layer (DAL), Business Logic Layer (BLL), and User Interface (UI). The three-tier architecture has advantages of zero maintenance

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.20

for client-side, high safety, good scalability, code reusability. It also can support distributed computing and be conducive to standardize.

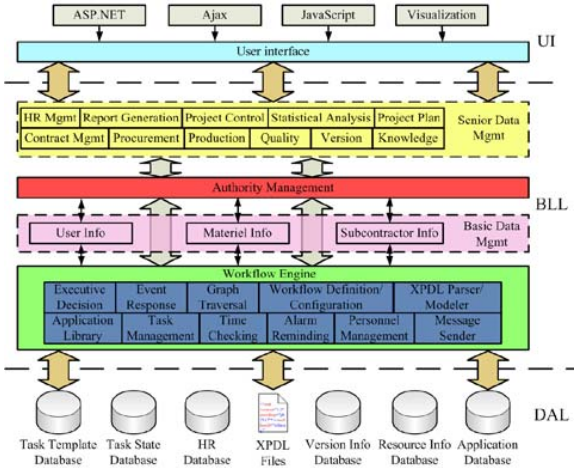


Figure 1. B/S three-tier architecture of CPMS

Data Access Layer: In this layer, CPMS mange accesses of databases. There are several databases, including task template database, task state database, human resources database, XPDL （XML Process Definition Language）file database, version information database, resource information database and application database. Task template database is used to store the definition of activities in workflow. Task state database is used to store the stage of the workflow and the state of activities. Human resources database is used to store the employee's information who is assigned tasks. XPDL files are used to store the order and logic of various activities in workflow. Version information database is used to store the version information of all the documents which are used in the project. Resource information database is used to store material information, subcontractor information, customer information and other basic data. Application database is used to store task information submitted by each step. All databases above will provide basic data for Business Logic Layer and User Interface Layer.

Business Logic Layer: In this layer, CPMS processes business logic and data. It is composed of four parts, i.e. workflow engine, basic data management, authority management and senior data management. Workflow engine implements workflow definition, configuration and implementation by using XPDL Parser/modeler module, decision-making module, time checking and alarm reminding module, task management module, personnel management module and workflow definition/configuration module. The basic data is managed by user information management module, materiel information management module and subcontractor information management module. Authority management module supplies user rights configuration functions and rights protection for definition and executer of workflow. It also provides rights guarantee for senior data management. Senior data management is composed of contract management, procurement management, production management, quality management, version management, knowledge management, human resources management, report generation, project control, statistical analysis and project plan management. It takes workflow as masterstroke and is based on basic data management and authority management. It provides users with a higher level of project management capabilities.

User Interface Layer: It is an interface between user and CPMS. It is what the user will see when he or she uses CPMS. In this layer, CPMS uses data that is provided by data access layer, accesses the function modules which are in business logic layer, and displays the final result to user. It uses ASP.NET [5], Ajax, JavaScript technology and visualization to improve the user-friendly interface and reduce the burden of the server.

III. COOPERATIVE WORKFLOW ENGINE

Cooperative workflow engine is a core part of CPMS. It is composed of executive decision-making, event response, graph traversal, workflow definition/configuration, XPDL Parser/Modeler, application library, task management, time checking, alarm reminding, personnel management, mail/message sender. Cooperative workflow engine implemented the main functions of project management, including definition, modification and execution of workflow, alarm reminding participants to process events through email or short message, and visualization of workflow design.

A. Executive Decision-making

Executive decision-making module is the core part of the cooperative workflow engine, which determines the status of the workflow. When the workflow engine receives an event, such as a keyboard event from the interface of workflow, event response module will call and pass the parameters of the event to executive decision-making module. Executive decision-making module uses the parameters to decide which activity should be activated. After above operations, executive decision-making module calls task management module writing those new activated activities into the task list.

B. Workflow definition / configuration

Workflow definition/configuration module implements visual definition of workflow. The visualization of workflow is mainly through conversion between XPDL and VML (Vector Markup Language). When users save the definition of workflow, this module will present the VML which is returned by user interface to XPDL Parser/Modeler module to converse and validate it. XPDL will be converted to a visual graphical interface when workflow is edited.

C. XPDL Parser/Modeler

XPDL Parser/Modeler module is a core module to implement definition and format conversion of workflow. The module will translate the abstract data of XPDL documents to cross-linked list, traversing them to verify the correctness of workflow diagrams. When XPDL is saved, the definition of workflow and the validated document of XPDL will be saved. Meanwhile, this

module also provides services for the executive decision-making module.

D. Application library and Task Management

Application library and Task Management is a core module to join workflows together. Executive decision-making module determines the activities which will be activated. It will check each processing mode of activities, which could be two types, automatic processing mode and manual handling mode. The activity will be submitted to relevant procedure of application library under the automatic processing mode. After the activity has been disposed, application library will inform executive decision-making module to submit it. Executive decision-making module will call task management module manual handling mode. Task management module will look for the task template of the activity, fill in massages into task template and insert the task into task list. The other function of task management module is checking database and submitting the state of activity to executive decision-making module. Task list management module shows the activated tasks on the user's personal home page in accordance with the identity of the user.

E. Time checking and alarm reminding

Time checking is a relatively independent module. The key technique in this module is thread. When time checking thread starts, it checks task list to find whether there is any task achieving the alarm time limit. If alarm time reaches the limitation, alarm reminding module will send alarm messages to related person, otherwise time checking thread will go to sleep. Alarm messages can be sent to related persons by e-mail, short massage and notice in personal homepage. So the project participants can get the alarm information from their email boxes, mobile phones and personal home pages.

IV. THE COOPERATIVE MECHANISM OF CPMS

In order to make workflow processed smoothly under the situation of multi-participant's operation, it is necessary to guarantee cooperative work among different departments and multi-participants. Some cooperative mechanisms are described in this chapter, including branch, parallel and approval mechanisms for implementing the correct performance logic, version management mechanism for keeping the information synchronized among multi departments and participants, and project plan management for defining and configuring project plan.

A. Branch, Parallel and Approval Mechanism

There are many branch, parallel and approval processes in the cooperative work among employees. So the most important function of cooperative workflow is that the executive track of project management can be automatic or man-made changed.

Branch is a basic structure of workflow. It is used to implement functions of path choice of workflow, parallel work, duplicate approval, and other functions. So divaricating and convergence of branches should be processed appropriately. The gateways of XPDL are used to implement functions of branch decision-making, parallel and approval in CPMS. There are four basic gateways [6] in XPDL, shown in Table Ⅰ.

TABLE I. FOUR GATEWAYS OF XPDL AND THEIR APPLICATION IN CPMS

Name	Icon	Function	Application in CPMS
Inclusive gateway		To make sure successor activities whether should be activated or not by independent condition, that is each successor activity is independent.	Complete the branch decision-making of cooperative work
Exclusive gateway		Using certain condition to determine which successor activity should be activated, that is successors are exclusive with each other.	Complete the approval of cooperative work
Parallel gateway		Activates all the successor activities.	Complete the parallel of cooperative work
Complex gateway		Allows complicated custom behavior.	Not use yet

Workflow engine defines workflow with XPDL. In it, inclusive gateway and exclusive gateway are considered as inclusive branch point; parallel gateway and complex gateway are considered as parallel branch point. In CPMS, Inclusive gateway completes judgment of branch decision-making. Exclusive gateway completes judgment of approval. Parallel gateway completes judgment of parallel.

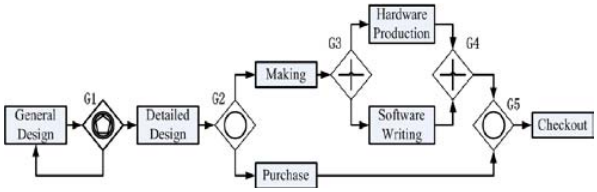


Figure 2. An example of the gateways application

Figure 2 is an example of using gateway to complete branch decision-making, parallel and approval. When a "general design" is submitted, G1 will be activated and inform handler. If the documents saved in "general design" are satisfying, the handler of G1 selects the path on the right side, letting the workflow enter into subsequent activities. If the documents are not approved, he or she can select the path going back to "general design" for modification. So an approval procedure is completed. After "detailed design" is submitted, G2 will be activated, handler will check "detailed design", if required material is enough, only activate the activity of "Making", else activate both activities of "Purchase" and "Making", so the branch decision-making of cooperative work is reflected in here. After "Making" is submitted, G3 will be activated and then "Hardware Production" and "Software Production" are activating at the same time. After both of "Hardware Production" and "Software Production" are finished, G4 will be activated, so the parallel mechanism is implemented in here. After all

activated activities are submitted, G5 will be activated, and its function is to re-merge the two branches at a vertex and jump to the next step "Checkout".

B. *Version management*

CPMS uses cooperative technology that break the traditional project management mode. It allows project manager, engineer, and customer and so on cooperates to complete a project design and execution. Version management is necessary for cooperative design system [7]. The function of version management is to let participants acquire effective relevant data in time and make sure version of data is latest. Another function of it is to keep the consistency of version information and provide effective modification of version.

Multi-version, dynamic, correlations are the three features of multi-participants cooperative work [8]. The implementation of version management is shown in Figure 3.

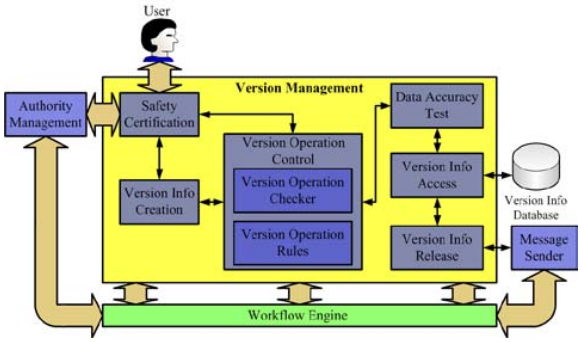


Figure 3. Implementation of Version Management

Safety Certification: When a user does some operations on documents, CPMS first need to check user's authorization on documents, such as read, write, copy, or delete. If this user owns the right to do this operation, this operation will be done. Otherwise, the message "exceed your authorities" will pop up on his screen, and no operation will be done. Safety certification module is to do what described above.

Version Information Creation: Version information of new design or new file will be created in this module. Version information is information mark of version in the system and the information basis of a series of version operations. It includes a unique version identifier and some other information about this document, such as version number, user information, version generated date.

Version Operation Control: There are operating rules and checker of version to take charge of version operation and data modification. The module makes use of the operating rules of version to constrain version operation. Such as, modification on document content can only be carried out on the copy one. If new version has not been approved as valid, the old one must not be removed. If there is conflict among versions, all versions will be tag as invalid. Version operation checker will check the version which is submitted by users whether its state information is same as which is saved in version information database. If the operation of the version meets the operating rules, it will be executed. Otherwise, the operation shall not be executed.

Data Accuracy Verify: It is responsible for maintaining the consistency of version. The system will establish a new version after the pending one is approved and the new one will be considered as the original version of the next step. Meanwhile the system must ensure that the new version replace the older one. Consistency of version can ensure the accuracy of data and avoid dirty data is written.

Version Information Access: It accomplishes the operations of version information and data, which like read, write, copy, and delete. In the processes of operations, the system will check integrality and rationality of version information and data.

Version Information Database: It is used to preserve the data of version information. It will record user's operations of reading or writing at each time. User can get back the older version or find out the processes of a project design by search version information database.

Version Information Release: It calls mail/message module to send the latest message of version to each participants. The purpose is to keep consistency of version and avoid dirty data is written.

C. *Project plan management*

Project plan management is another important part of CPMS. Microsoft Project has an absolute dominant position in project management system. In CPMS, there are exchange interfaces between CPMS and Microsoft Project through which data can be conversed with each other. The implementation of project plan management is shown in Figure 4.

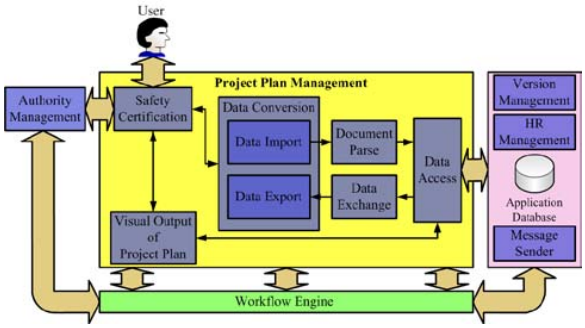


Figure 4. Implementation of Project Plan Management

Safety Certification: Safety certification module in project plan management provides safety certification for user's operation on project plan. Its principle is just the same as the one in version management described above.

Visual Output of Project Plan: This module calls data access module to show project plan information to users by combination of tree hierarchical structure and Gantt chart. The functions are extension and shrunk of project plan tree, edition of project plan on real-time, comparison of the execution progress and project plan have been implemented. The interface is simplicity and good interaction.

Data Exchange: This module is composed of data import module and export module. It is an interface module between CPMS and Microsoft Project. It imports the *.mpp file from Microsoft Project into CPMS and exports project plan data from CPMS into Microsoft Project.

Document Parse: This module calls methods of Microsoft.Office.Interop.MSProject Class to transfer Microsoft Project file into CPMS file and calls the data access module to save the file into database.

Data Conversion: It is reverse process of document parse. It calls methods of Microsoft.Office.Interop.MSProject Class to convert the project plan information into project file format of Microsoft Project. User can use these project files in their program.

Data Access: It accomplishes the operations of project plan, such as read, write, or delete. In the processes of operations, the system will check integrality and rationality of data.

V. THE IMPLEMENTATION OF CPMS

CPMS applied .net framework of Windows platform and developed with Microsoft Visual Studio 2008 and Microsoft SQL Server.

CPMS is based on B/S three-tier architecture. Through the ASP.NET, Ajax, JavaScript and visualization technology, it implemented the following functions, display and edition for cooperative workflow, decision-making branch, time checking and alarm reminding, display and edition for project plan, project progress monitoring and early warning, project data and document version management, contract information management, material purchase information management, explicit knowledge of the recording and management, human resource allocation and management. The visual edit interface for cooperative workflow design is shown in Figure 5. The visual interface for project plan and execution process is shown in Figure 6.

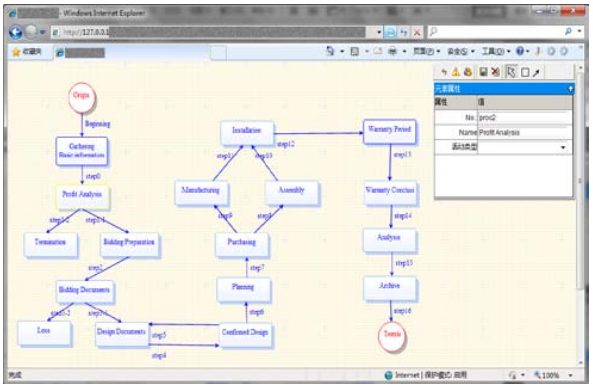


Figure 5. The visual edit interface for cooperative workflow design

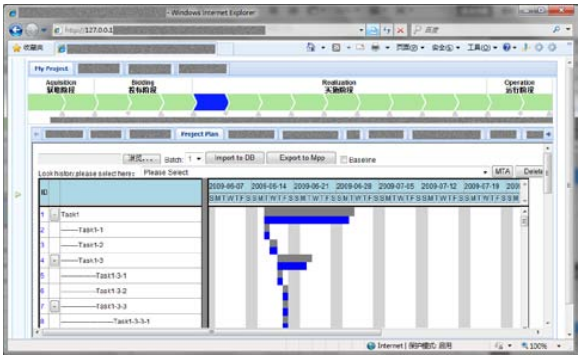


Figure 6. The visual interface for project plan and execution process

VI. CONCLUSION AND FUTURE WORK

CPMS is based on workflow and provides many functions such as visual workflow configuration, process control, task allocation, task submission and approval, project plan management, version management and human resources management. CPMS has already been used in an international enterprise that deals with thousands of projects each year. The application results demonstrate that CPMS can help the enterprise to manage large amounts of projects with high-quality, high efficiency and low cost.

The future works include following two aspects. First is to add more visualization functions. Second is to use Artificial Intelligence and Data Mining techniques to analyze task records, calculate the average execution time and the delay time of a project. Put all of the messages mentioned above on the workflow map, so that the user can get feedback from the map and optimize the business processes in later decision.

REFERENCES

[1] David I. Cleland, Roland Gareis. Global project management handbook. McGraw-Hill Professional,USA, 2006.

[2] Yi Chen, Jile Xin. A process-oriented configurable workflow system model for cooperative project management. The Fourteenth International Conference on Computer Supported Cooperative Work in Design (CSCWD 2010), 2010, pp.207-212.

[3] Greenberg, Marwood. Real time groupware as a distributed system: concurrency control and its effect on the interface. Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM Press New York, USA. 1994, pp. 207-217.

[4] Rav Anupindi, Sunil Chopra, Sudhakar D. Deshmukh, Jan A. Van Mieghem, Eitan Zemel, Sudhakar Deshmukh. Managing Business Process Flows: Principles Of Operations Management (2nd Edition), Prentice Hall,USA, 2005.

[5] Bill Evjen, Scott Hanselman, Devin Rader. Professional ASP.NET 3.5 SP1 Edition: In C# and VB, Wrox, UK, 2009.

[6] Workflow Management Coalition, Process Definition Interface—XMLProcess Definition Language Version 2.1a, Workflow Management Coalition 2008, WFMC-TC-1025.

[7] Rong Zhao,Weifeng Shi, Chao Zheng. Concurrency Control Mechanism in Cooperative CAD System, Computer Engineering, 2007,33(18): pp.106-108.

[8] Xin Jin, Jing Xu, Xuemeng Li, Fang Tan. Design and Implementation of XML-based Workflow Engine, Computer Engineering, 2007,33(23): pp.20-23.

A Distributed Management System With High Concurrency

YUNCHUAN LUO
China's ministry of culture

XIAO WU
China's ministry of culture

XIANGYE MENG
China's ministry of culture

YANJIE JIAO
China's ministry of culture

Abstract: The rapid development of computer network has greatly popularized online education, and the size of courseware is getting bigger. How to manage those resources is an urgent problem to be solved. This paper establishes a distributed courseware resource management system with P2P and CDN transmitting technology, providing highly concurrent VOD and distribution for numerous video courseware data, to meet the requirements of rising number of VOD users for online courseware. The system, like a huge memory, gives unified management for all courseware, and provides simple access interface. Meanwhile, the system provides management of users, authority, category and log, as well as searching, the administrator and students can look up the existing teaching scene and other references whenever and wherever possible.

Key words: courseware resource management; P2P; CDN

I. INTRODUCTION

Courseware is auxiliary teaching software providing relatively complete instruction of one or more knowledge points. Web courseware can run on the internet and be learned or downloaded through explorer.

Courseware management system is the main part of online education, being the premise and basis of integrating courses with information technology[1]. This paper puts forward a distributed architecture of courseware management system and illustrates the key feature and technique. This platform is running in Guangdong Study Website, providing great convenience for the site.

Online education saves much resource and manpower, and provides more options and flexible schedule for people in need of education. Thus, courseware can be used effectively and shared widely.

II. DESIGN OF DISTRIBUTED COURSEWARE MANAGEMENT SYSTEM

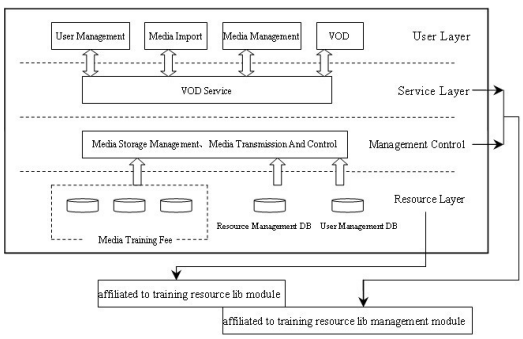


Figure 1 Architecture of the Courseware Management System

A. *Overview of CDN*

The core of distributed courseware management system is to establish a unified extendable architecture to enlarge user scope, realize the integration with learning system and guarantee access speed whenever and wherever possible. Under the principle of quick safe visit and less bandwidth occupation, configuration is implemented between nation centers and province centers, and among province centers.

According to the practical business process, this paper accomplishes strategy-based distribution and synchronization from nation centers to province centers, which ensures that the configuration of center resource is in accordance with the requirements of consistence, real-time and safety of business, and realize remote learning by multiple resource points supported by P2P, solving the problem of bandwidth shortage of single resource center and safe backup.

A unified resource management portal is built to exhibit learning resources of network academy concentratedly, provide effective resource for learning system and realize courseware VOD study based on P2P.

Unified management and application of resources are accomplished in the way that nationwide sharing learning resources are delivered to province centers by distributed resource management technique. Students are guided to access the entire resource according to different strategies, the pressure on nation and province centers can be balanced and students can visit nearby website conveniently.

The innovative courseware format released as stream file can enhance learning effect. P2P technique[2] is used to provide VOD and download service, with nation centers and province centers being data source, making each student a provider of learning resource while studying. This method takes full advantage of new network technique to solve the problem caused by dependence on servers and networks of nation centers and province centers while downloading and online learning, and enhance learning effect.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.3

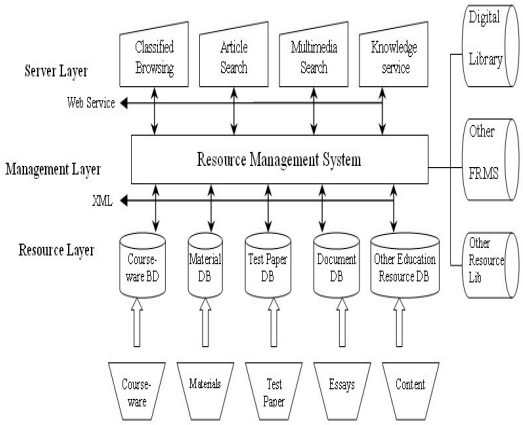


Figure 2 CDN Architecture

B. Design of Model

The operation platform is divided into 4 application function layers, namely resource service layer, resource management layer, operation management layer and operation maintenance layer.

1) Resource Service Layer

Resource service layer is service interface related to learning system.

Distributed courseware management system is designed with J2EE architecture, thus can be integrated with other systems conveniently and flexibly, guaranteeing the security and reliability of data interaction. It mainly provides interfaces for learning system as follows:

Resource selecting interface: providing optimal resource lists for administrators of learning systems in every province or branch institute according to settings of authority. While making up or releasing courses of learning systems, administrators can select learning resource of the course from the obtained resource lists that their province or branch institute is permitted to access through the interface, then the system will set the access path for the course automatically.

Interfaces are called in the form of Web Service, data are organized in format of XML, and transmitted data are encrypted with dual-key Rijndael symmetrical encryption algorithm.

Resource scheduling interface: scheduling requests for visiting or downloading from nation centers and province centers to a most suitable path for access, according to load and settings of resource accessing authority. When students request for accessing certain resource, the distributed resource management system will automatically compute the located server or server lists allowed to access and source server, then the intelligent scheduling system will return the path most suitable for access according to server load and network situation.

Resource transmitting interface: P2P is used in delivery and access of resource. Thus, resource is shared among distributing servers to accelerate transmitting rate while relieving pressure on central nodes and saving the cost of bandwidth of central backbone network.

2) Resource Management Layer

Resource management includes resource information management, uploading and releasing resource, deleting resource, valid status management of resource, resource distributing and sharing management.

Resource can be divided into 3 types according to location and access authority:

a) Nation center resources, which are distributed from nation centers to resource nodes of province centers, can serve all users,.

b) Province centers monopolize resource, and only serve users within the province.

c) Province centers share resources, which can be set by the administrator.

The distribution of resource includes pre-distribution and PULL-as-needed. Pre-distribution is that the distribution administrator of nation centers deliver resource to designated path of designated server according to distributing strategy and path, a PUSH way. PULL-as-needed is that when user visits certain resource, the intelligent scheduling server schedules it to an optimized server path. If the resource doesn't exist on the server, resource will be pulled from source server to that server in slices, which is a PULL way.

3) Operation Management Layer

Operation management layer provides statistical form, online management, settings for distributing resource and distributing path management of resource, etc.

Statistics of resource situation: according to resource centers and course-making time, recording course number, courseware number, courseware number contained by each course, course names, courseware names contained by courses, and generating statistical forms, which can be exported as excel file.

Record of utility status of resource: to record playing and downloading number of courseware according to time slot and resource centers, and generate statistical form able to be exported as excel file.

Record of resource distribution: to record the situation of courses and courseware delivered by distributed resource management system, including course name, courseware name, distributing time and the province center list to be delivered.

Distributing path management of resource: distribution administrator of province center sets the distributing path of resource within the province, according to servers and storage situation of the center. One resource distributing node can respond to more paths. One province center can have several resource servers, and each server can be set with several distributing paths. While distributing resources, they are delivered to one or more distributing paths on each server of province center.

Distributing strategy management: distribution administrators of nation centers make up distributing strategy, namely to select one or more proper resources and the one or more destination province centers, while choosing a distributing mode. If selected as auto-distributing, the distributing time should be set. When making distributing strategy, resources can be managed in the form of file or category, namely distributing

management of single file, multiple files and the entire category which can be netted with levels of subcategories.

Online management: to record currently logged-in user lists, displaying all presently online users and log-in time on different clients of the same user, connecting time, IP address of client machine, as well as the page being visited by users.

4) Operation Maintenance Layer

Operation maintenance layer includes traditional log management, safety management, flow management, data maintenance, authority setting, user setting and performance management, as well as information supervision, resource access supervision, resource website device supervision, etc.

Log management: to look up important operation logs of the system according to user or time. The logs should be backed up on a regular basis, and logs at designated time can be deleted when in need.

Data maintenance: data backup, transferring, load or unload, import and export of interface data.

Safety management: providing user identification based on CA; providing safe transmitting mode of SSL protocol; providing management of identification server.

Flow management: providing flow supervision and timely report, and access record of columns of each level, channels and website content.

Authority setting: providing allocation of managing authority on each level.

User management: managing admin-users of distributed resource management system in the form of role. Admin-users on the platform of nation centers include system administrator, distribution administrator, resource releaser, resource administrator. Admin-users of province centers include distribution administrator, resource releaser, resource administrator. User management is to create, modify, query and delete information of the above users, as well as authorizing them.

Server management: nation centers can manage servers of nation centers and province centers, so that the system can implement strategic distribution according to server status.

Performance management: providing performance-supervising report.

Supervision on networks and servers: supervising networks and servers of nation centers and province centers, to command and display the status of all nation centers and province centers, including service condition of networks and servers (CPU, memory and disks). Which application is occupying center resource, VOD and resource distribution can be displayed by a good user interface, showing normal, tips and alarm of servers. Data can be reserved according to certain time frequency, to be used in data analysis and system optimization.

C. Learning Procedure

The learning procedure is showed in the following figure, including 3 roles, students, learning system and distributed resource management system.

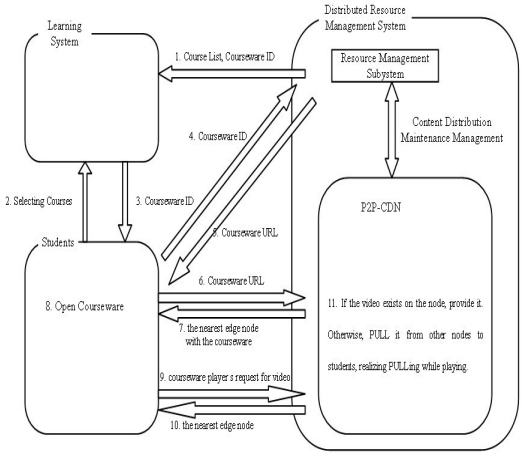


Figure 3 Learning Flowchart

III. KEY TECHNOLOGIES

A. P2P-CDN Accelerating Transmitting Resource

CDN has quick distribution speed and can allocate resources properly, ensuring that users can access the nearest good resources with high speed.

Transmitting data in the form of stream with P2P and CDN technique[3] makes the request of file block and transmission check more accurate, which can reduce lots of invalid requests and data delivery. In certain range, the more users visiting, the more stable is the system, avoiding the problem that servers are down and bandwidth shortage due to too many users requesting VOD of popular programs, to guarantee safety and stability of the system.

Clients can play video fluently and stably with P2P[4]. The more nodes connected, the better of the effect. Clients can exchange data, that is to say users can send data to others while playing courseware video. Thus, resource is shared between clients, and the pressure on servers to frequently and repeatedly process data can be reduced, to guarantee VOD effects. P2P can be managed flexibly, for example, P2P and non-P2P can be switched in complicated circumstances.

P2P used in edge nodes can realize mechanism of N+1 backup to ensure safety and stability of the system. As a result, even though one edge node is out of service, the system can still run normally.

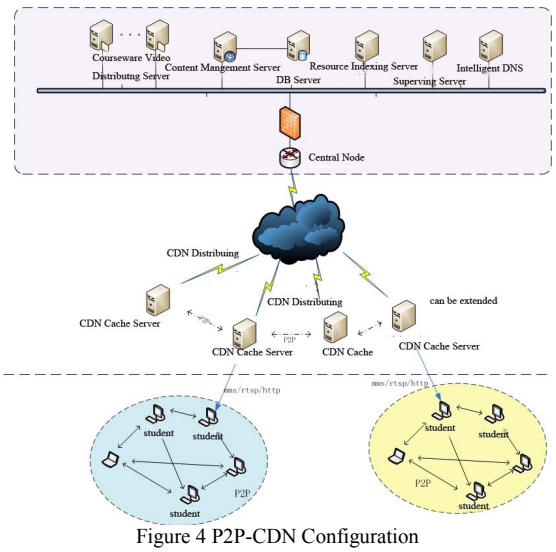


Figure 4 P2P-CDN Configuration

B. Intelligent Scheduling

Intelligent scheduling is mainly to balance global load of edge nodes and realize nearby visit.

First, a user gets access to learning site of stream media and select needed video. When the site receives user request, it will analyze the request and the origin of the user, then find the stream media server nearest to the user to provide service. That process can reduce waiting time of users. Moreover, because the server is nearest to the user and passes less nodes, it will occupy less resources of backbone network and lower the risk of package lost during transmission, in order to provide users with stable and fluent stream media.

Intelligent scheduling server correctly understand bandwidth and ISP type, load(CPU, memory and disk space) of each node to make sure that user request can be directed to proper node, balancing global load of the entire network.

When the load of server surpasses 80% , which can be customized, it will protect itself from responding to any new request. New users will visit other servers with light load.

If one edge node is down, users can quickly get access to others. Thus, user experience would not be affected, as they even can't feel the downtime, which makes the N+1 backup.

C. Stream Media Stored In Slices

With large video data, limited storage, when there is no video file on the nearest edge node, the edge node adopts PULL-as-needed strategy using sliced storage, sending one slice of data to the player once pulling and receiving it.

The total storage amount of sliced data can be set on each edge node. And when it is full, stored data is obsoleted with LRU algorithm.

IV. CONCLUSION

Distributed courseware resource management system is stably running with good performance and convenience.

The system can be installed and distributed onto different network nodes to form courseware resource net, ultimately satisfying demand of resource sharing.

[1] CELTSC. Specification of Learning Object Metadata, Beijing: http://www.celtsc.edu.cn/, 2003.

[2] CHRISTINA APERJIS, M. J. F., AND JOHARI, R. Peer assisted content distribution with prices. In CoNext'08: Proc. ACM SIGCOMM Conference on Emerging Networking Experiments(2008).

[3] Yang Chuandong. Research on Mixed Stream Media System With CDN and P2P. Computer Application, 2005.9.

[4] Stutzbach D,Rejaie R, Sen R.Characterizing Unstructured Overlay Topologies in Modern P2P File-sharing Systems[C]//Proc.of Intrenational Measurement Conference.[S.l.]:IEEE Press, 2005:267-280.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

A Parallel Packet Processing Method On Multi-Core Systems

Yunchun Li
Computer Science Department
Beihang University
Beijing, China
lych@buaa.edu.cn

Xinxin Qiao
Computer Science Department
Beihang University
Beijing, China
qiaoxinxin1126@gmail.com

***Abstract*— The demand of network packet processing is increasing as applications demand more and more bandwidth and computing capability. In this paper, the characteristics of packet processing and multi-core systems are analyzed. In order to analyze the differences of serial packet processing and parallel packet processing, the packet processing model is proposed. In the end, a DPI experiment on a multi-core system has been carried out to verify the analysis. The results show the proportion of the parallel part and the load balancing of computing resources affect performance.**

Keywords-prallel packet processing; multi-core; deep packet inspection

I. INTRODUCTION

With the rapid development of Internet, many network applications have appeared, such as firewalls, intrusion detection systems, encryption, traffic identification and traffic control. All of these network applications require higher network packet processing capabilities. As a result, people focus the packet processing capabilities on different systems.

Current network equipments can be divided into several categories: general-purpose processors, application specific integrated circuit (ASIC), heterogeneous and homogeneous multi-core network processors. Among these categories, multi-core network processors become the trend of network equipment for they have high performance and programmability.

While the proliferation of multi-core architectures has not expected fast packet processing, on multi-core systems, the following areas can affect performance, such as CPU starvation, task decomposition, CPU load balancing, and scalability of speedup. CPU starvation is mainly caused by lock competition of the sharing data. The task decomposition methods and the load situation of CPU will affect the efficiency of packet processing. In addition, if speedup can increase linearly when CPU cores increase linearly, it shows that the application has a good scalability.

In this paper, packet processing is divided into three basic stages: input, processing and output. To compare the differences of serial packet processing and parallel packet processing, a packet processing model is proposed to analyze the theory difference. Then we carry out some experiments, meanwhile the execution time and CPU load are recorded by using profile tools. Finally we analyze the performance of two kinds of packet processing and discuss the main factors affecting the performance on multi-core systems.

The remainder of the paper is organized as follows: In section II, Related works of parallel packet processing are introduced. In section III, we introduce the multi-core architecture and software architecture of packet processing. We build serial and parallel packet processing models in section IV. In section V, some DPI experiments are carried out, and then we analyze the performance. In section VI, we conclude the paper and propose future research directions.

II. RELATED WORKS

With the development of multi-core systems, parallel packet processing has drawn lots of interest. Some multi-objective evolutionary algorithms are designed to optimize the Intel IXP network processor [1]. In [2], they study the frame the next-generation internet architecture. Multi-core aware architecture is designed to exploit the processing power of multi-core systems [3, 4]. To improve the performance of packet capture, most capture packet technologies [5] use memory mapping based zero-copy techniques to carry packets from the kernel level to the user space. Researchers [6] have demonstrated that performance can be substantially improved by designing a novel multi-core aware packet capture kernel module which enables monitoring applications to scale with the number of cores.

Many researchers study the management of computing resources on multi-core systems. There are a rich set of literatures on accelerating DPI on multi-core systems. Some literatures [7, 8] focus on optimizing the representation of regular expression. In [9], a scalable multithreaded L7-filter algorithm is designed, it use affinity to improve performance. Programming of multi-core network processors is a challenge due to the complex interactions between the multiple processor cores, memory, and other shared components. Some approaches [10] provide a thin network processing operating system to simplify programming. Some systems provide the ability to quickly adapt multiple applications during run-time. [11] proposed an algorithm for adapting allocations of entire applications to processor cores to reduce power consumption. As computing resources are difficult to allocated balanced, runtime support for multi-core packet processing systems is designed [12].

III. MULTI-CORE SYSTEMS

A. Multi-core network processor architecture

According to architecture, multi-core network processors can be divided into heterogeneous and homogeneous multi-

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.11

core network processors. All cores in homogeneous multi-core processors are the same, such as Cavium OCTEON multi-core network processors. Cores in heterogeneous multi-core processors are different. This paper mainly studies parallel packet processing mechanism on homogeneous multi-core systems.

Homogeneous multi-core network processors' chip integrates a number of the same cores, and each core has its own private L1 cache. All cores share L2 cache and memory. There are some network processing units and hardware acceleration components, such as packet input unit, packet output unit, hardware regular expression engines. Fig. 1 shows the architecture.

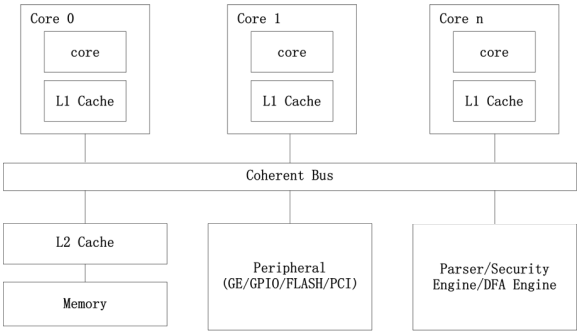


Figure 1. Homogeneous multi-core processor architecture.

To get high performance, we can use multi-thread, and let every core run at least one thread at the same time. But the threads are not the more the better. Too many threads' switch between cores will cause lots of cache misses, and these cache misses will reduce performance.

B. Software architecture

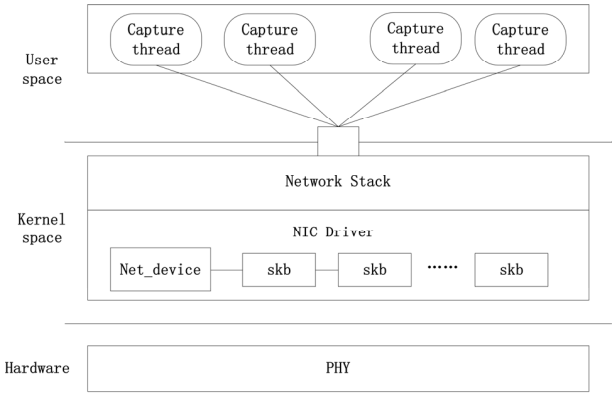


Figure 2. Network driver software architecture.

In this paper, Linux is used as an example, we introduce packet processing based Linux. As it shows in Fig. 2, when packets are received, packets are stored in a hardware queue. Then packets are sent to NIC driver. Normally these packets will be stored in a single queue. This design limitation is the major cause of performance bottleneck, because even if a user space application spawns several threads to consume packets, they all have to compete for receiving packets.

This competition will make packets' input in serial instead of in parallel. In the following we will analyze the packet processing mechanism, and our analysis is based on the single queue's design, so in this paper we only discuss the situation that packets are input or output in serial.

C. Deep Packet Inspection

DPI (Deep Packet Inspection) is mainly used to review the deep content of packets and classify the network applications. In spite of the simplicity and practicality of the classical port-based packet inspection, DPI is more and more widely used to meet the requirement of classifying P2P services in recent years. By matching against signature fields of various protocols, usually DPI applications use regular expression engine to obtain protocol type associated with the application layer data in the packet. In the following, a DPI experiment will be carried out to analyze the network packet processing mechanism.

IV. NETWORK PACKET PROCESSING MECHANISM

Packet processing is divided into three basic stages: input, processing and output. Assume time-consuming for the input stage is t_i, time-consuming for processing stage is t_p, time-consuming for output is t_o. And assume the other parts of the serial implementation such as initialization of hardware devices is a constant set as p, the CPU number is x, the total number of packets to be processed is n.

A. Serial packet processing

On serial packet processing systems, normally there is only one thread to deal with packets, and each packet is serially processed and each packet's three stages are serially executed like Fig. 3.

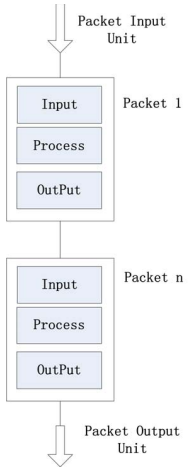


Figure 3. Serial packet processing.

The time-consuming of processing a packet is $(t_i + t_p + t_o)$ .The time-consuming of serial processing n packets is $n*(t_i + t_p + t_o)$.

The total run time of serial packet processing application is

$$T_{serial} = n*(t_i + t_p + t_o) + p\,. \quad (1)$$

B. Parallel packet processing

Lots of network packets transmit in the form of stream, and packets in some streams must be in order. In fact, to get higher performance, many people assign packets belonging to the same stream to the running queue of the same thread, which is dispatched to a dedicated core on a multi-core system. At the same time, usually we use multi-thread to let packets be processed in parallel. When a thread is running, the Linux scheduler timers interrupt periodically transfers focus to other threads. As it will increase cache misses and TLB misses, the thread switch will bring additional costs. In the packet processing model, we ignore the costs of thread switch. The mechanism of parallel packet processing shows in Fig. 4.

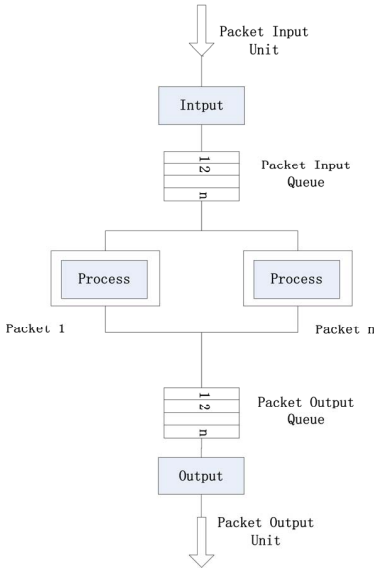


Figure 4. Parallel packet processing.

In this architecture the input packets will be stored in input queue, and then different steams will be dispatched to different cores, and their processing will execute in parallel. While the input queue is shared by different cores, usually we need to use lock to protect the input queue. These lock competition will bring delay. Assume the time for lock and unlock the queue is t_l, in input stage the total time for queue operation is $n*(t_l + t_i)$, time-consuming for parallel packet processing is $n*t_l + n*t_p/x + n*t_l$. In output stage the total time for queue operation is $n*(t_l + t_o)$.

The total run time of parallel packet processing application is

$$T_{parallel} = p + 4*n*t_l + n*(t_i + t_o) + n*t_p/x\,. \quad (2)$$

According to (1) and (2), the speedup is

$$S(n) = \frac{x*n*(t_i + t_p + t_o)}{x*(p + 4*n*t_l + n*(t_i + t_o)) + n*t_p}\,. \quad (3)$$

V. Experimental Results

A. Experiment platform

We choose Cavium Octeon 5860 [13] as our test platform. It has 16 MIPS cores, each core's frequency is 750MHz, and each core has 32KB instruction cache and 16KB data cache. 16 cores share a 2MB L2 cache. In the experiment the whole processing is divided into three stages, input, processing and output. DPI is used to process packets.

In the first experiment, the stream is processed in serial, and then we record the total time. In the second experiment we process the same stream in parallel and record time-consuming of each stage, record the CPU load. Then we analyze the differences of theoretical speedup and real speedup, and analyze the main factors that affect performance.

B. Experiment results

In the experiment, the input time-consuming t_i, the processing time-consuming t_p, the output time-consuming t_o and so on are recorded. According to (3), the theoretical speedup can be calculated. Meanwhile we record execution time of serial and parallel packet processing, and then calculate the actual speedup.

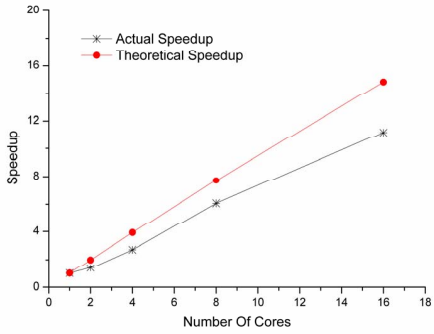


Figure 5. Theoretical and actual speedup.

As it shows in Fig. 5, speedup increases nearly in linear when CPU cores increase linearly. It illustrates when the number of cores is less than 16, the parallel packet processing mechanism has a good scalability. While we can see that the increase of the actual speedup is smaller than the theoretical speedup. To get the reasons, we use profile tools

to record the time of each stage and analyze their changes with the increase of cores.

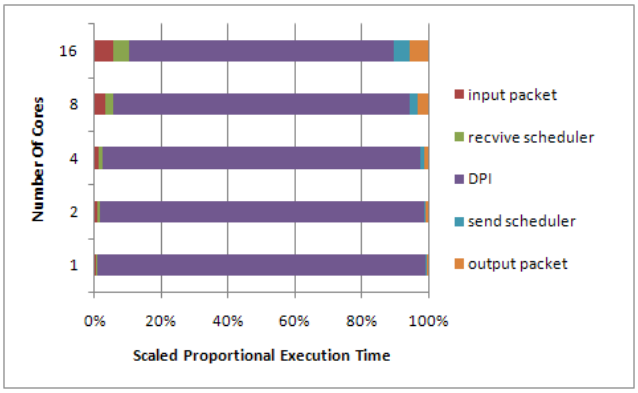


Figure 6. Scaled Proportional Execution Time.

Fig. 6 shows execution time proportion of each stage. The first observation from Fig. 6 is that the input and output incurs very limited overhead, while most of the execution time is for DPI. In the parallel packet processing experiment, the DPI stage is processed in parallel. With the increase of cores, the DPI's execution time reduces, and DPI's execution time proportion reduces. But the DPI's execution time does not reduce linearly as expected. So we use Oprofile tools to record the CPU load of each core. The CPU load of each core is showed in Fig. 7.

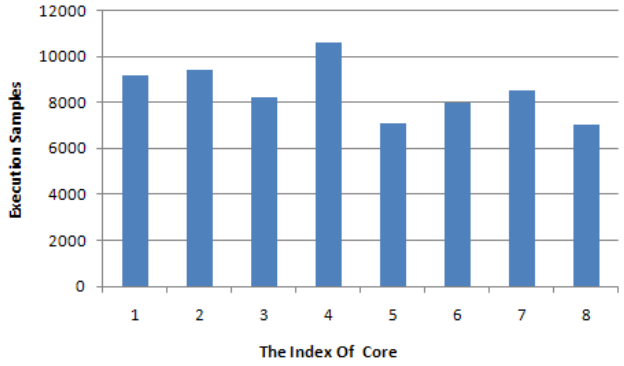


Figure 7. CPU load of each core.

There are 8 cores used to processing packets. As it shows in Fig.7, the load of each core is different. According to (2), the reduction of DPI's execution time is an average value, but in fact each core's load is not balanced, and the processing time will be the longest execution time of each core. So the DPI's execution time will be longer than the excepted time. That is the reason why the actual speedup is smaller than the theoretical speedup.

VI. SUMMARY

In this paper, we propose the packet processing model and compare the differences of serial packet processing and parallel packet processing. Then the DPI experiment is carried out. The results show it is possible to design and implement high-performance and multi-threaded parallel packet processing system on multi-core systems. Finally we analyzed performance, and found lock competition and CPU load affect the performance. In order to get higher performance, it is better to keep data independent and decrease the proportion of the serial part in packet processing application.

In the future, we plan to design a run-time system to support for multi-core packet processing systems. The run-time system can manage computing resources well, and reduce the lock competition.

ACKNOWLEDGEMENT

This work was supported by the National High-Tech Research and Development Plan of China under Grant No. 2007AA01A127 and No. 2010AA012404.

REFERENCES

[1] L.Noonan, C. Flanagan, An effective network processor design framework:using multi-objective evolutionary algorithms and object oriented techniques to optimise the intel IXP1200 network processor, 2006 .

[2] Xin Huang, S. Ganapathy, T. Wolf, "A Framework for Network State Management in the Next-Generation Internet Architecture," Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE, On page(s): 1 - 5, Volume: Issue: , Nov. 30 2008-Dec. 4 2008.

[3] V. Paxson, V. Sommer, and N. Weaver, "An Architecture for Exploiting Multi-Core Processors to Parallelize Network Intrusion Prevention," In Proceedings of IEEE Sarnoff Symposium, Nassau Inn, Princeton, 2007, NJ, 1—7.

[4] B. Haagdorens, T. Vermeiren, and M. Goossens, "Improving the Performance of Signature-Based Network Intrusion Detection Sensors by Multi-threading," In Proceedings of 5th International Workshop on Information Security Applications, Jeju Island, Korea, 2004, vol. 3325, 188--203.

[5] L. Deri, "Improving passive packet capture: beyond device polling," Proc. of SANE, 2004.

[6] F. Fusco and L. Deri, "High speed network traffic analysis with commodity multi-core systems," In: Proceedings of the 10th annual conference on Internet measurement. ACM. 2010, pp. 218–224.

[7] C. Hayes and Y. Luo, "DPICO: A High Speed Deep Packet Inspection Engine using Compact Finite Automata", ANCS 2007.

[8] P. Piyachon and Y. Luo, "Compact State Machines for High Performance Pattern Matching", DAC 2007.

[9] D. Guo, G. Liao, L. N. Bhuyan, B. Liu, and J. J. Ding, "A scalable multithreaded L7-filter design for multi-core servers," In Proceedings of the 4th ACM/IEEE Symposium on Architectures For Networking, 2008.

[10] G. Memik and W. H. Mangione-Smith. NEPAL, "A framework for efficiently structuring applications for network processors," In Proc. of Second Network Processor Workshop (NP-2) in conjunction with Ninth IEEE International Symposium on High Performance Computer Architecture (HPCA-9), Anaheim, CA, Feb.2003.

[11] R. Kokku et al., "A Case for Run-Time Adaptation in Packet Processing Systems," Proc. 2nd Wksp. Hot Topics in Networks, Cambridge, MA, Nov. 2003.

[12] T. Wolf, N. Weng, and C. H. Tai, "Run-time support for multi-core packet processing systems," IEEE Network, 21.(4), July 2007, 29-37.

[13] http://www.caviumnetworks.com/

A Web Page Classification Algorithm Based On Link Information

Zhaohui Xu[1,2]
1.School of Computer
Wuhan University of Technology
Wuhan,China
2 Grain Information Processing and Control Key
Laboratory of Ministry of Education
Henan University of Technology
Zhengzhou,China
Morsun@haut.edu.cn

Fuliang Yan
Grain Information Processing and Control Key
Laboratory of Ministry of Education
Henan University of Technology
Zhengzhou,China

Jie Qin
Grain Information Processing and Control Key
Laboratory of Ministry of Education
Henan University of Technology
Zhengzhou,China

Haifeng Zhu
Grain Information Processing and Control Key
Laboratory of Ministry of Education
Henan University of Technology
Zhengzhou,China

Abstract—**Effective classification of web pages can improve the quality of information retrieval. The traditional classification algorithms are basically based on the analysis of Web content, but the content of the webpage is complicated, filled with a large number of false, erroneous information, has seriously affected the accuracy of the classification of network information. To solve this problem, this paper presents a web page classification algorithm, Link Information Categorization(LIC). Based on the K nearest neighbor method, it combines information on the website features, to implement the Web page link to information classification. Experiments show that the algorithm can get higher efficiency and accuracy on the Web page classification**

Keywords- Web Page Classification; Link Information; Link Information Categorization

I. INTRODUCTION

Before the appearance of web pages, people have already carried on substantial research on the automatic text categorization[1,2]. With the rapid growth of web pages of the World Wide Web, automatic classification has been expanded from text to web pages and other network resources. Reasonable classification can help to organize and manage information , and help the users to find information quickly and accurately.

The main difference between the Web pages and text are the load of the information, and web pages to represent by html. Traditional classification algorithm is modified at the text preprocessing stage, it can be used to the classification of web pages. But the traditional classification algorithm has following disadvantages[2] :

(1) Excessively to dependent on the text content. The provider of web pages mixes a large amount of false information in the web pages for a certain purpose, such as a page, in order to improve its sequencing position in search engine, will usually be added the characters which have nothing to do the web content, and hided or revealed with the color of background. The traditional classification algorithm is incapable of that.

(2)When traditional classification algorithms to classify Chinese documents, Chinese word segmentation must be done to the text. But the websites changes rapidly, and new words followed each other. The emergence of a large number of specialized vocabulary makes the traditional Chinese Word Segmentation helpless.

(3)Text vector space dimension of the text is too large, amounting to thousands or even ten thousands, so it is difficult to make trade-offs between efficiency and accuracy.

(4)Most algorithms are only suitable for text classification, but they are useless to the videos, music and pictures.

This paper puts forward Link Information Categorization algorithm(LIC). It adopts the text link information to classify the web pages, and it can solve the above problems to a certain extent. The chapters are arranged as following: Chapter 2 analyses the existing text classification algorithms; Chapter 3 introductions the LIC algorithm and Chapter 4 we did the experiment and the comparison and analysis were carried out between the LIC algorithm and the traditional classification algorithms.

II. TRADITIONAL TEXT CLASSIFICATION ALGORITHMS

Some sophisticated text classification algorithms have been applied to Web Pages Classification, such as KNN algorithm, the Bayes, neural network classification algorithm and decision tree classification algorithm[1].

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.19

The Bayes[1]is suitable for dealing with incomplete data sets, and it can find the causal relationship among the data. However, it is large amount of calculation and efficiency is very low, and to some specific problem it is a more difficult problem to rationally confirm the priori probability of lots of variables. The Bayes supposes that characteristics to be classified are an orthogonal relation, and the polynomials are independent and identically distributed. Because there are no approaches to make sure some practical problems available to the rules, the application of it is very difficult.

The artificial neural network method [3] is more valid in the question of discerning on a small scale, but there is not a reliable t theory yet about fixing the topological structure of the network. It is easy to fall in local optima, and generalization ability to small sample's large-scale pattern recognition is bad.

Decision tree method is simple and clear, and the rule extraction ability is strong. It is suitable for dealing with problems associated with lots of examples. This method has some disadvantages, such as it depends too much on distribution and numbers of samples, too sensitive and ineffectually to process continuous attributes[4].

K nearest neighbor method (KNN)[4] based on the classification of instances is widely used in text classification . KNN method is simple, and does not require the training process. It can obtain higher classification accuracy for non-normal distribution and the unknown. It has the advantageous ability of robustness and conceptual clarity and it is also very easy to add new categories, it is also very easy to add new categories, and has a very small influence for the field. But in the text classification, KNN method has insufficient. KNN algorithm is a kind of lazy-learning algorithm, it need to store all samples in the computer and every decision requires to calculate the distance between testing samples and all training samples and compares. Therefore, storage and computing cost are large. Though the result of classification is better, its classification time is non-linear, when training texts or characteristics increase, classification time increases sharply.

In this paper we improve the KNN algorithm to increase the speed and accuracy of categorization.

III. A WEB PAGE CLASSIFICATION ALGORITHM BASED ON LINK INFORMATION

First, we discuss the organization characteristics of web sites and distinction methods of web information, and then we give the method about the settlement of classification. Finally, the web page classification algorithm baesd on link information is discussed.

A. The properties of web pages

The web pages on the network mainly have four ways to organize [5]: comprehensive web sites, professional web sites, the blogs and forums. It can further be divided into two major kinds according to the provider's characteristic: regular web sites and internet users' web sites.

Regular websites usually show us by means of comprehensive websites or professional websites. Its characteristic is that information is reliable and classification is very clear. The owners of the websites are mostly the government, relevant organizations or strong enterprises. Its structure is as shown in Figure 1.

The second kind is the Internet users' websites. The Internet users' websites talking about here does not mean the Internet users' private information, and refer to the websites which are published by the network users. They include the blogs, the forums and chatrooms. The form is varied. These websites have all kinds of information, the structures are almost different and. The quality of information varies widely and has a lot of errors and copies. The link information is complicated for these websites. The websites are influenced by the provider's interesting, interests and working, even the mood of the provider will influence the content and quality of websites. These characteristics have strengthened the difficulty of this kind of information classification. But because of respondent new knowledge and new information fast, it is easy to be accepted by users. By contrast, because this kind of websites are less likely to involve the commercial interests, it is comparatively just that the link information in the content.

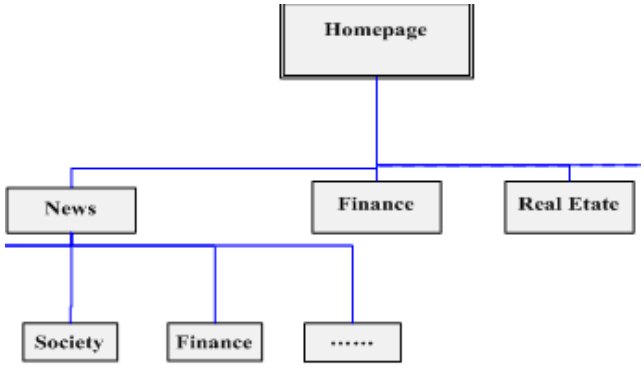


Figure 1. A comprehensive website

Network information was interwoven by these two kinds of information. The Regular websites themselves include the clear classification relation. The users' websites are comparatively fuzzy, but they can be classified by the link information. The structure of web pages is shown by Figure 2.

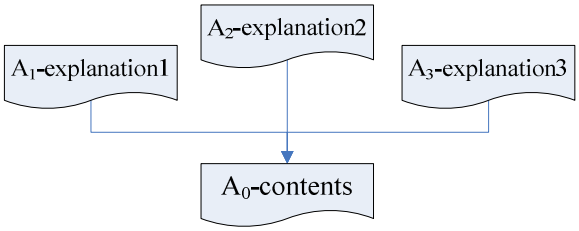


Figure 2. Logical relationships between Web pages

From Figure 2, we can conclude that website A1, A2 and A3 include links pointing to the website A0. These links can explain the website A0. In other words, the classification attribute of A0 is influenced by A1, A2 and A3. In other words, the classification attribute of A0 is influenced by A1, A2, and A3. So these links can help web pages classification and there are advantages of fewer features and high distinguishability.

B. Distinct sources of the web pages information

The websites of two kinds describe above interweave together, and quote, explain each other, form the complicated quotation relation. By distinguishing it, we can classify the web pages with them to improve the accuracy of classification.

The main methods of distinguishing the types of pages are as follows:

(1) We can judge the types of web pages through the web pages quantity. For example, private sites have fewer amounts of information and websites.

(2) We can also judge the types of web pages through the quantity of the servers and access speed.For example, the sites of the governments or the large enterprises have many servers and the access is fast. However, the small businesses and the private sites often have no economic strength to erect more servers and the access if slow.

(3) We can judge the types through URL. Because of being deficient of IP address and other reasons, there are no one's own servers and private websites of the Internet users and they publish the information through some other platforms. Some large internet firms, such as Google, Baidu, Sohu, Yeah, launched the blog and forum service. The URL usually includes domain name and some key words, such as blog, bbs, etc.

C. A. Setting-up and unity of the classification

The settlement of the classification is extremely complicated, and it directly influences exactness and rationality. It needs experts of different fields to participate in. Though large-scale websites have all classified the entries, but the criteria for classification is not unified.

The foreign representative criteria for classification have Dewey Decimal System, daily classification of USA's scientific research system, The International Standard Classification of Education (ISCED), etc. Because of the difference in respects, such as the culture background, habit of thought, these standards can't totally meet the classification [5] of the Chinese literature.

The representative classifications of domestic are Chinese Library Classification, subject classification and code. The former is mainly for the books category, and the latter for subject classification. Both of them do not apply to web page classification.

In this paper, we use the Open Directory Project (ODP)[7] as the forms of organization. It is the mapping target as the LIC algorithm. That is categories.

ODP is the biggest and most extensive human directory on Internet. It is created and maintained by the volunteers from around the world, and it is the world's largest global directory community. Currently it has 59,000 categories and contains more than 4,500,000 URLs. The directory structure is shown in Figure 3.

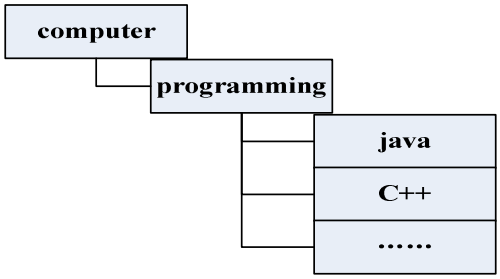


Figure 3. Classification structure chart of ODP

It can be seen from Figure 3 that ODP classification system has a clear hierarchy, such as Java is a sub-category of programming, programming is a sub-category of computer and so on. The traditional text classification systems are one or two structures. Comparing with traditional text classification systems, the ODP is more specific and it can quickly locate information. For example, when users need to find the "string interception methods "and other related issues, but java, C++, C and other fields contain the information in reality, and because of the traditional directory category too vague, the users have to find by category in "computer" or "programming language". This reduces the query efficiency. But fine-grained category increased the difficulty of the key technologies as the text feature selection and classifier design. Considering the query speed and accuracy, as a compromise, In this paper we selected the three-category as text classification and mapping of users' interesting.

D. Design of LIC algorithm

First, find the collection of pages in the training page with links to information vector to be classified in the K most similar articles page articles based on this website belongs to class K to determine the category a new page.

The Web Page Classification Based on Link Information is the improvement of the KNN algorithm [8]. First, we find K articles which are the most similar to the web pages to be classified at the training collection. Then we can determine the new pages' categories according to the K articles.

The main difference of LIC and KNN algorithm is that: the latter classifies relying on web contents; the former is dependent on its parent page reference information. The LIC algorithm determines the category attribute of the current page through the links which other web pages point to the current page. The algorithm calculates the importance of web pages belonging to a category, and endowed with different weights to each link (inbound).

The algorithm is list in detail as follow.

1)Set the limits of being classified documents

Not all pages on the network are to be classified. More pages are aggregated page and it does not provide

substantive information, mainly existing to link to other pages. The classification of this website is not necessary. Boundaries of the documents which will be classified is determined by Formula (1).

$$\varphi = \frac{L(d_i)}{O(d_i)} \tag{1}$$

Where $L(d_i)$ is the length of the document d_i， $O(d_i)$is the out link degree of the document. The ratio φ is to be used to indicate the possibility of classification of the document collection D. When the value φ is greater than a threshold α, we can determine that the d_i is to be classified document and it belongs to the document collection D to be classified.

2)Link Notes preprocessing

We analysis them and set the initial category. Get the pages' all direct parent link information which will to be classified and are expressed as the vector space $\{(t_1, w_1), (t_2, w_2), ..., (t_n, w_n)\}$, Among them, t show that chains information, w expresses the weight, which is calculated by Formula (2).

$$w(t,d) = \frac{\sum_{i-1}^{i=n} q_i}{N(t)} \tag{2}$$

Where $N(t)$ shows the number of links that contains the keyword t, n is all the link numbers of web pages to be classified. q_i is the weight of direct father's web pages the web page to be classified, q_d、q_u、q_r respectively. If the domain name of the website to be classified is the same with domain names of its parent page, then we set its weight q_d. Otherwise, if it's father's webpage is the first category, the weight is q_u; If it's father's web pages are second category then the weight is qr, obviously $q_d > q_u > q_r$.

3)The text classification based on link information

We select K links information vector in the training set d, which are the most similar to the web page that is expressed by x. It is calculated by Formula (3).

$$sim(x,d) = \frac{x \cdot d}{|x||d|} \tag{3}$$

Where, $sim(x, d)$ is similar degree of cosine between the vectors of link information.

4)Among the k neighbours of web pages to be classified, sequentially calculate the weight of each category.

Formula (4) shows the calculation.

$$p(x \mid c) = \sum_{i=1}^{i=k} sim(x, d_i) y(d_i, c) \tag{4}$$

Among them, $y(d_i,c)$ is function of the classification attribute, that is, if d_i belongs to category c, the function is 1, otherwise 0.

5)Compare the weight of each category and the web pages are mapped to the greatest weight category.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experiment environment

To verify the effectiveness of the algorithm, we use java language, in a pc with Intel Pentium (R) D CPU 2.80GHz and 1G RAM, in Java SE, Windows XP programming environment, to realize the KNN algorithm, support vector machines, and the LIC algorithm.

B. Experimental data

We get experimental data from integrated portal Sohu, Netease, Yahoo and professional website javaeye, csdn. We totally collect 3430 pages, which contains 9 ordinary categories and 9 professional categories.

We use the document frequency of feature extraction methods to preprocess the data and every select 200 features, a total of 1800-dimensional feature dimension to the traditional classification algorithms. Performance evaluation of web page classification we use three indicators, precision, recall, and the assess value of F1. To the overall performance evaluation of the classification system we calculate the average macro [8].

C. Experimental result

We collect twenty percent of the web pages from the preprocessed web pages as a test set, the remaining as a training set .Three experiments, LIC, KNN and SVM, were conducted separately.

The experimental result of comprehensive website is shown in Table 1.

TABLE I. EXPERIMENTAL DATA FOR COMPREHENSIVE WEBSITE

Category	Rate of accuracy			Recall rate			F1 testing value		
	LIC	***KNN***	***SVM***	***LIC***	***KNN***	***SVM***	***LIC***	***KNN***	***SVM***
Sports	0.995	0.999	0.999	0.769	0.156	0.786	0.867	0.271	0.880
Health	0.999	0.982	0.999	0.685	0.683	0.364	0.813	0.806	0.533
Amusement	0.997	0.995	0.999	0.676	0.522	0.395	0.806	0.685	0.566
House property	0.999	0.998	0.999	0.742	0.305	0.653	0.852	0.467	0.789
Education	0.999	0.996	0.999	0.727	0.667	0.531	0.842	0.799	0.693
Car	0.999	0.985	0.999	0.738	0.465	0.558	0.849	0.633	0.716
Computer	1	0.995	0.222	0.689	0.819	0.949	0.816	0.898	0.360
Science and technology	0.999	0.997	0.449	0.757	0.358	0.563	0.862	0.527	0.500
Finance and economics	0.977	0.988	0.999	0.626	0.428	0.227	0.763	0.598	0.369
	Average of accuracy			***Average of recall rate***			***average of F1 value***		
	0.996	0.993	0.852	0.713	0.489	0.558	0.830	0.632	0.601

The results show that the LIC has a higher and recall than other when classifying the web pages of the portal sites. The value of F1 is also greater, but the difference is not prominent. When we classify the professional sites, the precision, recall and F1 of LIC is significantly higher than other algorithms. The reason is that the traditional classification algorithms classify the sites depending on the content, and each language, especially SQL statements, is extremely similar. Therefore, it leads to the low precision and recall.

The experimental result for professional website is shown in Table 2.

TABLE II. THE EXPERIMENTAL RESULT OF PROFESSIONAL WEBSITE

Category	Rate of accuracy			Recall rate			F1 testing value		
	LIC	*KNN*	*SVM*	*LIC*	*KNN*	*SVM*	*LIC*	*KNN*	*SVM*
java	0.978	0.799	0.798	0.564	0.482	0.415	0.715	0.601	0.547
C/C++	0.989	0.879	0.786	0.753	0.356	0.598	0.855	0.507	0.679
C#	0.976	0.775	0.686	0.695	0.327	0.374	0.812	0.460	0.484
python	0.987	0.863	0.879	0.792	0.420	0.453	0.879	0.565	0.598
Linux	0.976	0.759	0.875	0.574	0.568	0.431	0.723	0.650	0.577
Ajax	0.987	0.873	0.768	0.958	0.428	0.563	0.972	0.575	0.650
Oracle	0.956	0.659	0.677	0.626	0.468	0.498	0.757	0.547	0.574
MySQL	0.963	0.684	0.527	0.633	0.319	0.549	0.764	0.435	0.538
SQL Server	0.959	0.654	0.644	0.678	0.357	0.458	0.794	0.462	0.535
	Average of accuracy			**Average of recall rate**			**average of F1 value**		
	0.975	0.772	0.738	0.697	0.414	0.482	0.8084	0.5341	0.576

In summary, LIC is more suitable for Web information classification, especially for professional web site, it has better classification results than other traditional classification algorithms, such as KNN, SVM.

V. CONCLUSION

This paper presents a web page classification algorithm, Link Information Categorization algorithm (LIC). LIC is based on KNN method, make full use of the characteristics of Web information. Compared with traditional classification algorithm, LIC algorithm has the following advantages: (1)Can effectively prevent the page has been artificially processed deception; (2) the classification algorithm is mainly based on link information, make a little improvement, it can achieve the classification of non-Web documents, such as video, music and other non-document information. (3) slightly modified the algorithm can also achieve the classification of information for different languages.

Merge LIC with personalized search engine technology, and improving the efficiency of LIC is our future work.

REFERENCES

[1] Q. Xiaoguang, B. D. Davison. Web page classification: Features and algorithms. ACM Computing Surveys, 2009,41(2)

[2] W. Zhong-Ying, W. Ming-Wen. Automated text classification model based on projection pursuit regression. Journal of Chinese Information Processing, 2005, 29 (4)

[3] H. Kim, P. Howland, H. Park. Dimension reduction in text classification with support vector machines. Journal of Machine Learning Research, 2005,6(1): 37 −53

[4] S. JinShu, Z. BoFeng, X. Xin. Advances in Machine Learning Based Text Categorization. journal of software, 2006,17(9), pp.1848 −1859

[5] Z. Miao. On the basis of supporting the Chinese text sort research of the vector quantity machine[D]. Henan University of Technology, 2008

[6] Web Surveys.http://www.iresearch.com/.2009-10

[7] Open Directory Project. http://www.dmoz.org/about.html.2008-07

[8] Z. Haifeng. The studies on personal search engine [D]. Henan University of Technology, 2009.

Calculation Of Quantum Entanglement

Jingshui Yu
School of IoT Engieering
Jiangnan University
Wuxi Jiangsu 214122, China
e-mail: shinging0@126.com

Wenbo Xu
School of IoT Engieering
Jiangnan University
Wuxi Jiangsu 214122, China
e-mail: xwb@jiangnan.edu.cn

***Abstract*—Quantum entanglement which is used as the basis for such applications as quantum cryptography, quantum teleportation, and quantum computing, is a multifaceted property that has attracted much attention. The calculation of quantum entanglement therefore has become important. But the calculation of entanglement has become much more complex as systems that use entanglement have evolved. Based on the relative entropy of entanglement, a general method was developed for the calculation of entanglement for n-qubit or n-qudit system, which using an optimization algorithm technique such as genetic algorithm or Quantum-behaved Particle Swarm Optimization. The method using different optimization algorithm was tested on a two qubit system, for which there are some known points, and compared to exact calculations using the entanglement of formation and with each other. Advantages and disadvantages of the method are discussed and directions of future research are exposed.**

Keywords-Quantum Entanglement; Genetic Algorithm ; Quantum-behaved Particle Swarm Optimization(QPSO) Algorithm; Calculation; Measure; The Relative Entropy Of Entanglement

I. INTRODUCTION

Quantum entanglement has been known since 1935 when Einstein, Podolsky and Rosen, (EPR) [1], and Schrödinger [2] investigated the counterintuitive properties of quantum systems. Through the research of the question of expected locality of the entangled quantum systems raised by EPR, Bell developed his famous inequalities which serving as a test of strange properties of the simplest entangled wave function represented by a singlet state. Any discuss of entanglement measures must begin with a question of what entanglement is, and how we actually use it. Entanglement can loosely be described as the quantum correlations that can occur in many party quantum states. Then, how do we define quantum correlations, and what differentiates them from classical correlations? Classical correlations can be defined as those that can be generated by LOCC (Local Operations and Classical Communication) operations. If we observe a quantum system and find correlations that cannot be simulated classically, then we attribute them to quantum effects, and label them quantum correlations. Suppose that we have a noisy quantum state, and we process it using LOCC operations. If in this process we obtain a state that can be used for some task that cannot be simulated by classical correlations, such as violating a Bell inequality [3], then we must not attribute these effects to the LOCC processing that we have performed, but to quantum correlations that were already present in the initial state, even if the initial state was quite noisy. This is an extremely important point that is at the heart of the study of entanglement. Unfortunately, despite a great deal of work by a large number of well-funded and very bright people, entanglement and the calculation of entanglement are not yet fully understood. So, we need a convenient and accurate method for general entanglement calculation. In this thesis, applied to the relative entropy of entanglement, we develop a general method for calculating the entanglement, based on a genetic algorithm (or a quantum-behaved particle swam optimization algorithm) technique. We test and apply it to the relatively simple two-qubit system, in order that we will be able to determine whether it is working; however, its advantage is that it is easily extendible to multiple qubit systems and to qudit systems.

II. THE RELATIVE ENTROPY OF ENTANGLEMENT

The regularized version of the relative entropy of entanglement is a measure that lies between entanglement cost and entanglement of distillation [4,5,6]. The quantum mutual information is

$$I(\rho_{AB}) = S(\rho_A) + S(\rho_B) - S(\rho_{AB}) \quad (1)$$

Employing the quantum relative entropy

$$S(\rho \| \sigma) := tr\{\rho \log \rho - \rho \log \sigma\} \quad (2)$$

, which is a measure of distinguishability between quantum states, one may then write the quantum mutual information as

$$I(\rho_{AB}) = S(\rho_{AB} \| \rho_A \otimes \rho_B) \quad (3)$$

The total correlations are quantified by a comparison of the state ρ_{AB} with the closest separable state, a classically correlated state devoid of quantum correlations. The general definition of the relative entropy of entanglement [4, 5, 6, 7] with respect to a set X as

$$E_R^X(\rho) := \inf_{\sigma \in x} S(\rho \| \sigma) \quad (4)$$

This definition leads to a class of entanglement measures known as the relative entropies of entanglement. In the bipartite setting the set X can be taken as the set of separable states, states with positive partial transpose, or non-distillable states. In the multiparty setting there are even more

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.27

possibilities but for each such choice a valid entanglement measure is obtained as long as the set X is mapped onto itself under LOCC.

III. ALGORITHM

A. Genetic Algorithm

Genetic algorithm [8] is a method for solving both constrained and unconstrained optimization problems that is based on natural selection, the process that drives biological evolution. A genetic algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm selects individuals at random from the current population to be parents and uses them to produce the children for the next generation. Over successive generations, the population evolves toward an optimal solution. The genetic algorithm uses three main types of rules at each step to create the next generation from the current population:

- Selection rules select the individuals, called parents, which contribute to the population at the next generation.
- Crossover rules combine two parents to form children for the next generation.
- Mutation rules apply random changes to individual parents to form children.

A genetic algorithm is a computational technique based on the evolution of the species. A possible solution to the problem is coded in a binary string, called a chromosome. An initial population of chromosomes is created at random and processed by natural selection. During reproduction: crossover, the exchange of parts of the binary string between chromosomes; mutation, the inversion of the bits of the binary string; and inversion, all take place at random position in the binary string. An evaluation of the fitness, how good the solution is, takes place for all individuals in the population. With these rules, the features of one solution can be transmitted to next generation of chromosomes and better solutions can be found. Natural selection and reproduction are probabilistic stages and, hence, a genetic algorithm is a random process. The best solution cannot be predicted, but a number of generations must run to find the best solution.

B. Quantum-behaved Particle Swam Optimization (QPSO) Algorithm

Quantum-behaved Particle Swam Optimization (QPSO) [9,10,11] was proposed keeping to the philosophy of PSO[12], which is a Delta Potential well model of PSO in quantum world. A particle is composed of three vectors:

- The x-vector records the current position (location) of the particle in the search space.
- The p-vector records the location of the best solution found so far by the particle.
- The v-vector contains a gradient (direction) for which particle will travel in if undisturbed.

, and two fitness values:

- The x-fitness records the fitness of the x-vector.
- The p-fitness records the fitness of the p-vector.

In QPSO the particles move according to the following equation:

$$mbest=\frac{1}{M}\sum_{i=1}^{M}p_i\left(\frac{1}{M}\sum_{i=1}^{M}p_{i1},\frac{1}{M}\sum_{i=1}^{M}p_{i2},\ldots,\frac{1}{M}\sum_{i=1}^{M}p_{id}\right) \quad (5)$$

$$p_{id}=\varphi * p_{id}+(1-\varphi)*p_{gd},\varphi=rand() \quad (6)$$

$$x_{id}=p_{id}\pm\beta*\left|mbest_d-x_d\right|*\ln\left(\frac{1}{u}\right),u=rand() \quad (7)$$

, where $mbest$ is the mean best position among the particles. φ and u are a random number distributed uniformly on [0,1] respectively and β is the only parameter in QPSO algorithm.

IV. CALCULATIONS AND RESULTS

We'll use the genetic algorithm and Quantum-behaved Particle Swam Optimization algorithm to calculate the entropy of entanglement for large classes of states, using three different distance measures: the Bures distance, the Hilbert –Schmidt distance, and the von Neumann distance. These results are compared to the entanglement of formation, which is calculated according to the exact formula for two-qubit systems.

In graphed results, there are seven different entanglement calculations shown:

- Entanglement of formation as calculated according to the exact formula, the equation is defined as

$$E_F(\rho)=s\left(\frac{1+\sqrt{1-C^2(\rho)}}{2}\right) \quad (8)$$

$$s(x)=-x\log_2 x-(1-x)\log_2(1-x)$$

- The entropy of entanglement, using the Bures distance to the set of separable states, calculated using the genetic algorithm or QPSO. The Bures distance is defined as

$$D_B(\rho_1,\rho_2)^2=2-2Tr\sqrt{\sqrt{\rho_1}\rho_2\sqrt{\rho_1}}=2-2\sqrt{F(\rho_1,\rho_2)} \quad (9)$$

, where ρ_1 and ρ_2 are density matrices.

- The entropy of entanglement using the Hilbert – Schmidt distance to the set of separable states is calculated using the genetic algorithm or QPSO. The Hilbert – Schmidt distance is defined as

$$H-S(\rho_1,\rho_2)=\sqrt{Tr\left[(\rho_1-\rho_2)^2\right]} \quad (10)$$

, where ρ_1 and ρ_2 are density matrices.

- The entropy of entanglement using the Von Neumann distance to the set of separable states is calculated using the genetic algorithm or QPSO. The von Neumann distance is defined as

$$VonNeumann(\rho,\sigma)=\rho\log\frac{\rho}{\sigma} \quad (11)$$

, where ρ and σ are density matrices.

The results are as follows:

Fig. 1 shows results for of |00> + |11> + γ|01>. The normalized density matrix is

$$\left(\frac{1}{2+\gamma^2}\right)\begin{pmatrix}1 & 0 & \gamma & 1\\ 0 & 0 & 0 & 0\\ \gamma & 0 & \gamma^2 & \gamma\\ 1 & 0 & 0 & 1\end{pmatrix} \tag{12}$$

Fig. 1 is missing the Von Neumann results because the density matrices turn out to be singular and logarithm calculation did not converge. The Hilbert – Schmidt distance was found to be one for all values of γ for this state and was therefore left off the graph.

Fig. 2 shows results for M'(γ), which is a fully entangled state with a varying amount of mixture contamination, that is, a density matrix given by

$$\left(\frac{1}{2+\gamma}\right)\begin{pmatrix}\gamma & 0 & 0 & 0\\ 0 & 1 & 1 & 0\\ 0 & 1 & 1 & 0\\ 0 & 0 & 0 & 0\end{pmatrix} \tag{13}$$

Fig. 2 is also missing the Von Neumann results because the density matrices turn out to be singular and logarithm calculation did not converge.

Fig. 3 is a calculation of entanglement for the Werner states, which are EPR triplet states with noise. The Werner states are defined as

$$Werner(F) = F\begin{pmatrix}0 & 0 & 0 & 0\\ 0 & 0.5 & -0.5 & 0\\ 0 & -0.5 & 0.5 & 0\\ 0 & 0 & 0 & 0\end{pmatrix} + \frac{1-F}{3}\begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 0.5 & 0.5 & 0\\ 0 & 0.5 & 0.5 & 0\\ 0 & 0 & 0 & 1\end{pmatrix} \tag{14}$$

Fig. 4 shows results for the Bell triplet plus white noise.

$$Bell\ (p) = \frac{p}{2}\begin{pmatrix}1 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 1\end{pmatrix} + \frac{1-p}{4}\begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix} \tag{15}$$

Fig. 5 shows results for the EPR triplet state with white noise.

$$EPR\ (p) = \frac{p}{2}\begin{pmatrix}0 & 0 & 0 & 0\\ 0 & 1 & 1 & 0\\ 0 & 1 & 1 & 0\\ 0 & 0 & 0 & 0\end{pmatrix} + \frac{1-p}{4}\begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{pmatrix} \tag{16}$$

Fig. 6 shows results for the Bell triplet plus near white noise.

$$Bell\ (\gamma) = \frac{1}{2+4\gamma}\begin{pmatrix}1+\gamma & 0 & 0 & 1\\ 0 & \gamma & 0 & 0\\ 0 & 0 & \gamma & 0\\ 1 & 0 & 0 & 1+\gamma\end{pmatrix} \tag{17}$$

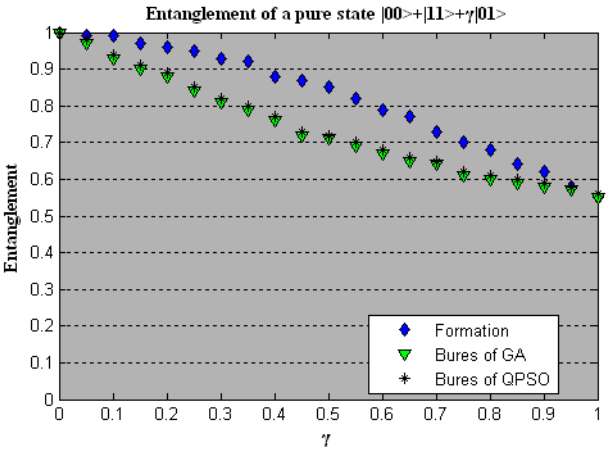


Figure 1. Entanglement of a pure state|00> + |11> + γ|01>

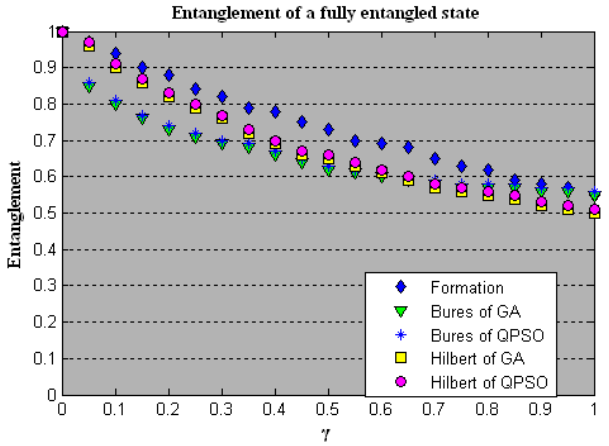


Figure 2. Entanglement of a fully entangled state

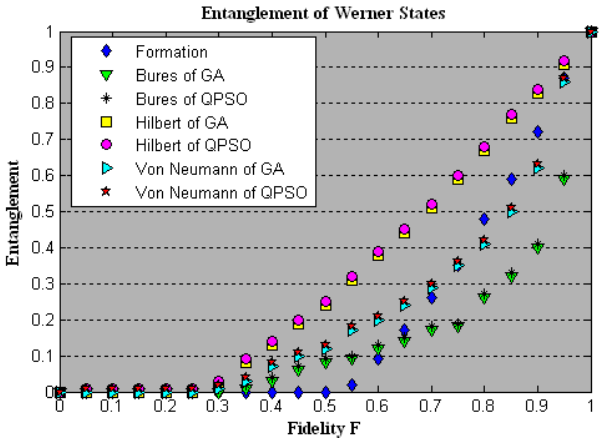


Figure 3. Entanglement of Werner States

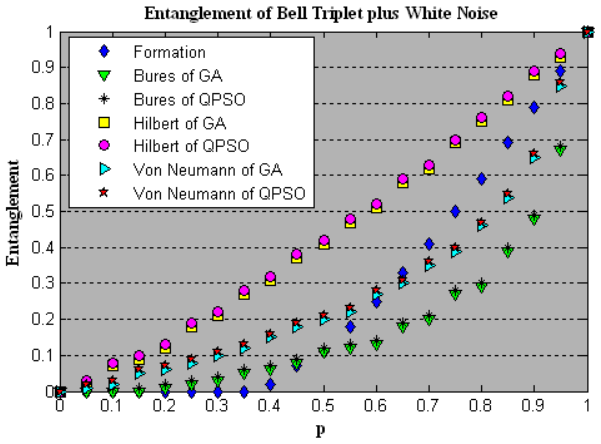


Figure 4. Entanglement of Bell Triplet plus White Noise

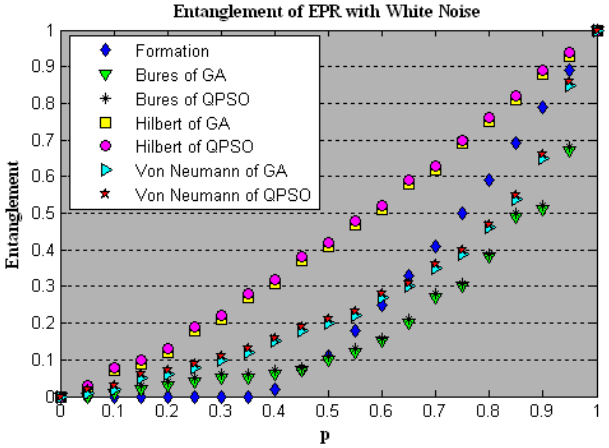


Figure 5. Entanglement of EPR with White Noise

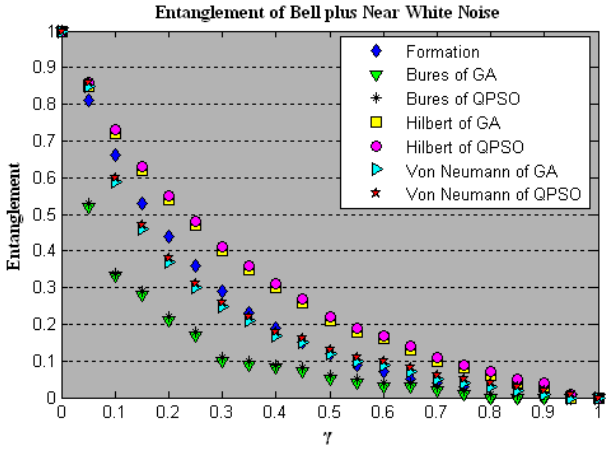


Figure 6. Entanglement of Bell plus Near White Noise

V. CONCLUSIONS AND DISCUSSION

It is always possible that there is another matrix in the set X which we did not find, which is closer still to the system's density matrix, and, thus, that the correct value for the entanglement is smaller. However, the fact of two algorithms that the Bures numbers are always the lowest, then the von Neumann, then the Hilbert – Schmidt, gives us some confidence that the method is getting good minimization, since if we were not, the rankings would be arbitrary and probably not consistent. As it is, the rankings can be ascribed to the differences among the distance measures. It is also good to see that for each of the Fig. 1 – Fig. 6, the curves generally have the same shape. The Bures distance has a bit more noise, but there multiple square roots of matrices involved, so one would expect more problems with these calculations.

The major advantage of using our method, a genetic algorithm or a QPSO algorithm to find the minimum entanglement, is that it is very easy to change what you are minimizing. Not only the genetic algorithm but also the QPSO algorithm do not care about the complexity of the system; if it can calculate the results it will find a minimum.

The major disadvantage of using a genetic algorithm or a QPSO algorithm method is that, since it is a random process, it does take a long time processing. Sometimes with a complicated equation, the genetic algorithm method can get stuck in a local minimum, instead of the global minimum desired. Multiple attempts may have to be tried until the global minimum is found (but the QPSO dose not).

There are a number of items that could be done next. So far, entropy of entanglement has been tried, but other entanglement calculation such as entanglement distillation could also be tested. Also, three or greater qubit systems could be tried. The same procedure can easily be used with minor modifications. And also we can extension it to qudit systems.

REFERENCES

[1] Einstein, A., Podolsky, B. and Rosen, N., "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?," Physical Review, Vol. 47, May 1935, pp 777-780.

[2] E. Schrodinger, E., "Discussion Of Probability Relations Between Separated Systems," Proceedings of the Cambridge Philosophical Society, Vol. 31, 1935, pp 555-563.

[3] Thaller, B., Advanced Visual Quantum Mechanics, Springer, New York, NY, 2005.

[4] Vedral, V, and Plenio, M.B., "Entanglement Measures And Purification Procedures," Physical Review A, Vol 57, No. 3, March 1998, pp. 1619 – 1633.

[5] Vedral, V., Plenio, M. B., Rippin, M. A., and Knight. P. L., "Quantifying Entanglement," Physical Review Letters, Vol. 78, No. 12, March 1997, pp 2275 - 2279.

[6] Vedral, V., Plenio, M. B., Jacobs, K., and Knight, P. L., "Statistical Inference, Distinguishability of Quantum States, and Quantum Entanglement," Physical Review A, Vol. 53 No. 6, December 1997, pp 4452 – 4455.

[7] Plenio, M. B., Virmani, S. and Papadopoulos, P., "Operator Monotones, The Reduction Criterion And The Relative Entropy," Journal of Physics A: Mathematical and General, Vol. 33, No. 22, June 9, 2000, pp L193 – L197.

[8] MATLAB Genetic Algorithm and Direct Search Toolbox 2 User's Guide, The MathWorks, Inc., Natick, MA, 2007.

[9] SUN J, XU WB, "A Global Search Strategy of Quantum-behaved Particle Swarm Optimization[A]," in Proceedings of IEEE Conference on Cybemetics and intelligent Systems[C], 2004.111~116.

[10] SUN J, XU WB, “Particle Swarm Optimization with Particles Having Quantum Behavior[A],” in Proceedings of 2004 Congress on Evolutionary Computation[C], 2004.325~331.

[11] Jing Liu, Jun Sun, and W.B. Xu, “Solving Constrained Optimization Problems With Quantum Particle Swarm Optimization[A],” DCABES and ICPACE Joint Conference on Distributed Algorithms for Science and Engineering[C],2005: 99~103.

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

Leveraging Internet Marketing to Extend Business Reach: An Empirical Analysis of One Language Training Centre

Ai Xu
International Business Faculty
Beijing Normal University Zhuhai
Zhuhai, China
e-mail: gdxuai@163.com

Shufeng Gao
College of Computer Science & Technology
Beijing Institute of Technology Zhuhai
Zhuhai, China
e-mail: gsfdl@163.com

Xiaohong Li
School of Management
Dalian Polytechnic University
Dalian, China
e-mail: lxh0717@yahoo.com.cn

***Abstract*—Internet Marketing can help companies get more traffic to their website and extend their business reach. Taking one language training centre named as Zhuhai New-Concept Training Centre (ZHNCTC) as an example, this paper details some Internet marketing tactics, including approaches to increase website traffic, methods to improve the "time on site" and "pages viewed", proposed activities to drive more sales from existing customers and some relevant evaluation methods. The paper also puts forward some recommendations about successful implementation of these Internet marketing tactics.**

Keywords-Internet Marketing; business reach; tactics; empirical analysis

I. INTRODUCTION

Zhuhai New-Concept Training Centre (ZHNCTC), founded in 2000, specializes in English language education and training for adults and students and has certain popularity in Zhuhai area. But in order to extend the business reach and promote the development of ZHNCTC in the fierce competition, it is necessary to leverage Internet marketing tactics and take a series of actions to achieve three objectives: (1) to increase website visits, (2) to improve the "time on site" and "pages viewed", and (3) to drive more sales from existing customers.

This paper begins with the analysis and evaluation of the current website traffic of ZHNCTC, and then details the approaches to increase website traffic, including search engine promotion, social media promotion and mail list promotion. The specific practices for search engine promotion are also presented, i.e. search engine optimization (SEO), adding URL to search engine and directories, making keywords advertisement, setting up backlinks and exchange links. Then how to improve the "time on site" and "pages viewed" is studied. Improving the website design and contents are considered to be very important. The detailed methods are proposed respectively. In order to drive more sales from existing customers, the report designs some activities, such as "5-day Free Trial for On-line English Learning", "Parents, Children— Learn English Together" and "10% Off while Learning together with Your Classmates". Some evaluation methods are discussed too, i.e. undertake website evaluation by using some analysis tools, obtain feedback through customer surveys, and calculate and measure ROI for Website Promotion. At last, the report also puts forward some recommendations about successful implementation of the digital marketing tactics.

II. ANALYSIS AND EVALUATION OF THE CURRENT WEBSITE TRAFFIC OF ZHNCTC

Firstly, it is wise to review and evaluate the current website traffic of ZHNCTC. But while logging on Alexa website to see the traffic rank of ZHNCTC in China, there is no data available for it. While on the website of China Website Ranking, it can be found that the rank is 1289853, with only 1 reach per million users and 1 page views per user. Although the traffic rank doesn't reveal the actual traffic of one website absolutely, but it can really show that the current traffic of ZHNCTC is very low.

After a preliminary usability analysis, it can be found that there are some usability issues on the ZHNCTC website.

A. No Site Search

There is often a search engine within the website to index the information on the web site or intranet, making it easier for visitors to find what they are looking for. But no search engine tool exists on the webpage of ZHNCTC, which would result that visitors are hard to find information they need.

B. Legibility Problems

Some text is illegible on the landing page. The color depth varies on the same area, and the color of the date jointed with the headline is too light to identify.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.55

C. Not Changing the Color of Visited Links

Color doesn't change for the visited links, which will make visitor not know where they have already visited and where to go next.

In general, the website of ZHNCTC is not easy to navigate and clearly organized so that visitors can't easy to find the information they're looking for.

III. Approaches to Increase Website Traffic

In order to attract more visitors' attention and extend business reach, it is essential for ZHNCTC to improve the Internet presentation. There will be many creative ways, some of them will cost money, and some won't. The following three channels are strongly suggested here.

A. Search Engine Promotion

1) Search Engine Optimization: ZHNCTC should improve its search engine ranking by focusing content on keywords, which will help people find out the website. The keywords should highlight English education and training, especially New-Concept English training, and also should emphasize the area of Zhuhai, because the training will be mainly delivered to students and adults in Zhuhai. The keywords should not only go into headings and page names but also into "meta tags". It would be advantageous by using Free META Tag Generator (http://www.submitexpress.com/free-submission.html) for best placement. Keywords prominence, proximity and density should be taken into account, because they are the three basic rules of keyword deployment. Claiming the website listing on Google Maps & Places would be helpful for local search.

2) Add URL to Search Engine and Directories: It is a good choice for ZHNCTC to add website URL to search engines, such as Google, Baidu, Yahoo, etc. Now there are automatic search engine registration and URL submission services available on the Internet, such as Submit Express (http://www.submitexpress.com/free-submission.html), which will be helpful. Meanwhile, it is necessary to submit the website to some important directories, such as Open Directory Project, GoGuides.Org, JoeAnt.com, etc. All of these will help the website to be searched by search engines.

3) Buy Baidu Keywords Advertisement: Some advertisements on the web are worth considering, and Baidu Phoenix Nest (Baidu PPC) is preferred. The reasons of choosing Baidu.com are very simple. The overwhelming majority of ZHNCTC customers are Chinese people and Baidu.com has already become the largest search engine in China. At the same time, it is helpful to control the budget by using Pay Per Click advertisement. In view of fund limitation, the budget should be set no more than 50 RMB everyday and for one month. If the online courses are not available soon, the advertising area will be only restricted at the city of Zhuhai.Webmaster and managers of ZHNCTC should take the time to review the reports provided by the PPC services. These reports will provide information in regard to how well each of the keywords are performing and enable them to adjust their keywords or change the budget accordingly.

4) Set up Backlinks and Exchange Links: Building high quality backlinks will boost the Pagerank and Search Engine Position. Therefore it is necessary to build more effective internal backlinks. Relevant links should be made to other websites with the topic closely related to English education & training, such as the local education institution, primary & secondary school, etc. Other ways are recommended, e.g. motivating staffs to post answers to other people's questions or ask questions on some forum such as Baidu Zhidao, or to submit articles, make comment, initiate discussion on BBS, Spaces and Blogs, all of which can create useful backlinks and increase traffic.

B. Social Media Promotion

Promotion through social media is one of the most effective and relatively new. There are dozens of social media sites active across the Internet today, within which QQ space, Renren, Kaixin, Baidu Zhidao are very popular in China. One effective way to promote on these social media site is to add "JisThis Share" button to ZHNCTC website, which is a social share tools provided by JisThis. This button can make users easily share the content of ZHNCTC website to some very popular social media websites.

C. Mail List Promotion

It is necessary to collect contact email address of the current students and their parents, even the potential customers visiting the reception, and then set up a mailing list in Outlook, so that managers of ZHNCTC can send emails to a group of people easily. ZHNCTC should develop some kind of E-zines relating to methods about how to learn New-Concept English, Oral English Expressions, English stories, etc., which will be welcomed by students and increase satisfaction and loyalty, thus resulting many return customers. New initiatives would also easily reach customers through email.

In general, the website visits can be greatly increased through the above-mentioned tactics with low expenditure.

IV. Methods to Improve the "Time on Site" and "Pages Viewed"

It is widely recognized that rich content is the best guarantee to attract people to visit the website. Therefore more efforts should be spent in improving the design, structure and especially the contents of the website of ZHNCTC so as to make it more attractive to visitors, improve their "time on site" and "pages viewed", and encourage them to return.

A. *Improve Website Design*

The existing usability issues should be solved as soon as possible. Meanwhile some other actions helpful for increasing website usability should be taken.

- Add a Search Engine on the website to provide easy navigation functionality for visitors and encourage them to stay longer. This can be achieved by download the free search code of some search engine providers and add the code into the webpage of ZHNCTC.
- Ensure the pages easy to read by choosing suitable font size, breaking up blocks of text and creating short paragraphs.
- Make the color changed for those visited links and keep the color consistent for the same items, e.g. the headings.
- Place important information on the top left of the page so as to catch more attention.
- Ensure the pages load as quickly as possible.
- Clean the website up by removing all the unnecessary visual elements to allow important items to stand out.
- Highlight some important headlines by using different color or signals to attract attention.
- Add clear click paths to indicate the current position and provide link to the homepage.

B. *Improve Website Contents*

The key to retain customers is to continue to add high quality content to the website and keep it up to date and relevant.

- Try to offer high-quality content on the website. The more contents ZHNCTC have, such as methods & tips of English learning, the more often and time visitors will be on the website of ZHNCTC. Those selective contents only for current students should with great value and the contents opened to the public should also be attractive. Furthermore, for repeat visits, it is crucial to provide regular updates to the website contents, especially in frequently viewed zones.
- Open a chat room or BBS board on ZHNCTC website. This will be helpful not only for creating an interactive communication platform, but also providing a place to deliver on-line courses, which will be with a great market prospects.
- Add video to landing pages. Good video can improve conversions and page ranking more than most any other item added to the pages.
- Set up some English contest. People who enter will continually revisit the website to get the results.
- Offer free e-books related English learning or relative software. While people download the e-books, they may browse other contents on the website.

V. INITIATIVES TO DRIVE MORE SALES FROM EXISTING CUSTOMERS

The above-mentioned approaches are not only helpful to attractive more visitors and increase website traffic, but also conductive to benefit existing customers. In addition to this, it is vital to initiate some activities to drive them register more courses and other payable services. The following initiatives should be good choices, which can be best deployed through emails.

A. *5-day Free Trial for On-line English Learning*

Nowadays, online English learning become more and more popular. So it is necessary for ZHNCTC to set up this kind of services. This free trail activity will attract more attention from people, especially the existing customers and encourage them to apply for free trial and then register payable on-line courses.

B. *Parents, Children—Learn English Together*

Parents should understand and believe in the importance of English language skills, which would affect the attitudes of children. Offering a discount for parents will motivate them to learn and practice English together with their children. This would encourage children to work harder.

C. *10% Off While Learning together with Your Classmates*

This activity will encourage the current students to continue their study at ZHNCTC and not switch to other, at the same time, bring more new customers from their own school & class to save their own money.

VI. MEASUREMENT OF THE EFFECTIVENESS

A. *Undertake Website Evaluation by Using Some Analysis Tools*

There are many tools for website traffic and usability & conversion analysis, some of them are even free. Therefore ZHNCTC can choose some to measure the effectiveness of the deployment. Google Analytic is undoubtedly one of the most useful tools to undertake website analysis and evaluation. Considering Google's departure from the Chinese market, Baidu Tongji, a free analytics tool from Baidu, will be the best way for ZHNCTC. Through Baidu Tongji, ZHNCTC can track user's actions and get to know the how, where and when of traffic flow as well as time spent on each page, and then make decisions about how to further improve the website and increase the ROI. It's also worthy to use other useful analytical tool, such as 4Q Online Survey, Attention Wizard, ClickDensity and Website Optimiser. It should be noticed that, Baidu.com launched its official web page heat map service on 10 January 2011, claiming it is the first free heat map analysis service in the world. Therefore ZHNCTC can make use of this to see which part of a link is being clicked more, reveal poorly performing areas of the website or look for "high click traffic" spots on the page, and consider further to put an advertisement there for maximum results.

B. Obtain Feedback through Customer Surveys

Customer interviews and surveys will be useful ways for ZHNCTC to obtain feedback and evaluation directly from customers.

C. Calculate and Measure ROI for Website Promotion

The ROI of the keywords advertising should be considered and assessed, e.g., the ROI on Baidu PPC should be subdivided and further analyzed, because there are several keywords, each of them has their own contribution. Therefore web managers of ZHNCTC should identify ROI of the promotion plan, the keywords groups and each keyword. And then web managers can optimize keywords according to their ROI, e.g. increasing investment to the keyword with high ROI and improve the landing page to increase conversion rate.

VII. Recommendations about Implementation

To ensure the Internet marketing tactics to be successful, proper implementation is very important. The following issues should be taken into consideration.

A. Management Support and Commitment

Management backing will ensure sufficient funding and resources for the project, make it go well and to be successful.

B. Detailed Action Plans

It is necessary to clarify who, where, when, and how different individuals are going to implement this digital marketing plan. If the current staffs are insufficient, new recruitment and training are necessary.

C. Organizational Change

There may need some changes for ZHNCTC in structures, people and business process, e.g. establishing a new section in charge of those online learning courses, hiring more qualified teachers and clerks, improving the present operation process about booking, delivering and evaluating courses, etc. Therefore, change management should be taken into account.

D. Monitoring and Control

There should be a person assigned to be responsible for the entire plan. Tasks should be monitored to ensure they meet the schedule and budget. If a plan falls behind, activities should be rescheduled and budget reallocated to keep the plan moving forward.

VIII. Conclusions

In general, deploying Internet tactics is essential for ZHNCTC to survive and prosper in future. In this paper, three Internet marketing channels, i.e. Search Engine, Social Media and Email, are highly recommended. Through these proposed free and cost-saving approaches, ZHNCTC are sure to increase their website traffic, while with low budget. Meanwhile the enhancement of the website construction will extend the business reach and improve the benefits to both existing and potential customers. And it is obvious that ZHNCTC will increase their revenue and improve customer loyalty by deploying this Internet marketing tactics.

References

[1] D. Chaffey, F. Ellis Chadwick, K. Johnston, R. Mayer, Internet Marketing - Strategy, Implementation and Practice, 3rd Ed., England: Pearson Education Limited, 2006, pp. 349-446.

[2] J. Nielson, Designing Web Usability: The Practice of Simplicity, Indianapolis: New Riders Publishing, 1999.

[3] J. Nielson, H. Loranger, Website Optimization: Construct User-satisfied Website by Improving Web Usability, L. Zhang Transl. Beijing: Electronic Industry Press, 2007.

[4] J. Nielsen, "Top Ten Mistakes in Web Design of 2007," Jakob Nielsen's Alertbox, http://www.useit.com/ alertbox/9605.html, 2007.

[5] A. Jafari, "Putting Everyone and Every Course On-line: The Oncourse Environment," WebNet Journal, Vol. 1, Jan. 1999, pp. 37-43.

[6] N. Turner, "Ten Ways to Improve the Usability of Your Ecommerce Site," http://www.webcredible.co.uk/user-friendly-resources/web-usability/ecommerce-usability.shtml, 2005.

[7] T. Zhang, "Optimization Strategy of Searching Engine for Enterprise Website," Journal of Hubei University of Technology, vol.24, Oco. 2009, pp.61-63.

[8] J. Kyrnin, "Searching Your Site: Adding Search Functionality to Your Web Site," http://webdesign.about.com/od/administration/a/aa091399. htm.

Game Model between Enterprises and Consumers in Green Supply Chain of Home Appliance Industry

Ai Xu, Xiangpei Hu
School of Management
Dalian University of Technology
Dalian, China
e-mail: gdxuai@163.com

Shufeng Gao
College of Computer Science & Technology
Beijing Institute of Technology Zhuhai
Zhuhai, China
e-mail: gsfdl@163.com

***Abstract*—Green supply chain management is a kind of effective management approach for home appliance industry to pursue the strategy of sustainable development and improve the product international competitiveness. In this paper, a game model is proposed to study the relationship and game status between enterprises and consumers in green supply chain of home appliance industry. Through the equilibrium analysis of the game model, some countermeasures are put forward, which can be adopted by the government to improve the construction of the green supply chain of home appliance industry.**

Keywords-Green supply chain; home appliance industry; static games of complete information; Nash equilibrium

I. INTRODUCTION

Since the beginning of the 21st century, to protect the ecological environment and realize sustainable development has become a topic of common concern around the world.

In recent years, many countries in the world formulate more strict environmental protection laws and regulations to strengthen the environmental protection and management.

WEEE directive 2002/96/EC (Waste Electrical and Electronic Equipment, as amended by 2003/108/EC) and RoHS directive 2002/95/EC (Restriction of Hazardous Substances) are the important two regulations relating to household electrical and electronic equipment used by consumers. The objective of the WEEE directive is to improve the level of environmental protection within the European Union through the reduction of waste from household electrical and electronic equipments. Equipment producers are responsible for the management of takeback and disposal of waste starting from 13 August 2005. The RoHS directive aims at harmonization of the legislation in the EU Member States on the restriction of the use of hazardous substances in household electrical and electronic equipment. The general rule is that equipment containing a certain level of lead, mercury, cadmium, hexavalent chromium, PBB's and PBDE's may not be placed onto the market after 1 July 2006. As a main country to manufacture and export home appliances, China will be absolutely affected by these regulations. Therefore, there is no time to delay for Chinese home appliance industry to establish green supply chain, which will be with great significance to improve our product competitiveness in the international market, protect the environment and implement the strategy of sustainable development.

The green supply chain of home appliance industry is more complex than traditional one, with diversifying components and operational objectives, therefore the coordination and management are with great difficulties. The mutual relationship between the main stakeholders, especially the relationship between the enterprises and consumers, will influence the effective implementation of the green supply chain of home appliances.

II. RESEARCH STATUS OF GAME ANALYSIS OF GREEN SUPPLY CHAIN

In 1996, the National Scientific Funds (NSF) in USA provided $400,000 financial aid to the Manufacture Research Consortium (MRC) in Michigan State University to conduct a research project named "Environmental Responsible Manufacture" and then the definition of Green Supply Chain is proposed firstly [1]. Then more and more scholars researched the green supply chain. The so-called green supply chain is a kind of modern management mode which takes the environmental impact and resource efficiency into consideration comprehensively within the entire supply chain. Taking green manufacturing theory and supply chain management technology as the foundation, it involves suppliers, manufacturers, distributors and users, with the purpose to make the environmental impact (negative effect) minimum and resource efficiency maximum during the whole process from material acquisition, processing, packaging, storage, transportation, usage to Scrapping [2]. The basic objective of green supply chain management is to protect environment and make use of resources effectively.

So far, game analysis between the participating subjects in the green supply chain is very limit. Most of the existing research results are about the game analysis between governments and enterprises, e.g., Pantumsinchai (1992) [3], Viswanathan (1997) [4], Zhu and Dou (2007) [5], Li and Liu et al (2007) [6], Xu and Zheng (2008) [7] and Cao and Wen (2011) [8] studied the relationship between governments and enterprises in the green supply chain, set up game models and analyzed their behavior and equilibrium strategies.

Some researchers analyzed the relationship between the enterprises in green supply chain, e.g., Wang and Yan (2009) [9], or the relationship between enterprises in green supply

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.46

chain and those in traditional one, e.g., Hou and Wang (2010) [10].

There are also a few scholars tried to conduct preliminary analysis of the multilateral game relationship between government, enterprises and consumers in the green supply chain, e.g., Wang (2004) [11] and Yu and Liu (2011) [12].

Some scholars analyzed the game relationship in the green supply chain of specific industries, e.g., Zhou and Zhang (2007) [13] and Feng and Wang (2010) [14].

It can be found that the research results about the game analysis between the main stakeholders is so limit and there is no research result found about the game relationship analysis connected with the background of home appliance industry. Concerning of the pressure from the more strict environmental rules such as WEEE and RoHS, it will be essential to study the game relationship between the main stakeholders in the green supply chain of home appliance industry, which will be helpful to promote the construction and development of the green supply chain.

III. A GAME MODEL BETWEEN HOME APPLIANCE ENTERPRISES AND CONSUMERS

A. Basic Assumptions and Definitions

The construction of the green supply chain of home appliance industry requires the involvements of the enterprises, consumers, the government and the society members. In order to facilitate analysis, here we assume the model only include two major stakeholders, i.e. home appliance enterprises (simply as "enterprise") and consumers. Enterprises refer to those engaged in home appliances manufacturing or sales and consumers get home appliances from enterprises. Meanwhile, we assume that both enterprises and consumers are rational economic men, who take the benefit maximization as their goal. The players of the game know the strategies and payoff with each other. In short-term equilibrium, the game problem can be regarded as a kind of static games of complete information and to seek Nash equilibrium.

In the current market conditions with green home appliances and traditional home appliances coexist, the enterprises have two strategies: one is to offer green home appliances, e.g., home appliance manufacturers actively develop ecological design and introduce some available technologies to manufacture green home appliances, or retailers actively promote and sell the green home appliances; the other one is to offer traditional home appliances by using the traditional methods to design, manufacture and sale home appliances, and in this circumstance enterprises will be punished by the government and pay for penalty because of not meeting the requirements of environmental protection. Consumers also have two strategies to choose: i.e., acceptance or rejection. Here we assume that consumers should return the waste home appliances while they decide to accept green home appliances, and they can get an amount of compensation, which just reflect the current policy of "trading in old appliances for a new one".

According to existing literatures [3-14], we make some assumptions and definitions about the benefits and costs for enterprises and consumers while adopt different strategies. While enterprises choose to offer traditional home appliances, R_B and C_B respectively represent the total revenues and total costs, and F_B represents the payable penalty to the government due to not meet the requirement of implementing green supply chain. While enterprises choose to offer green appliances, R_B' and C_B' respectively represent the total revenues and total costs in this case, and S_B represents the subsidy that the enterprises can obtain from the government. As for consumers, according to western theory of cost-benefit analysis, the benefits include indirect benefits and direct benefits. Indirect benefits can be expressed by the difference between the price consumer willing to pay psychologically and the actual one, and R_C and R_C' respectively represent the indirect benefits from traditional home appliances and green ones. Here direct benefit mainly refers the subsidy S_C while the consumers return the waste home appliances at the same time of choosing green home appliances. Here, the revenue may be increased while enterprises choose to offer green home appliances, but the cost will be increased correspondingly. Normally there will be such relation as $R_B'-C_B' < R_B-C_B$, because if not, enterprises will inevitably choose to offer green appliances, because they can not only obtain an extra subsidy, but also not to be punished by the government. Meanwhile $R_C'+S_C >0$, because our intention is to motivate consumers to accept green home appliances, therefore the benefits to consumers including the subsidy should be at lease greater than zero.

B. Payoff Matrix of Enterprises and Consumers

Based on the above assumptions and definitions, we set up a game model between enterprises and consumers, which is represented by the payoff matrix shown in Table□. The payoff matrix shows the two players of the game (enterprises and consumers), strategies (B_1, B_2 for enterprises and A_1, A_2 for consumers) and their payoffs with every possible combination of actions.

TABLE I. PAYOFF MATRIX OF ENTERPRISES AND CONSUMERS

		Enterprises	
		Offer Green Home Appliances (B_1)	Offer Traditional Home Appliances (B_2)
Consumers	Acceptance (A_1)	$R_C'+S_C$, $R_B'-C_B'+S_B$	R_C, $R_B-C_B-F_B$
	Rejection (A_2)	0, $-C_B'+S_B$	0, $-C_B-F_B$

IV. EQUILIBRIUM ANALYSIS OF THE GAME MODEL BETWEEN ENTERPRISES AND CONSUMERS

Now we conduct detailed equilibrium analysis of the game model proposed above.

A. Pure Strategy Nash Equilibrium

1) If $R_B'-C_B'+S_B>R_B-C_B-F_B$, then as for enterprises, this means the payoff of offering green appliances surpasses the amount of offering traditional appliances. This condition can be expressed in another way, i.e., $(C_B'-C_B)-(R_B'-R_B)-S_B<F_B$, which indicates that the increased cost of offering green appliances are compensated by the increased revenue and government subsidies and the surplus is less than the penalty F_B. Because we have assumed $R_C'+S_C>0$, thus the game has a unique pure-strategy Nash equilibrium. The stable strategies combination is (Acceptance, Offer Green Home Appliances).

2) If $R_B'-C_B'+S_B<R_B-C_B-F_B$, then as for enterprises, this means that the payoff of offering green appliances is less than the amount of offering traditional appliances. This condition can also be expressed in another way, i.e., $(C_B'-C_B)-(R_B'-R_B)-S_B>F_B$, which means that the increased cost of offering green appliances will be greater than the penalty F_B even if have being compensated by the increased revenue and government subsidies. Next we consider two circumstances:

a) If $R_C>0$, i.e., consumers can gain benefit from traditional home appliances, then there will be a unique pure-strategy Nash equilibrium and the stable strategies combination will be (Acceptance, Offer Traditional Home Appliance). In this case, the penalty from the government maybe lesser, therefore driven by interests, enterprises will choose to offer traditional appliances.

b) If $R_C<0$, i.e., the benefit to consumers of acceting traditional home appliances is nagetive, then it will be discuss further as following two circumstance:

If $-C_B'+S_B<-C_B-F_B$, i.e., $C_B'-C_B-S_B>F_B$, which means that increased cost of offering green appliances can not be compensated by increased revenue because of consumers' rejection and will be greater than the penalty F_B even if have being compensated by the government subsidies. In this case, enterprises would rather choose to offer traditional home appliances and pay for the penalty than offer green home appliances. Then the game has unique pure-strategy Nash equilibrium and the stable strategies combination will be (Rejection, Offer Traditional Home Appliance).

If $-C_B'+S_B>-C_B-F_B$, i.e., $C_B'-C_B-S_B<F_B$, which indicates that the increased cost of offering green appliances is less than the government penalty F_B after being compensated by the government subsidies, although there is no revenue from the sale. It is undoubtedly that the enterprises would be willing to offer green home appliances and then consumers would change their strategy again. Therefore there will be no pure-strategy Nash equilibrium and the players will choose mixed strategies.

B. Mixed Strategy Equilibrium

If no pure-strategy Nash equilibrium exists, the players will choose mixed strategies, where a pure strategy is chosen at random, subject to some fixed probability. Here we assign consumers the probability p_1 of playing A_1 (Acceptance) and $(1-p_1)$ of playing A_2 (Rejection), and assign enterprises the probability p_2 of playing B_1 (Offer Green Home Appliances) and $(1-p_2)$ of playing B_2 (Offer Traditional Home Appliances), where $0\le p_1\le 1$ and $0\le p_2\le 1$, then the expected revenue for enterprises and consumers can be determined by the following equations.

$$E_B(p_1,p_2)=p_2[p_1(R_B'-C_B'+S_B)+(1-p_1)(-C_B'+S_B)]+(1-p_2)[p_1(R_B-C_B-F_B)+(1-p_1)(-C_B-F_B)] \quad (1)$$

$$E_C(p_1,p_2)=p_1[p_2(R_C'+S_C)+(1-p_2)(R_C)]+(1-p_1)[0\,p_2+0(1-p_2)] \quad (2)$$

$$Make\ \frac{\partial E_B}{p_2}=0,\ then \qquad p_1=\frac{(C_B-C_B')+S_B+F_B}{R_B-R_B'}. \quad (3)$$

$$Make\ \frac{\partial E_C}{p_1}=0,\ then \qquad p_2=\frac{1}{1-\dfrac{R_C'+S_C}{R_C}}. \quad (4)$$

1) From (3),we can safely draw the following conclusions.

a) While $R_B>R_B'$ *and* $(R_B-R_B')>(C_B-C_B')+S_B+F_B$, p_1 *will be meaningful.* Since $R_B'-C_B'+S_B<R_B-C_B-F_B$, i.e., $(R_B-R_B')>(C_B-C_B')+S_B+F_B$, meanwhile $-C_B'+S_B>-C_B-F_B$, i.e., $(C_B-C_B')+S_B+F_B>0$, therefore $R_B-R_B'>0$, and now p_1 will be meaningful and mixed strategies will be feasible. If not, enterprises would rather choose pure strategy "Offer Green Home Appliances".

b) P_1 *is a increasing function of* S_B .When S_B increases, enterprises would be willing to offer green home appliances in order to obtain subsidy from the government to compensate the costs. Then the probability of consumers to accept green home appliances will increase correspondingly along with the increase supply of green home appliances onto the market.

c) P_1 *is a increasing function of* F_B .When F_B increases, enterprises would be willing to offer green home appliances in order to avoid to be punished. Then the probability of consumers to accept green home appliances will increase correspondingly.

d) P_1 *is a increasing function of* C_B-C_B'. Since $C_B<C_B'$, so $C_B-C_B'<0$. Then when C_B-C_B' increases, the cost difference between offering green appliances and traditional ones decreases, which means there will be not too much cost increase for offering green appliances. Taking further considering of the government penalty, enterprises would have more motive power to offer green home appliances. Then the probability of consumers to accept green home appliances will increase too.

e) P_1 is a decreasing function of $R_B - R_B'$. Since $R_B - R_B' > 0$, then when $R_B - R_B'$ decreases, which means offing green home appliances would not decrese the revenue too much. Therefore enterprises would choose to offer green appliance and the probability of consumers to accept green home appliances will increase similarily.

2) From (4),we can safely draw the following conclusions.

a) Since $R_C < 0$, p_2 will be meaningful only while $R_C < R_C' + S_C$. This means that only when benefits from accepting green home appliances surpass the amount from accepting traditional ones, the consumers will possibly choose to accept green home appliances and the mixed strategies will be feasible. This also indicates that the resonable price of green home appliances and effective subsidy will affect the adoption of mixed strategies.

b) p_2 is a increasing function of R_C'. More R_C' means more benefits to consumers and the probability of accepting green home appliances will be increasing. Affected by the increasing demand of green home appliances, enterprises would be willing to offer more green home appliances to the market. Therefore p_2 will increase.

*c) p_2 is a decreasing function of R_C .*When R_C increases, consumers will have more tendency to accept traditional home appliances and enterprises would be driven to choose offering traditional home appliances.

V. CONCLUSIONS

In general, in the static game of complete information between enterprises and consumers in the green supply chain of home appliance industry, there are three pure-strategy Nash equilibriums, i.e., (Acceptance, Offer Green Home Appliances), (Acceptance, Offer Traditional Home Appliance) and (Rejection, Offer Traditional Home Appliance). It is obvious that the solution of (Acceptance, Offer Green Home Appliances) is what we desire. The conditions are as follows:

$$\begin{cases} R_B' - C_B' + S_B > R_B - C_B + F_B \\ R_C' + S_C > 0 \end{cases} \tag{5}$$

To meet the above conditions, the government should carry out more preferential policies to ensure consumers to gain a certain amount of subsidy S_C in order to encourage them to accept green home appliances and ensure the enterprises to gain subsidy simultaneously so as to compensate their increased cost while choose to offer green products. At the same time, the penalty F_B should be big enough so that the punishment for those enterprises choosing to offer traditional home appliances should be effective.

In the long run, government will play a very import role in the construction of green supply chain of home appliance industry. The government should begin with improving laws and regulations about environmental protection, cultivating green price system and green market in order to create favorable external conditions to improve the construction of the green supply chain of home appliance industry. Meanwhile the role of market in the aspect of resource allocation should effectively come into play and lead the enterprises in consciously pursuing scientific management approaches and techniques of green supply chain. Only by doing so, the home appliance industry can realize sustainable development.

REFERENCES

[1] R. B. Handfield, S. V. Walton, S. A. Melnyk, and L. Seegers, "Green supply chain: best practices from the furniture Industry," Proceedings of the Annual Meeting of the Decision Science Institute, Nov. 1996, pp. 1295-1297.

[2] B. Dan and F. Liu, "Study on Green Supply Chain and Its Architecture," China Mechanical Engineering, vol. 11, Nov. 2000, pp. 1233-1236.

[3] P. Pantumsinchai, "A Comparison of Three Joint Ordering Inventory Policies," Decision Science, vol. 23, Jan. 1992, pp. 111-127.

[4] S. Viswanathan, "Periodic Review (s, S) Policies for Joint Replenishment Inventory Systems," Management Science, vol. 43, Oct. 1997, pp. 1447-1454.

[5] Q. H. Zhu and Y. J. Dou, "An Evolutionary Model between Governments and Core-enterprises in Green Supply Chains," Systems Engineering-Theory & Practice, vol. 27, Dec. 2007, pp. 85-89.

[6] J. Y. Li, W. L, and G. P. Cheng, "Dynamic Game Analysis about Reverse Logistics between Government and Enterprise," Chinese Agricultural Mechanization, vol. 6, Nov. 2007, pp. 24-27.

[7] W. Xu and Y. F. Zheng, "Game Analysis between Governments and Corparations with Prosecution in Green Supply Chain Management," Chaniese Journal of Management Science, vol. 16, Oct. 2008, pp. 450-454.

[8] H. Y. Cao and X. Q. Wen, "Government-oriented Green Supply Chain Management Study Based on the Game Analysis," China Business and Market, No. 2, Feb. 2011, pp. 33-37.

[9] S. L. Wang and G. L. Yan, "Evolutionary Model between Suppliers and Core-enterprise in Green Supply Chains," Science-Technology and Management, vol. 11, May. 2009, pp. 59-62.

[10] L. N. Hou and H. Y. Wang, "The Evolutionary Game Analysis on Sustainable Strategies of Green Supply China and Traditional Supply China," Technology and Innovation Management, vol. 31, Nov. 2010, pp. 677-680.

[11] Y. E. Wang, "A Study on Setting Up Reverse logistic System of Enterprise", Chengdu: Southwest Jiaotong University, 2004.

[12] S. Q. Yu and J. Liu, "Game of the Government, Enterprise and Consumer in Green Supply Chain System," The World & Chongqing, vol. 28, Jan. 2011, pp. 45-47.

[13] M. Zhou and S. W. Zhang, "Three Players Game Analysis between Government, Coal and Electricity Power," Coal Economic Research, No. 4, Apr. 2007, pp. 33-34.

[14] Y. H. Feng and S. L. Wang, "Game Analysis for the Construction of Green Supply Chain in Construction Industry," Journal of Xi'an Univ. of Arch. & Tech.(Natural Science Edition), vol. 42, 2010, pp. 674-678.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Management Stratagem and Models in Emergence Response System

Anna Dai
Department of Construction Management and Real Estate
Tongji University, TJU
Shanghai, China
dansyf@163.com

Da Xu
Department of Supporting and Development
Shanghai Renjie Riverside Real Estate Co. Ltd.
Shanghai, China
cherryjackie@163.com

***Abstract*—This paper mainly studies the emergency management decision-making model of city government. Decision-making behavior of the government in emergency management is analyzed from the perspective of decision objective, decision criteria and decision-making behavior. Then, the principles of emergency classification are discussed and a city emergency classification and grading system is established accordingly. Eventually, dynamic fuzzy model of city emergency incident classification is established to lay a sound foundation for emergency management stratagem.**

Keywords-management stratagem; decision-making behavior; dynamic fuzzy model; emergency system

I. INTRODUCTION

Government emergency management (EM) refers to the related activities during the process of emergency prevention, incident response, and recovery when the government takes series of necessary measures to protect public lives and property, promote social harmonious healthy development through the establishment of necessary response mechanisms. EM is the whole process management of public emergency according to public emergencies of four stages, namely early warning, occurrence, mitigation and rehabilitation. EM can be divided into four parts: prevention and preparedness, monitoring and early warning, emergency response and rescue, recovery and reconstruction after the process.

City EM is management procedures and methods preparing for emergencies, rescuing and recovery. Through pre-disaster plans and emergency measures and by using all possible force points, disaster development should be controlled immediately after it happens. With residents rescuing, disaster mitigation and consequences control, disaster losses will be minimized. EM is a dynamic management, the four stages of EM interface to constitute the measure layer, and to withstand disasters in order to ensure the realization of the target layer, namely city public safety [1]. EM is a complete engineering system that can be summarized as public emergency contingency plans, emergency response mechanisms, institutions and legal system.

II. THE MAIN BODY ANALYSIS OF CITY EM

A. Main Body of the City EM

The main bodies of city EM are organizations, institutions and personnel dealing with unexpected events. The main bodies of EM are relevant. Before the occurrence of unexpected events the city should fully prepare and plan. After the emergency, they should control the situation and reduce disaster losses. Later, we should deal with the aftermath and rescue, as well as try to establish a scientific and careful emergency prevention system. From the perspective of decision-making model, we should first determine the system's decision makers. City emergency response actors include: the government, the affected population, rescuers and related media.

B. Behavioral Characteristics of the Affected Population

In the types of the behavior bodies above, we should focus on the government in particular. Figure 1 can express government behavior and the behavioral characteristics of the affected population under the conditions of urban emergency. Decision-making behavior of the decision making bodies have the following characteristics:

1) The Incompleteness of Decision-making Information

In essence, the occurrence of unexpected events the city is difficult to predict, and its evolution is very complex, involving multiple decision actors. In this process, the decision makers cannot get sufficient information in a relatively short time, so the information is of high uncertainty and asymmetry.

2) The Timeliness of decision-making

Under emergency conditions, both government and other relevant stakeholders are required to make the right reasonable decisions and timely responses under incomplete information.

3) Game Behavior and Compound Group Decision-making

Game stresses the existence of multiple decision-makers target their own interests, and there is the strategy or conduct of their own choices. Each of these strategies or behavior is with interactive features. City EM involves the central government and all levels of government, the public, community agencies and other emergency organizations. Even the executive branch of government has a relationship with public security, health, labor, urban management, and civil affairs departments---a comprehensive co-ordinate among the decision-making bodies to better serve the city EM services.

4) Dynamic Decision-making

In the city EM decision-making behavior, along with the evolution of emergency, the decision-making body

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.56

constantly absorbs new information to constantly modify the a priori assumptions. Meanwhile, with the continuous evolution and development of emergency needs at each key decision point, there is no once and for all decision-making behavior. Therefore, the city emergency decision-making behavior is also dynamic.

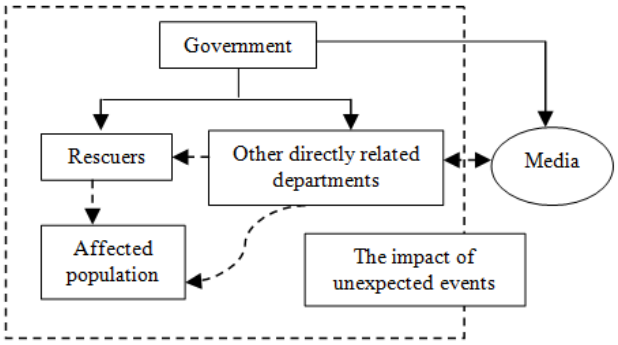


Figure 1. The relationship of behavior bodies under urban emergency

III. GOVERNMENT EMERGENCY DECISION MAKING

A. Government Policy Objectives

In the unexpected event, all levels of government often face multiple decision-making objectives, such as economic and cultural development, infrastructure construction, and social security, etc. The emergency preparedness and prevention can be only one of many economic goals. With the evolution and the loss of specific events, the government will increase the attention on unexpected events, or even fully respond to emergencies, as shown in Figure 3-2.

TABLE I. GOVERNMENT EM GOAL UNDER EMERGENCY OCCASIONS

Objective of government EM	Improve emergency response capabilities (Normal state)	Improve the predictive ability of relevant emergency
		Establish a sound system of laws and regulations
		Improve infrastructure reliability and resource ensuring level
		Improve complete and scientific contingency plans
		Strengthen disaster prevention culture
	Response, reducing and elimination of disasters (Emergency)	To ensure maximum safety of personnel
		Strengthen infrastructure construction and rehabilitation
		Accurate and rapid assessment of emergency
		Raise the level of information dissemination
		Improve the capability of disaster control, reduce and eliminate
		Reduce the negative impact on government and society

In addition to emergencies caused by natural properties, highlighting the government performance and reduce the negative government response to the adverse effects from the perspective of government action preferences, increase the effectiveness of other aspects (such as increasing government revenue, increasing other aspects of economic and social development level of facilities, etc.) and other pursuits, in some extent, can also influence government decisions on the objectives of emergency.

B. Government Decision-making Criteria

The uncertainty and complexity of unexpected events makes the government tend to take a rational decision-making model to minimize the impact of psychological factors in their decisions. Therefore, the decision-making criteria adopted should include the following.

1) The Maximum Expected Utility Criterion

The expected utility criterion is the cornerstone of classical decision theory and is the most comprehensive decision model theory approach. In actual decision-making applications, due to a series of theory axioms of lottery selection involved in the decision of decision-makers, its application in practice is limited.

2) Strict Uncertainty Decision

The so-called strict uncertainty decision problem is uncertain and decision makers can know what the state of nature may be, so it can only be possible to give a list of the state, ignoring the size of the possibility in emergency of various states. That is to say the probability of various natural states cannot be estimated [2], including pessimistic criteria, optimistic factor criteria.

C. Government Decision Making

Government decision-making behavior for emergencies covers infrastructure construction, emergency prevention and control, preparation of contingency plans, each of which involves a number of specific issues and affairs from the details of the discussion. Figure 2 shows the city EM decision making process related to the main content of the government.

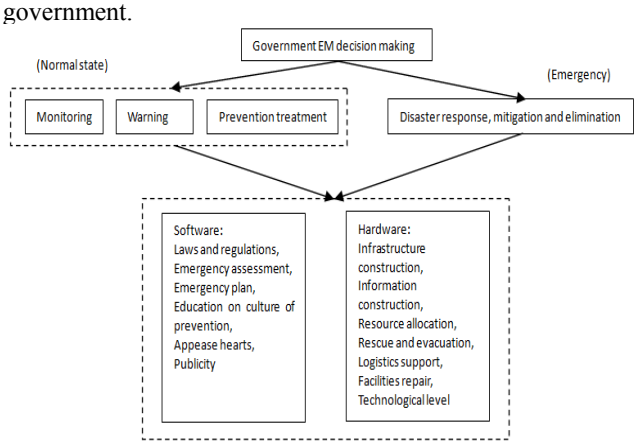


Figure 2. Government management decision under emergency

From the different attributes of things to be solved, the implementation of government decision-making can be divided into decision making and strategy-based decision-making. Implementation of decision making related to

specific emergency projects response, while strategic decision making is for general problems in emergency decision-making, as shown in table Ⅱ.

TABLE II. EXECUTIVE DECISIONS AND STRATEGIC DECISIONS

	Decision making	Strategy-based decision-making
Decision-making body	Management	Decision-making
Decision object	Specific issues in emergencies	The general question of emergency
Information release	Generally do not need	Open to departments and the public
Main participants	Technical staff based	Senior decision-makers based

IV. CLASSIFICATION OF DECISION-MAKING MODEL

Classification of urban emergency is the main government EM decision-making tasks. Based on the type and level of emergency, the government takes measures of great significance for the government to adopt a scientific and rational decision and reduce disposal costs of the event. Emergency of scientific urban classification is not only the basis of EM, but also the premise of the development of contingency strategies and plans.

Classification is based on the principles of urban emergency and according to the nature of the city emergency can be divided into five categories [3], for detailed descriptions of each category are as follows.

a. Natural Disaster. This includes flood and drought disasters, meteorological disasters, geological disasters and forest fires and major biological disasters.

b. Accidents category. This includes major transport event, all kinds of major security incidents, and the loss of a significant impact on urban lifeline events, nuclear radiation events, major environmental pollution and ecological destruction events.

c. Categories of public health emergencies. This mainly refers to the sudden onset, caused or may cause the project participants or serious harm to the health of person, the major infectious diseases, mass unexplained diseases, etc.

d. Emergency social safety event class. This includes major criminal cases, foreign emergencies, terrorist attacks, unrest, rebellion and other large-scale mass incidents.

e. Economic crisis category. Mainly refers to resources, energy and life, producing a serious shortage of essential items, the financial impact of the credit crisis and other serious financial disorder.

According to the Government's capacity to respond, the scope of emergency, the loss of the consequences and other standards, the city emergency can be divided into four levels, namely, in general, large, important and very important. For a detailed description of the type of urban types of emergencies, we constructed a city emergency classification three-dimensional diagram, as shown in Figure 3.

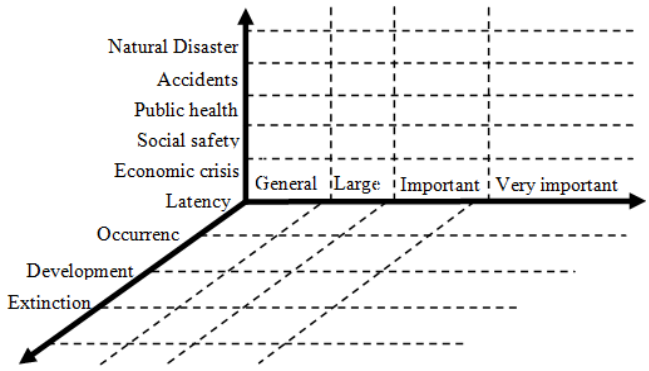


Figure 3. **Classification and cluster of urban emergencies**

V. THE DYNAMIC FUZZY MODEL

Previous studies on the classification of emergency classification methods are static [4] [5] [6], some work is retrospective evaluation of the incident; in the development of EM guidance is not strong enough. The development of emergency situation is constantly changing, and need dynamic classification work. Dynamic classification is based on information available the situation now to make some emergency-level assessment of future point.

Because it is assessment of the state for the future, the dynamic classification is uncertain. Multi-expert scoring has brought the event elimination of a single expert in subjective bias, and with fuzzy numbers to represent the weight factors and the assignment can be more accurate response to uncertainty. Figure 4 describes the basic flow of dynamic fuzzy classification model.

After the incident, firstly obtain the type of event under the k-means algorithm and k-modes algorithm, and then based on the mechanism analysis and analysis of previous cases to determine the dynamic classification elements and the selected factors; using the method of multiple experts scoring given each factor triangular fuzzy number. The left minimum in triangular fuzzy numbers mean experts believing that this factor may be the minimum, the middle number indicate the most likely value while the maximum right value indicate the possible maximum. Then, obtain the fuzzy weight of each factor based on the fuzzy Delphi [7], and obtained the fuzzy comprehensive evaluation of the optimal solution according to Kerre [8]. Lastly, using pessimistic decision-making principle, select the most pessimistic value as a final determination of the optimal solution value from the fuzzy values, and make use of the corresponding level of classification to determine the final level.

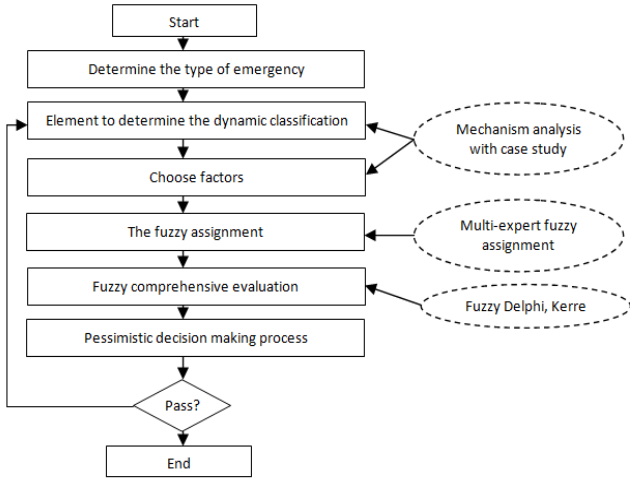


Figure 4. **Program for dynamic fuzzy classification algorithm**

Determining the level of future time calculate state transition probability of future time in order to verify the accuracy of determining the dynamic level. Development of sudden events is random, assuming that the level of emergency numbers have "no after-effect", and the time point in calculating the event level is discrete, so all levels of the state constitutes a Markov chain. Given past value experience in state transition matrix and in accordance with Komogorov-Chapman equation [9], state transition probability in arbitrary discrete-time points can be calculated.

Dynamic fuzzy hierarchical fuzzy comprehensive evaluation consists of three main parts: identification of fuzzy weights, determine the optimal solution (the most realistic emergency needs of the program) and the calculation of state transition probability.

Determine the fuzzy weight:

a. A number of experts score the importance between the two selected factors, a number of expert evaluation weight can be derived by AHP.

b. Get the fuzzy weight of each factor by fuzzy Delphi method.

The formula to determine fuzzy weight:

$$W=(e_i, f_i, g_i)$$

W---the fuzzy weight of factor i, w is the certain weight of factor i by expert j.

$$e_i = \min \omega_i^j, g_j = \max \omega_i^j, \quad f_i = \{\prod_{j=1}^{n} \omega_i^j\}^{\frac{1}{n}}$$

c. Determine the best solution.

First of all, normalize the value of fuzzy factor matrices. Secondly, on the normalized basis, the program gives each expert to determine the right fuzzy maximum set of factors. Thirdly, Hamming distance between the factors that determine the various options in the right-fuzzy values and the maximum right fuzzy set.

Factors obtained in the steps above, then determine a right fuzzy maximal set and calculate the Hamming distance between these factors and the right of great fuzzy set with the principle above.

d. Select the smallest Hamming distance of the corresponding program elements, and make use of Bonissone approximate formula [10] to calculate the final score value of the program. Bonissone approximate formula is used to calculate the product of two fuzzy numbers:

$$wj * xij = (ac, a\zeta + a\alpha - \alpha\zeta, a\delta + c\beta - \beta\delta)$$
$$wj = (a, \alpha, \beta), \quad xij = (c, \delta, \zeta)$$

Use pessimistic fuzzy principles to make the right fuzzy value of triangular fuzzy number of value as the final assessment level.

e. Calculate the state transition probability. Using past cases and expertise to determine the state transition matrix, according to the principle of Markov chain and Komogorov-Chapman equation, calculate the state transition probability of discrete time points.

f. Test whether the largest transfer of state (level) of the state transition probability matrix, meet assessment results of the initial level of the previous steps. If yes, algorithm end; if not, further feedback adjustment is needed.

By comparing the results of State transition probability and the dynamic fuzzy classification algorithms, it can further demonstrate the dynamic accuracy of fuzzy classification algorithm. If the above two results are inconsistent, then it need to take steps to find the reason, adjusting the state transition matrix of the experience or expert scoring assignment.

ACKNOWLEDGMENT

I want to give my deepest appreciation to Tongji University for offering me such a good environment for studying and communication. And I want to give my deepest appreciation to my partner and my tutor who enlightened me when I am stuck. Thank you all, my friends.

REFERENCES

[1] Feng Kai, XU Zhisheng, Feng Chunying, Wang Dongsong. Urban public safety planning and the integration of disaster emergency management (J). Natural Disasters, 2008 (4): 17 ~ 19.

[2] Yue Chaoyuan. Decision theory and methods. Science Press. Beijing. 2009: 71 ~ 78.

[3] Robert Heath [US]. Crisis management, CITIC Publishing House, 2007: 12 ~ 15.

[4] Department of State. Countries the general public emergency contingency plans. January 8, 2006 issued.

[5] Wu Zongzhi, Liu Mao. Contingency plans for major accident classification, classification system and its basic content. China Safety Science Journal. 2007: 80 ~ 82.

[6] Lv Xin Chi. Classification of urban fire and rescue force response plan. Shenyang Fire Research Institute Ministry of Public Security. 2009.11: 67 ~ 71.

[7] Han Shilian, etc. Multicriteria Fuzzy AHP Model Evaluation and Selection of Logistics Center. Systems Engineering Theory and Practice. 2010.7: 50 ~ 53.

[8] Fang Zhaoben. random process. China Science and Technology Press, 2008: 41 ~ 45.

[9] Kerre, EE. The use of fussy sets theory in electrocardiogram in MM. Guptaand E. Sanchez. “Approximate Reasoning in Decision Analysis”, 2009:301~312.

[10] Bonissone P.P. M.T. Lamta and S Moral Decision making problems in a general environment, “Fuzzy Sets and System”. 1988: 110~115.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Game of The Parties to The Transaction Taobao

Chang Qing
School of Management
Inner Mongolia University of Technology
Hohhot, China
changqingimut@126.com

Suo Hong, Jiao lei
School of Management
Inner Mongolia University of Technology
Hohhot, China
suohong_101@163.com

***Abstract*—In this article the trading behavior of buyers and sellers in Taobao was analyzed from the perspective of game theory. The three trade expectations and their possible trade links and the corresponding solutions and the transaction model of Taobao Game of the parties were proposed. Moreover the measures of successful e-commerce transactions can not do without the protection of various effective was presented, which could prevent the occurrence of fraud, but also prompt the formation of Nash equilibrium game model, so that trading between the two sides was forced to choose to complete the transaction integrity of the transactions.**

Keywords-Game; Nash equilibrium; Transactions; Taobao

I. INTRODUCTION

With the rapid development of information technology and the popularity of the Internet, online shopping has become more and more consumer choice. And Taobao in terms of the number of registered users, or sales, have become China's largest C2C online trading platform, was first set up Taobao in 2003 was 30 million registered users, with annual sales of 7.3 billion, to In 2010, registered users has reached 370 million, with annual sales of 4,000 billion[1]. From these data we can see that the attitude of consumers shopping online has been suspected from the past into the present trust, the reason is mainly Taobao through a series of measures to continuously improve its market environment for trading Both sides in the online shopping process can be effectively protect their own interests, effective way to eliminate the occurrence of fraud, and create a shopping atmosphere of good faith. This article will analyze Taobao conduct of the parties in the transaction from the perspective of game theory, and find game ideas which are behind the various measures of the transaction process.

II. C2C E-COMMERCE REVIEW THE GAME

C2C e-commerce at home and abroad on research concerning the game are mostly focused on how to credit rating and reputation by establishing mechanisms to achieve integrity aspects of trading. Zeng Yong, Mao Xu Wei is the first from the perspective of the game between the seller and C2C transactions, credit analysis mode selection [2]. Without integrity behavior of the buyers and sellers the most fundamental reason is that the asymmetry of information, effective way to solve this problem is to build credibility evaluation system. In China, for the credibility of the system of C2C are: Jingbo Xian, Xu Feng, reputation-based trust model design and analysis [3]; Pu Chunhui, quiet, Fang Meiqi analysis of the shortcomings of existing evaluation model was established to improve the C2C e-Credit evaluation model business website and a new credit evaluation algorithm [4]. Abroad, Sulin Ba proposed: asymmetric information (information asymmetry) is the leading C2C e-commerce fraud in the primary cause of [5]. For the credibility of the System, Cynthia G. McDonald and V. Carlos SlawsonJr on eBay, the personal credibility of individual transactions and the issue of an empirical study, found that those with high credibility and low credibility in the price and there are different aspects of behavior[6].

In summary, for the Game of C2C e-commerce, whether domestic or foreign, mainly for the evaluation process and the reputation of the cumulative part of a study to solve the problem of asymmetric information, but very few buyers and sellers on the behavior of the transaction process Analysis. However, the author's personal experience of many Taobao buy and read a lot of relevant information found: transaction buyers and sellers have been able to succeed mainly because of the transaction process both are in a Nash equilibrium of the game model, and Taobao in On completion of the transaction of various measures to promote the formation of a Nash equilibrium.

III. NASH EQUILIBRIUM BASED GAME TAOBAO BUYERS AND SELLERS TRADE

Nash equilibrium (Nash equilibrium), also known as non-cooperative game equilibrium, game theory is an important term to John • Nash named. Nash equilibrium is defined as: Suppose there are n individuals to participate in games, strategy given the conditions of other people, each person to choose their own optimal strategy (optimal strategy may depend on the individual may not rely on strategies of others), all they choose the strategy together constitute a strategic combination of (Strategy Profile). Nash equilibrium refers to the combination of such a strategy, this strategy combination of all participants by the composition of the optimal strategy, that is, given the case of other strategies, no single participant has a positive sexual selection and other strategies to no one has the enthusiasm to break this equilibrium[7]. Interpretation of a classic case of Nash equilibrium is the "prisoner's dilemma ", in this case the non-cooperative game theory to lay the theoretical foundation, and it can be used as many real-life phenomenon of an abstraction.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.47

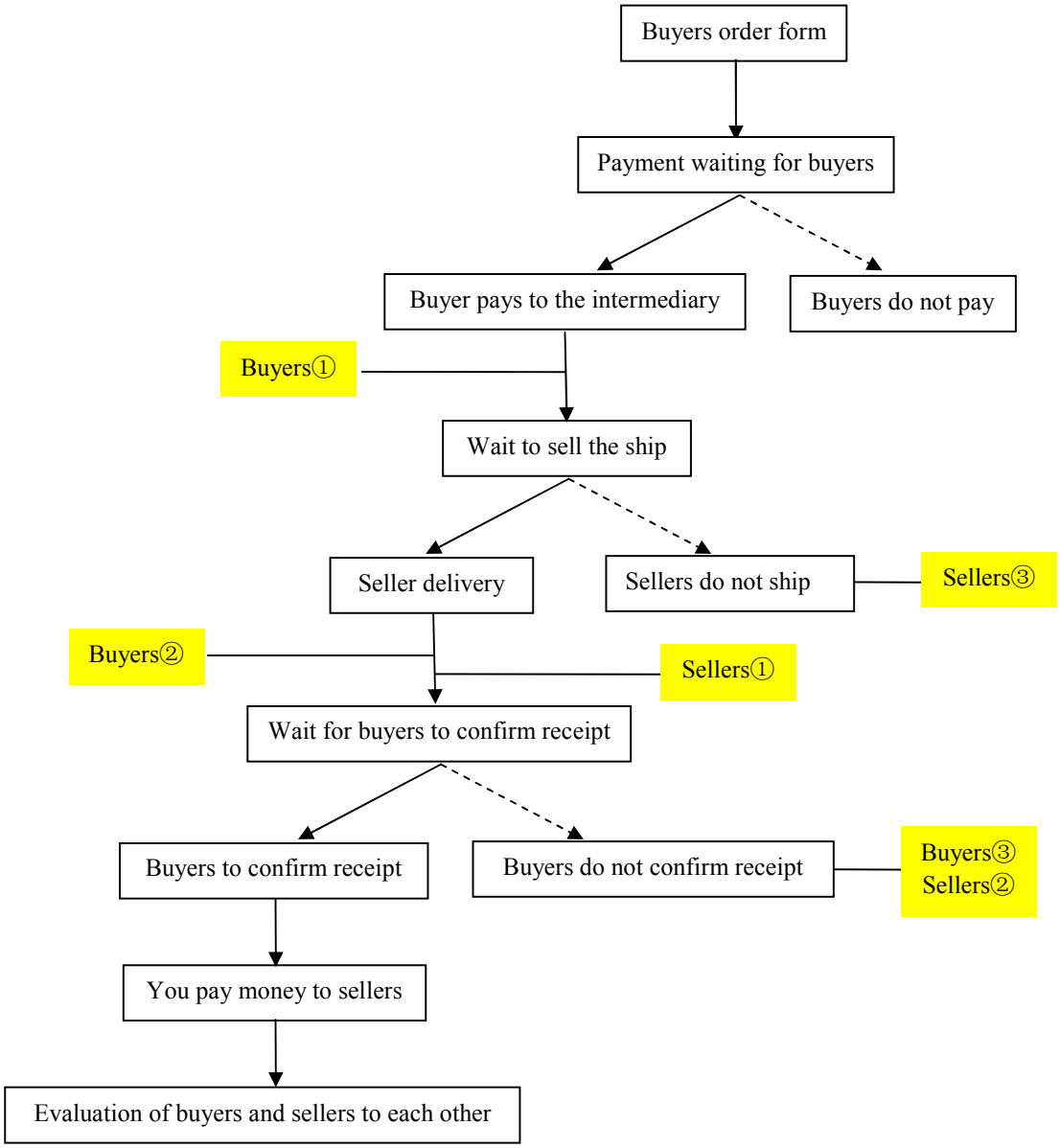


Figure 1. Taobao buyers and sellers transaction process

Fig. 1 Taobao transaction process between buyers and sellers, by the solid arrow path of the transaction is successful.

According to Taobao this transaction process, buyers and sellers on the transaction there are three types of results were expected, as shown in Table Ⅰ. Eliminate the occurrence of the behavior of buyers in the following two transactions may occur in areas: ① "buyer payment to the intermediary, " after the seller does not ship (Fig.1, the "buyer ① " below); ② "buyers pay Money to the agency "after" the seller shipped ", but not within the prescribed time the goods to the buyer in the hands (Fig.1, the "buyer ② "below). Buyer fraud may occur in shipping but the buyer receives the seller fails to transaction procedures, "confirm receipt", but a refund to the Seller (as shown in Fig.1, the "buyer ③ " below). The seller to prevent the behavior may occur in the following two trading links of: ① "seller's delivery" and the goods were to the purchasers, buyers and apply for a refund (as shown in Fig.1, "the seller ① " below) ; ② "seller's delivery" and the goods were to the purchasers, but within the prescribed time, "the buyer does not confirm receipt"(as shown in Fig.1, "the seller ② " below). Seller fraud may occur in the "payment to the intermediary buyers" after "the seller does not ship"(as shown in Fig.1, "the seller ③ " below).

TABLE I. TAOBAO IN THE EXPECTATIONS OF BUYERS AND SELLERS ON THE TRADING RESULTS

Participants / Transaction expected	Buyer	Seller
Eliminate the occurrence of	Payment can not get the goods	Delivery shall not receive
Fraud	Do not pay to get the goods	No payment is received Delivery
Normal	Payment to get the goods	Payment is received Delivery

Taobao trading above expectations for a series of solutions developed to ensure the smooth progress of the transaction, the corresponding measures in Table Ⅱ :

TABLE II. FOR THE EXPECTATIONS OF BUYERS AND SELLERS TRADING SOLUTIONS

	Transaction expected		Solutions
Buyers	***Eliminate the occurrence of***	"Buyers paid to the intermediary, " seller does not ship after (Buyers ①)	At this point the transaction status as "Buyer Payment" wait to sell the ship. " If the seller does not ship, the transaction will remain in a "buyer paid" status, the transaction state is not out, does not automatically play money to the seller, the transaction will not succeed and will not close automatically, you need to buy Home initiative to apply for a refund, to return the money, buyers apply for a refund, the refund if the seller has been untreated, the buyer 2 days after requesting a refund, the system default buyers and sellers have been reached by the buyer refund agreement and refund to the buyer. The transaction time-out state, the system will automatically refund to the buyer.
		"Buyers paid to the intermediary, " after "the seller shipped", but not within the prescribed time to the purchasers the goods (Buyers ②)	At this point the transaction status as "the seller has shipped" waiting for buyers "to confirm receipt. " Buyers can first contact the seller to verify shipments, please help extend the trading out the seller if the seller does not help extend the transaction timeout, the buyer can immediately apply for a refund, to avoid time-out play money, if refund 5 days after the seller does not respond, the system default buyers and sellers to refund the buyer has reached a refund agreement and refund to the buyer.
	Fraud	But the failure to receive the shipment buyer seller transaction process, "confirm receipt", but a refund to the Seller (Buyers ③)	At this point the transaction status as "the seller has shipped" waiting for buyers "to confirm receipt. " Based logistics company, "the goods to the signing" principle, the seller may be related to logistics in the website of the goods reach the buyer is truly the hands of fraudulent refund if the buyer by virtue of the evidence provided by logistics companies refused to refund the buyer or Proceedings to the buyer.
Sellers	***Eliminate the occurrence of***	"Sellers shipping" and the goods were to the purchasers, the buyer and refund (seller ①)	With the "buyer ③ " measures to address
		"Sellers shipping" and the goods were to the purchasers, but within the prescribed time, "the buyer does not confirm receipt" (seller ②)	At this point the transaction status as "the seller has shipped" waiting for buyers "to confirm receipt. " If the buyer within a specified time without any reason and no refund request confirmation of receipt, the system will confirm receipt according to types of goods within the time fixed confirm receipt, and to fight money to the seller.
	Fraud	"Buyers paid to the intermediary, " after "the seller does not ship" (seller ③)	With the "buyer ① " of solutions

Table Ⅱ solutions resulted in the buyers and sellers in the trading game Taobao in the process of the formation of Nash equilibrium, game behavior analysis shows that buyers and sellers: The buyer in the transaction process, there are two active behaviors, namely, "payment do not pay "," confirm receipt, do not confirm receipt "; seller in the transaction process has an active behavior, the "delivery, not delivery. "Assuming both parties are economic, that are in the trading process to maximize their own interests to pursue and eliminate the occurrence of acts set to pay for the -1, 1

payment fraud, payment behavior under normal conditions of 0, due to Taobao Provided for payment only after the buyer, the seller can be shipped, so the case of non-payment buyers will be unable to continue trading. Accordingly, the parties to the transaction Taobao game model shown in Table Ⅲ:

TABLE III. GAME TRANSACTION BUYERS AND SELLERS IN TAOBAO

Payment / Buyer / Seller	Payment	
	Confirm receipt	***Do not confirm receipt***
Delivery	0,0	1,-1
No shipping	-1,1	0,0

Note: ① The first number is the buyer's payment, the second number is the seller of the payment; ② Table 4 in the "payment" refers to payments to the intermediary, then the seller did not receive any money, "Confirm receipt "means you pay money to the seller, the seller really time payment is received.

Analysis of Table Ⅲ shows, the parties to the transaction Taobao Nash equilibrium is: (the buyer does not confirm receipt of payment, the seller does not ship), indicating that each of fraud when buyers and sellers expectations, the transaction is not their dominant strategy. If the transaction continues, we can see from Table Ⅱ, (the buyer to confirm receipt of payment, the seller does not delivery) and (the buyer does not confirm receipt of payment, the seller shipped) have both been Taobao fraud the corresponding solutions are developed to avoid, although this method is to avoid passive, but to ensure the transaction goes smoothly. Therefore, as long as the buyers and sellers in Taobao does not terminate the transaction, even if the process of fraud of any party's expectations will be limited due to various measures, so that they are forced to give up hope, turn to (buyers pay Confirm receipt, the seller shipped) game state, until the transaction is successful.

IV. CONCLUSION

Taobao, as China's largest C2C online trading platform, its comprehensive transaction security mechanism is to win much key Taobao consumer choice. Based on the transactions between buyers and sellers and the transaction process analysis of various measures, summed up the expectations of buyers and sellers each of the three transactions, each expects the trading process and the possible solutions to avoid the expected happened, come to the "Electronic The success of business transactions can not do without the protection of various effective measures, these measures prevent the occurrence of fraud, but also prompted the formation of Nash equilibrium game model, so that buyers and sellers are forced to choose to complete the transaction integrity of the transactions, "the Conclusion.

ACKNOWLEDGMENT

This paper gets the following fund support: The Program for New Century Excellent Talents in University (NCET-08-0869).

REFERENCES

[1] LONDON, http://content.caixun.com/NE/02/ed/NE02edr5.shtm, 2011-1-20

[2] Zeng Yong, Mao Xu Wei. e-commerce between buyers and sellers of credit Mode Game Theory [J]. Science and Technology Progress and Policy, 2004, (12) :147-149

[3] Jingbo Xian, Xu Feng, Wang Yuan, Lv Jian. for C2C reputation-based trust model design and analysis [J]. Computer Applications, 2007,27 (6) :1349 -1352

[4] Pu Chunhui, quiet, Fang Meiqi. C2C e-commerce website credit evaluation model and algorithm [J]. Information Science, 2007, (8) :105-107

[5] Sulin Ba, PAPavlou. Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior [J]. MIS Quarterly, 2002,26 (3) :243-269

[6] Cynthia G McDonald, V Carlos Slawson Jr.Reputation in an Internet Auction market [J]. Economic Inquiry, 2002,40 (4) :633-650

[7] Weiying. Game Theory and Information Economics [M]. Shanghai: Shanghai People's Publishing House, 2004

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Research on Bounded Rationality and Evolutionary Game Analysis of e-business transaction

Chun Chu, De-shan Tang
Business School of Hohai University
Hohai university
Nanjing China
Chuchun126@126.com

***Abstract*—With the rapid development of the global economy today, e-business model play an important role in the rapid development of modern enterprises. However, due to its characteristics of e-business reasons, transactions integrity problems associated with its rapid development and become increasingly highlighted of the rapid development bottleneck. Research the problems in two aspects use the evolutionary game theory under the bounded rational, one aspect is the integrity problems of games between enterprises and another is about the enterprises under government regulators of e-business transaction integrity. Finally, to provide a viable theoretical basis for the formulation of government policy and regulation.**

Keywords—evolutionary game theory; e-business; bounded rationality; the regulatory

I. INTRODUCTION

Along with the economic globalization and the rapid development of science and technology, e-business model has become the main mode of modern business model. Both of their platforms through the network to promote the development of traditional industries and technological upgrading, while new e-business can bring the rise of the network industry in China has formed a good development prospects, living and working in the community playing an increasingly important role. However, due to the model itself characteristics of e-business models, the harm in the trading caused by the credit crisis gradually revealed, and has become the bottleneck of the development of e-business models.

Therefore, this paper use the evolutionary game theory to study evolutionary of e-business transactions in function of integrity management in the construction of research and analysis, to built at present various existing deception and dishonesty in the process of the development, and to establish integrity trading game model and regulatory model, in hopes that the integrity management work for the government provide feasible theory basis.

II. THE EVOLUTIONARY GAME THEORY AND E-BUSINESS

A. Game Theory and Evolution

Game theory is the theoretical study of people decision problem but is different from general rigorous methods, game theory emphasizes the decision-making strategy choice between mutual conditionality, reflected the unity of opposites between cooperation and conflict relationship. The game problem gambling sides as a rational behavior subject called bounded rationality. Strengthen the fundamental realities of game theory method is based on the limited rational game square as the game analysis of foundation. Limited rational first means gambling sides often cannot or will not be adopted under the condition of complete rationality, means that the optimal strategy is the strategy equilibrium between gambling sides is often the result of learning adjustment instead of disposable choice results, and even reached the equilibrium also might again deviation. Therefore, to the traditional theory of completely rationality in the hypothesis bounded rationality questioned, evolutionary game has altered the gambling sides of the completely rational game hypothesis. Based on bounded rationality, the basis of the theory of evolutionary game opened the new direction of development, is the development and extension theory. It is different from traditional game theory that the evolution game research gambling process of human behavior strategy of dynamic adjustment and learning process. The key of evolutionary game analysis is to determine the learning and strategy game adjust model, according to the specific circumstances to constructing dynamic learning evolution model.

The evolutionary game is a traditional game with the dynamic evolution process, it assumes bounded rationality as the premise, and research the evolutionary process of learning and strategy adjustment. Evolution game is often as the dynamic evolution process repeated game, but it compared with the traditional game theory is concerned, have overcome a game in the process of research completely rational from actual shortcomings, and establish rational basis in limited, consequently, evolutionary game makes its theoretical basis more solid, practical stronger. Although evolutionary game theory are with traditional occupies an important position, but they have a fundamental difference. The evolutionary game theory of evolution gamble sides act as a time evolution system, it is important to study the adjustment process game behavior. Traditional game theory research method is for game one action of the message, the key is the main basis for research in expectations game the

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.76

decision result of information under the difference, they mainly displays in:

- Different rational basis.
- Different understanding of the dynamics.
- Different equilibrium concept.
- Different mathematical tools used.
- Different targets.
- Different process to reach equilibrium.

Traditional game theory and evolutionary game theory differences make built on conventional Nash analysis based on the theory of evolutionary game theory are no longer apply, the key is to design a game in the process of people in times of dynamic learning mode game. Therefore, this article will bring e-business integrity into trading mechanism of evolutionary game, then to establish suitable model for e-business enterprises in the transaction of honesty and faith model.

B. Evolutionary Game Theory and E-business

E-business enterprise through the Internet as a platform for business activity patterns have become the mainstream of the majority of enterprises to global business model, through the electronic commerce development condition analysis shows that enterprise mode of electronic business affairs will gradually replace the traditional pattern of the majority of enterprises has become a major operation mode. With the continuous development of the mode of operation mechanism, will gradually get perfect, therefore we can also gradually as the ecological environment evolution and perfect a process. Meanwhile, e-business enterprise in this environment also continues to grow and evolve, and constantly adapt to the change of environment, the strategy has to do with the experience that they obtain the accumulation of gradually evolve.

In this case, we could bring the e-business environment as a big group game environment, in unified environment under the rules, e-business enterprises as large groups of individual members, they repeated game between random pairing, equivalent to e-commerce enterprise groups between different individual in long-term repeated exchanges and cooperation, or a small group of small and medium range between individuals of mutual learning and cooperative coevolution phenomenon.

Participate in e-business enterprises as a profitable business organization is for the purpose of maximizing profits, therefore we can regard it as bounded rational individuals. Meanwhile, because the repeated decision-making process of individual, not each time all can take a completely rational optimum equilibrium strategies, often require repetitious repeated games can eventually achieving equilibrium state. Therefore, using completely rational basis of game analysis method and balanced concept, also is the establishment in reasoning based on the analysis of one-time or repeated game selection method of bounded rationality to analysis the game problem gambling sides usually not applicable, because cannot reflect the evolution game both parties, nor can learn about game both parties behavior and take the strategy for the dynamic evolution process, analytical is very poor and didn't foresee ability of specific groups of individuals repeated confrontations between the members lack of effective methods. The population characteristics, just meet e-business enterprises integrity transaction mode between evolutionary game models feature.

Through the bounded rational game between enterprises of e-business to analysis the method of trade honestly evolution of process. The result of the game of enterprise each relative to the previous game result from behavioral characteristics and optimizing degrees in the dynamic changing process are analyzed, and through the dynamic equation are described herein. Here is the change process refers to the optimal strategy to change, because the reality of the need for a lot of repetitions, enterprise by chance in game of single game won't affect overall final defeat the equilibrium, and can ultimately realize the whole group strategy equilibrium results, the optimal also accord with the actual development in e-business model in an ideal target state.

III. THE MODELING AND ANALYSIS

A. Bounded Rationality and integrity of e-business transactions between the Game

Treat all the enterprise implementing e-business models as a learning-oriented large groups and all the enterprises in the group repeat transactions at random, mainly taking two strategies of good faith and bad faith in the trading process .Assume the proportion of the enterprises of good faith is x ,while that of the enterprises of bad faith is $(1-x)$ .If both parts of trade are integrity enterprise, then the benefit is A .If trade takes place between integrity enterprises and un-integrity enterprises at random, then the benefit of the former is b and that of the latter is r . The transaction cannot be completed between un-integrity enterprises, that is , the benefit of both parts is 0。The benefit matrix is shown in Table 1.

TABLE I. GAME MATRIX OF INTEGRITY TRADING BETWEEN ENTERPRISES

	Integrity	Un-integrity
Integrity	(a,a)	$(-b,r)$
Un-integrity	$(r,-b)$	$(0,0)$

In the absence of the premise given the circumstances, to benefit from the specific matrix which represents the Nash equilibrium is not clear.

However, it is not hard to see that the expected return of two kinds of business groups and the average expected return are

$$E_A = ax + (-b)(1-x) = (a+b)x - b \quad (1)$$

$$E_B = rx + 0(1-x) = rx \quad (2)$$

$$\overline{E} = xE_A + (1-x)E_B \quad (3)$$

In accordance with the idea of dynamic evolution and copy, during the transaction, the enterprises whose strategy brings lower benefit will gradually transform their strategy to that cause higher benefit. Thus, the proportion of enterprises of different strategies will also change, and variation velocity of proportion of some certain strategy is proportional to its proportion and to the benefit increment compared with the average benefit. Therefore, the change velocity of credible enterprises can be expressed as

$$\frac{dx}{dt} = x(E_A - \overline{E}) = x[E_A - xE_A - (1-x)E_B] = x(1-x)(E_A - E_B)$$
$$= x(1-x)[x(a+b-r)-b] \quad (4)$$

Eqn. (4) can be simply denoted by $\frac{dx}{dt} = F(x)$ (5)

We call equation (4) the dynamic replication equations. According to the nature of differential equations, we can solve the fixed points or stationary points when $F(x)=0$, Which means x keeps stable.

For the above equation, there are three fixed points. They are $x^*=0$、$x^*=1$ and $x^*=\frac{b}{a+b-r}$, respectively. As the third point may be one of the first two points, there are probably only two. The first two stable points mean that the two enterprises having different proportions of strategy tend to adopt the same strategy, namely all of good faith or all of bad faith. The last point means that some enterprises will occupy a certain percentage of different strategy.

The two points correspond to pure strategic equilibrium under absolute rational game, while the third point corresponds to mixed strategic equilibrium.

By take credit strategy that the optimal stable point x^*, Take no integrity strategy for the stability of the $1-x^*=1$、$1-x^*=0$ and $1-x^*=\frac{a-r}{a+b-r}$ points respectively. According to the actual situation in the market analysis of the third point knowable,

When $a+\varepsilon\theta-r=0$,namely $a+\varepsilon\theta=r$,are $1-x^*=0$,Now take the un-integrity strategy group for 0, the proportion of businesses reached an ideal state of integrity. Its practical significance for, when the enterprise take dishonest strategy, the income gained by subtracting the government just equal to the income when they both take the integrity strategy, such integrity strategy is the strategic equilibrium point. But in practice, not always maintain in the balance, and the government's supervision and not be completely over un-integrity enterprise, the excess profit will inevitably have enterprise adventure take dishonest strategies to break the balance, that stability, and tends to move the strategic equilibrium point $1-x^*=\frac{a-r}{a+b-r}$. Similarly, when b tend to 0 , there are $1-x^*=1$,Now take no integrity market strategy for all proportion of business enterprise, but it is do not accord with the actual situation, so market must to the stable point $1-x^*=\frac{a-r}{a+b-r}$ moving, and stability in here. Thus, e-business enterprises in the evolutionary game stability point depend on all locations of initial state .

B. Integrity trading Strategy of the Government Supervision

In the actual market environment, the government is in constant increase of the e-business market transaction efficiency optimization and dishonest behaviors, therefore the supervision of government regulation factors, introduced e-business enterprises and government game analysis.

Assuming trading both sides is pursuing benefit maximization rational individuals, the cost of government implemented the supervision for C, the punishment of enterprise for εC, ε for punishment coefficient, enterprises to choose the earnings of a credit transactions, right now the government to increase social welfare for A; when enterprise chooses not to credit transactions can obtain excess profit for b, now earnings for $a+b$, and causes losses to society $-L$.

TABLE II. GOVERNMENT REGULATION GAME MATRIX

	Perfect regulation	Not Perfect Regulation
Integrity	(a, A)	(a, A)
Un-integrity	$(a+b-\varepsilon c, \varepsilon c - L)$	$(a+b, -L)$

For e-commerce companies, in perfecting government supervision under the condition of credit system, electronic business enterprise whether still carries with it the risk of suffering punishment choose not to honest behavior, this should see trading party won by cheating the excess profits b, whether on more than punishment loss εC .When $b > \varepsilon C$, enterprise get the the excess profits gained for cheating more than the punishment expectations. Therefore, the enterprise in order to pursue the more big profits in trading, still can choose dishonesty, even so cheat behavior may be regulators found and suffered punishment. So, Only when the $b < \varepsilon C$, that is the punishment more than the excess profits gained for cheating expectations, enterprise will abandon dishonest behavior in trading process.

Therefore, if the government hopes that by implementing integrity regulatory means to perfect the market credit system construction, while in considering optimization penalty costs at the same time must improve the hope value, punishment can be effectively reduce dishonest behavior.

IV. SUMMARY

In e-business transactions mode, integrity plays a very important role. Especially in informationization highly developed modern market economy, the good faith principle is improving transactional efficiency, one of the most basic principles, and good sincerity is the key of e-business mode existence and rapid development.

Through the game analysis of above, we can see enterprise as a bounded rational individuals through constant repetition of trading to optimize its benefits, achieve the optimal stable throughout the market point. Although through to government regulation can optimize the market transactions and improve but considering the cost and market regulation, it is impossible to achieve the ideal honesty behavior of the all and stable there. Therefore, the government should focus on establishing and gradually improve integrity system, leading the market and enterprises, and the conception of good faith can be spent and relatively less regulation. Optimize the Transaction efficiency by the market to improve the enterprise cooperation benefits, achieve the optimal market overall interests.

REFERENCES

[1] Zhao Hong-xia, Yang Jiao-ping, "Moral Hazard and Reputation Model of B2C Electronic Commerce Intermediary," Journal of University of Electronic Science and Technology of China, Vol. 28, pp. 83-86, November 2009.

[2] Williamson S, Wright R. "Barter and monetary exchange under private information," The American Economic Review, Vol. 39, pp.149-172, March 1994.

[3] Chen Yun, Wang Huan-chen, Shen Hui-zhang, "Study on the Price Competition between E-commerce Retailer and Conventional Retailer," Systems Engineering-Theory & Practice, Vol.26, pp. 35-41, January 2006.

[4] Yu Zhong-hua, "The Game of Credit Between Buyer and Seller in Electronic Business," Commercial Research, Vol. 339, pp. 67-70, July 2006.

[5] Rajiv Sethi. Somanathan E. "Understanding reciprocity," Journal of Economic Behavior & Organization, Vol. 50, pp. 1-27, 2003.

[6] Anderhub, V. Engeilmann, D., Guth, W, "An experimental study of the repeated trust game with incomplete in formation," Journal of Economic Behavior & Organization, Vol. 48, pp. 197- 216, 2002.

[7] Gth, W. "Incomplete information about reciprocal incentives: an evolutionary approach to explaining cooperative behavior," International Journal of Game Theory, Vol. 24, pp. 323-344, 1995.

[8] Pierre Picard, "Auditing claims in the insurance market with fraud: The credibility issue," Journal of Public Economics, Vol. 63, pp. 27-56, 1996.

[9] Petit ML, Tolwinski B, "R&D cooperation or competition," European Economic Review, Vol. 43, pp. 185-208, 1998.

[10] Xie Shi-yu, "Evolutionary Game Theory Under Bounded Rationality," Journal of Shanghai University of Finance and Economics, Vol. 3, pp. 3-9, 2001.

Design and implementation of a demand and supply system about sponsorship between enterprise and university based on S2SH

HUANG Run-peng, ZUO Wen-ming, JIANG Zhen-hao, PAN Ming, ZHENG Ling-xiao
School of Economics and Commerce,
South China University of Technology,
Guangzhou, P.R.China
hrunpeng@gmail.com

***Abstract*—By integrating Struts2, Spring and Hibernate lightweight frame technology, using ExtJS technology to build rich client applications, setting up HTTP server of Apache and Tomcat load balance, a demand and supply system about sponsorship based on B/S structure between enterprise and university is designed and implemented. Its functional requirements, part of core codes and key technology are provided. As an electronic business platform, this system provides opportunities for online communication and cooperation for univsersities and enterprises. And it can be applied to solve the problem of effectively integrating student associations in universities and enterprise resources.**

Keywords-Struts2; Spring; Hibernate; RIA; load balance

I. INTRODUCTION

"Enterprises - universities" demand and supply system about sponsorship plays a role of e-business platform providing communication and cooperation opportunities for student associations in universities and enterprises. With it, the problem that it's difficulty for student associations to look for sponsors and enterprises to implement effective sponsorship and publicity program can be solved. In the process of solving the problem of sponsorship supply and demand, universities, student associations and enterprises can get further promotion, communications and cooperation.

In this platform, student associations and enterprises are able to browse the other's information. They can determine their best partners by auction or autonomous selection. Besides it, serving as an intermediary, this platform will establish a norm in supervision, feedback and assessment of bilateral credit, characteristic and capacity[1-2]. Meanwhile, the platform will also provide campus activities planning, publicity program of enterprises and its brand and other services, which can improve effectiveness of cooperation between student associations in university and enterprises.

II. REQUIREMENTS ANALYSIS

Questionnaires were issued to collect relevant data and then functions of "enterprises - universities" demand and supply system about sponsorship are analyzed and discribed as Figure 1.

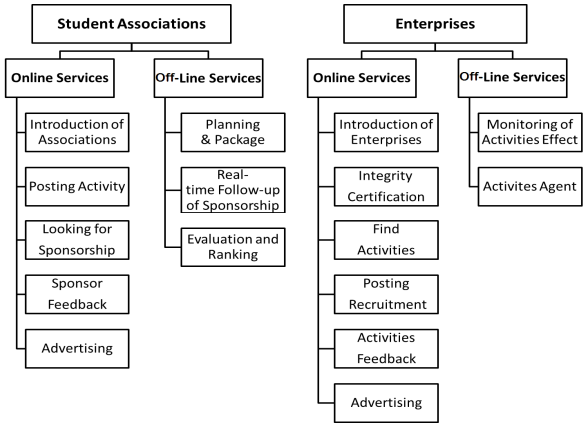


Figure 1. Architecture of Platform Functions.

A. Management System of Sponsorship Projects

Information of activities to be held by student associations and corporate tender projects can be displayed on the platform. Students can manage their activities and companies can manage their sponsorship. At any time they are able to add content for projects on-shelf, update projects information, and cancel or recover display. Correlation among projects is established through the system and then automatic referral function can be implemented. Visitor comments and related media introduction about the project can be added on project details introduction page of. The platform will also log entries clicks of each project, in order to build projects list and classification of the total popular ranking. Each project will have value-added services tips, suggesting that customers can enjoy these services[3].

B. User System

Regular members as free members are able to apply for presentation of students' activities or enterprises tenders, which may be displayed on the platform after audit by administrator. Members have permission to manage their own projects. There are two kinds of members which are ordinary members and certified members. Vetted and approved by administrator, ordinary members can become certified members. Member votes and complaints were recorded. Administrators can forbid members complained to

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.36

log on and publish their information. Each member's credit, characteristic and capacity assessment will also be recorded by the platform in order to establish a sound monitoring system. Platform also provides a separate page for each member, which is a platform to show corporate recruitment information, corporate tender information, project information from student associations and their advocacy information.

Paid members can enjoy all rights of regular members and privileges, including top ranking priority service, integrity and authentication services, professional buyer list, unlimited products releases, pictures bulk upload, the PPC services, home advertising, home page Gold Booth, discounts and so on. Meanwhile, the platform will periodically send e-mail to paid members, and push the latest project information meeting needs from users by using data mining techniques.

III. PLATFORM DESIGN

By integrating Struts2, Spring, Hibernate and ExtJS and building Apache and Tomcat load balancing, the "enterprises - universities" demand and supply system about sponsorship are comprehensively planned. 5 levels including UI layer, business layer, persistence layer, domain model layer and web server will be described as follows.

A. JavaEE Four-Level Architecture

Most of web-based applications based on JavaEE can be divided into four layers, including UI layer, business layer, persistence layer, and domain model layer (Figure 2)[4]. Among them, the UI layer is responsible for responding to user's request, returning a response result. Business layer is responsible for processing for business logic of platform and providing functions' interface to UI layer. Persistence layer is responsible for data persistence and data tables will be mapped into Java objects. Domain model layer encapsulate data access and is responsible for data transfer between different layers. Reasonable and appropriate design of the hierarchical model can improve system flexibility, reusability and maintainability, and the purpose of achieving decoupling and reducing system complexity can be achieved.

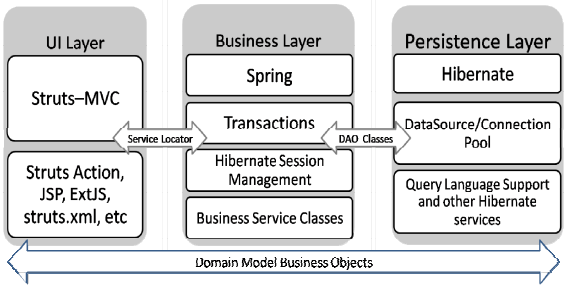


Figure 2. Figure 1 JavaEE Architecture.

B. web Server

web server is an application running on the server side, responsible for responding to client requests. When the web server receives an HTTP request, based on contents of the request, it will return an HTML page or image, or entrusted the request to other applications such as JSP script, ASP script, and then generate HTML pages and responses.

There are 2 schemes to build web server. One is using a single web server respond to static pages and dynamic page requests at the same time. The other is using multiple web servers to respond to different requests, implementing load balance. With business increasing, traffic and data volume growing, a single web server cannot bear such a heavy load. Therefore, building a web server load balance can enhance network capacity and increase network scalability and security.

The platform will use Apache and Tomcat to implement load balance. Due to Apache's stability, rapid response capabilities and ability to respond to high concurrent requests, Apache will serve as the front-end web server, which is responsible for responding to user's request and returning static distribution of resources such as HTML pages or images. Tomcat works as the web back-end server, JSP and Servlet container, which is responsible for compiling JSP and Servlet and generating dynamic pages.

IV. IMPLEMENTATION OF PLATFORM

Based on before-mentioned platform design, the overall platform architecture is proposed as shown in Figure 3. To elaborate implementation of the platform, implemention of project exbition module which is as an example is described as follows.

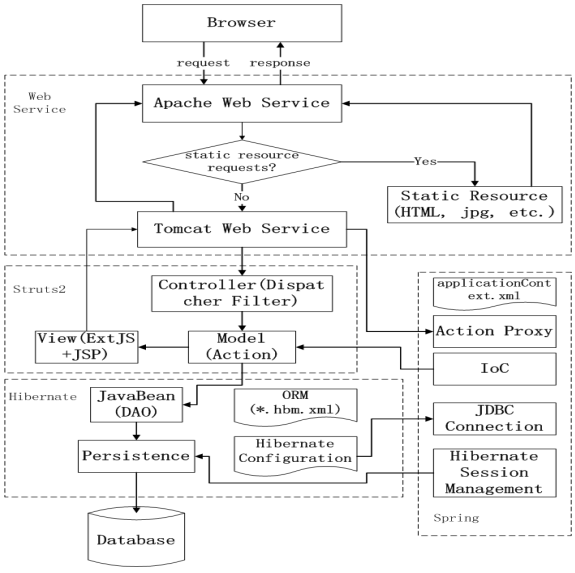


Figure 3. Platform Architecture.

A. Implementation of UI Layer

When a user makes a request to access to a sponsor project, Struts2 interceptor is to intercept user requests, according to struts.xml configuration information, instantiated the viewProject.action[5]. Struts.xml configuration information is as follows:

```
<action name="viewProject" class="action.ViewProject">
  <result> /viewProject.jsp</result>
```

```
</action>
```

ViewProject.action is responsible for handling user requests. Based on user's request, it will call business layer components to obtain sponsorship information through Spring carrying out IoC. Then according to the results, the response jumps to viewProject.action specified page.

Method of combining technology and ExtJS.JSP for page implementation is used. JSP generate the page display by using Struts2's TagLib library to obtain sponsorship data. Meanwhile, the introduction of ExtJS file (resources / css / ext-all.css) and the ExtJS library files (adapter / ext / ext-base.js and ext-all.js) can be applied to build rich client applications[6].

B. Implementation of Business Layer

ProjectService class as sponsorship model in viewProject.action which is responsible for show of sponsor projects is built based on business modeling. It encapsulates sponsorship and operation rights management. When business layer receives a call to business logic components because of performance of the request, according to configuration in applicationContext.xml, Spring will instantiate ProjectService and then inject into viewProject.action, achieving decoupling purposes between UI layer and business layer. ApplicationContext.xml configuration information is as follows.

```
<bean id="viewProject" class=" action.ViewProject">
  <property name="projectService">
      <ref bean="projectService" />
  </property>
</bean>
<bean id="projectService" class="dao.ProjectServiceImpl">
  <property name="projectDao">
    <ref bean="projectDao" />
  </property>
</bean>
<bean id="projectDao" class="dao.ProjectDaoImpl">
  <property name="sessionFactory">
    <ref bean="sessionFactory" />
  </property>
</bean>
```

Based on configuration information in applicationContext.xml, it is shown that Spring is also responsible for instantiating data access classes ProjectDao, and injecting into ProjectSerivce components. ProjectService executes operation of sponsorship project by calling ProjectDao. Implementation of ProjectDao (i.e. the domain model layer) is explained as follows.

C. Implementation of Domain Model Layer

ProjectDao encapsulates details of accessing data. In S2SH Framework, Spring uses sessionFactory to bear the transaction management of Hibernate persistence. Through inheriting data access helper class - HibernateDaoSupport, ProjectDao gets injection of sessionFactory instance, thus obtains the persistent entities of sponsorship - operation with project. ProjectDao provides a series of get and set methods. By calling these get and set methods business layer achieve operations for Project.

D. Implementation of Persistence Layer

Based on preparation of the corresponding Java persistent classes of database tables and the corresponding mapping file, relational database tables can be mapped into Java persistent classes using Hibernate, which can operate database tables with the idea of object-oriented. The mapping file of sponsorship information is as follows.

```
<hibernate-mapping>
  <class name="model.Project" table="project">
      <id name="id" type="java.lang.Integer">
        <column name="id" />
        <generator class="native" />
       </id>
      <property name="message" type="java.lang.String">
          <column name="message" />
      </property>
  </class>
</hibernate-mapping>
```

E. Implementation of web Server Load Balancing

The following describes realization of web server load balance using Linux operating system as an example[7]. This platform uses a program of mod_jk to integrate Apache and Tomcat[8].

Step 1. Install Apache and Tomcat.

Step 2. Download mod_jk.so plug-ins. Compiled, copied to Apache's module directory.

Step 3. Create a new file named workers.properties in conf directory of apache, which is as follows:

```
workers.tomcat_home=/usr/share/tomcat6/
workers.java_home=/usr/lib/jvm/
worker.list=ajp13
worker.ajp13.port=8009
worker.ajp13.host=localhost
worker.ajp13.type=ajp13
```

Step 4. Create a new file named mod_jk.conf in conf directory of apache, which is as follows:

```
JkWorkersFile "/etc/httpd/conf/workers.properties"
JkLogFile "/etc/httpd/logs/mod_jk.log"
JkMount /servlet/* ajp13
JkMount /*.jsp ajp13
JkMount /*.action ajp13
```

Step 5. Configure httpd.conf file. In the end of httpd.conf, import mod_jk.so and configuration files:

```
LoadModule jk_module modules/mod_jk.so
Include /etc/httpd/conf/mod_jk.conf
```

Step 6. Configure Tomcat configuration file - server.xml:

```
<Context path="" docBase="/var/www/html" debug="0"
reloadable="true" crossContext="true" />
```

F. Implementation Results

Through realization of integration of Struts2, Spring, Hibernate and ExtJS, load balance of Apache and Tomcat, the platform laid a solid foundation for future development and operation. Figure 4 shows the enterprise and student associations in university are on an online discussion.

Figure 4. Online Discussion.

V. CONCLUSION

Through analyzing the demand and supply situation between student associations in university and enterprises, "enterprises - universities" demand and supply system about sponsorship is proposed in this paper from application layer and server layer. Using an example of realization of demonstrating a sponsor project, realization of solution is explained in detail. Struts2 and ExtJS implement UI layer. Spring implements business layer and decoupling between various levels. Hibernate is responsible for data persistence. Apache is responsible for responding to user requests and distributing requests. Tomcat is a JSP and Servlet container. Solution of this system can provide technology reference to build a similar eletronic business platform.

ACKNOWLEDGMENT

This paper is supported by National Student Innovation Experiment Program (091056140), Ministry of Education, Humanities and Social Science Foundation Project (07JC630045), Guangdong Province Soft Science Research Project (2007B060401078), South China University of Technology, The Fundamental Research Funds for the Central Universitis (2009SZ0029).

REFERENCES

[1] Shu Zi. When the Expo case of the Internet. Internet World, 2010 (4) :18-21.

[2] Shu Wenqiong. PC makers to push the Internet company behind the free sponsorship. Communication World, 2005 (41) :11-11.

[3] Chen Peng. CBA League official site communication and project risk management. Beijing University of Posts and Telecommunications, 2010.

[4] Jian-Wen Wang, Han Li, Zhang Junming. Oilfield based on multi-tier architecture design and implementation of material management system. Computer Engineering and Design, 2011,32 (1) :162-165, 292.

[5] Zuo Wenming, Wang Qingqiang, Wu Ying Liang, Zhou Pu Xiong. Struts-based land management information system . Computer Engineering and Design, 2008,29 (19) :4972 -4974, 4978.

[6] Xiao Jie, Chen Xiang, Hehai Jiang, Shao-Gang Cui. AJAX and Struts-based web application design and implementation. Computer Engineering and Design, 2009 (8) :1934 -1937.

[7] Du Ding, GAN Ren-chu, Lei Yusheng. Linux environment, web-based information system construction. Computer Applications, 2004,21 (6) :53-54, 89.

[8] Hao Haifeng, Xu Haifeng, Ye Jun. JSPWEB server based on Windows platform, set up. Computer and Digital Engineering, 2007,35 (8) :68-70.

The optimal design model of logistics network for scrap copper recycling

Pin Peng
School of Economics and Management
Jiangxi University of Science and Technology
Ganzhou, China
pengpin@126.com

Yarui Zhang
School of Economics and Management
Jiangxi University of Science and Technology
Ganzhou, China
zhangmisue@hotmail.com

Gui Pang
School of Economics and Management
Jiangxi University of Science and Technology
Ganzhou, China
IE_panggui@163.com

***Abstract*—Based on the analysis of the scrap copper recycling, the paper builds a three-stage logistics network including the scrap copper recovery nodes, reprocessing center and demanding nodes. At the same time, the paper constructs the decision model of recycle and distribution in reverse logistics, and the maneuverability of the model is verified by an example.**

***Key words*—scrap copper; recycling; reverse logistics; network design**

I. INTRODUCTION

The reuse of the waste and the energy is becoming important all over the world as the attention to the environment protection and resources recovery. The resource recycling industry develops rapidly with the support of the related policy. The resources are the fundamental conditions for sustainable development. However, the contradictions between the lack of copper resources and the growing consumption of copper alloy are more prominent [1], so it is an important way to remedy insufficient copper resources by recycling scrap copper. The research on the reverse logistics of the scrap copper and the optimization recycling process is not only to increase recovery and multipurpose utilization and reduces production cost, but also reduces resource waste and realizes the sustainable development of copper industries.

The reverse logistics has been analyzed and studied extensively and deeply by many scholars. Spengler T. [2] concentrated the recycling procedure of waste electrical appliance into a MILP model and made the dividing decision of routine product's treatment. Liqing Da [3] reviewed the reverse logistics from its problems and methods. Zhiqiang Lu [4] aimed at minimizing the whole cost, discussed the integrated logistics network of remanufacturing, and built a mixed integer linear programming for single facility location and three facility location. Using the mixed integer linear programming, Zujun Ma [5] built up a design model of logistics network for single and capacity restricted products with the view of minimizing investment cost of facilities and the operating cost.

Based on the station and feature of the scrap copper, this paper established a three-gradation construction for scrap copper recycling. The construction involved recovery node, reprocess center and demanding node. The article built a capacity restricted linear programming model for the scrap copper recycling. And the distributive quantity of the material in all the paths of the scrap copper recycling network was adopted by LINGO 8.0. At last, an example would be used to verify the model.

II. PROBLEM DESCRIPTION

A. Problem Definition

The reverse logistics is generally used for description of recycling, treatment of waste disposal and hazardous materials management [1], while from a much wider perspective， it can also be used for resources conservation, recycling, substitution and reuse of materials. With reusing, remanufacturing, repairing, recycling, returning product and treatment of waste disposal, the reverse logistics is a set of activities which has also emerged with collection, transportation, inventory management and distribution together. The design of the logistics network is a whole construction of the channel starting from origins of waste to consumption，which includes types, quantity and situation of the facilities and modes of transportation [5].

The reverse logistics of scrap copper has Recovery Node, Reprocess Center and Demanding Node, which is responsible for recycling, disassembling, selecting, classifying, utilizing and distribution. The recovery node of the scrap copper which aims to purchase the scrap copper is built in the consumption area. The reprocess center is responsible for detection, classification and regeneration. The demanding node is made up of the recycled materials markets, the secondary markets and new product markets. For these a three-stage logistics network of the scrap copper recycling is established. (See Figure 1)

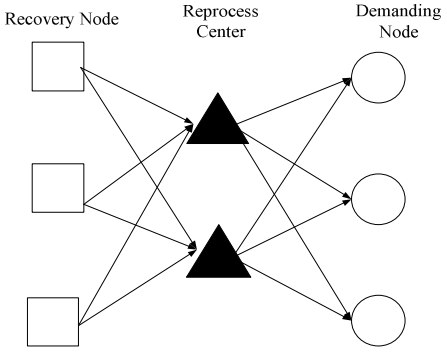


Figure 1. The reverse network of scrap copper recycling

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.92

B. Fundamental assumption

1) The quantity and location of the recovery nodes，the reprocess centers and the demanding nodes are known.

2) The recycling capabilities of the recovery nodes and the handling ability of the reprocess centers are known.

3) The distance and the transportation cost per unit between the recovery nodes and the demanding nodes are known.

4) The $1^{\#}$, $2^{\#}$ and $3^{\#}$ scrap copper are recycled by recovery nodes, and the recovery price is different in the recovery nodes.

5) After reprocessed and smelted, the scrap copper turns into electrolytic copper and copper alloy. The $1^{\#}$ scrap copper are all used to product the copper alloy, and the $2^{\#}$ and $3^{\#}$scrap copper are all used to product the electrolytic copper.

6) The quantity and the price of the reprocessed product in every demanding node are known.

7) Without considering the spillage of material in the recycling and the processing, the conversion rate of scrap copper and the treatment cost are known.

8) The government would give subsidy for the scrap copper recycling, but just for the electrolytic copper and the copper alloy.

9) Without considering the inventory of product in the reprocess center, the electrolytic copper and the copper alloy are all sent into the demanding nodes.

III. THE CONSTRUCTION OF THE MODEL AND PROBLEM SOLVING

Based on the considering of factors including the price, the subsidy of the reprocessed product, the processing price, the total transportation cost and the price of scrap copper, this paper designed a optimal model of logistics network for scrap copper recycling and then it can decide the handling capacity and freight volume between every node of the reverse logistics network.

A. Parameter and variable

1) Symbol definition

I：the types of the scrap copper, $i=1,2,\cdots,I$，and $I=3$，I stands for the $1^{\#}$, $2^{\#}$ and $3^{\#}$ scrap copper .

J：the types of the reprocessing product, $j=1,2,\cdots,J$，and $J=2$，J stands for the electrolytic copper and the copper alloy.

R：The quantity of recovery nodes, $r=1,2,\cdots,R$.

T：The quantity of reprocess centers, $t=1,2,\cdots,T$.

K：The quantity of demanding nodes, $k=1,2,\cdots,K$.

2) Decision variable

x_{irt}：The quantity of the $1^{\#}$, $2^{\#}$ and $3^{\#}$ scrap copper sent from the r recovery node to the t reprocess center.

y_{jtk}：The quantity of the electrolytic copper and copper alloy sent from the t reprocess center to the k demanding node.

3) Parameter

p_{ir}：The price of scrap copper in the r recovery node.

C_{ir}：The maximum of scrap copper in the r recovery node.

d_{rt}：The distance between the r recovery node and the t reprocess center.

c_{rt}：The transportation cost of scrap copper sent from the r recovery node to the t reprocess center.

q_{ij}：The conversion rate between the i scrap copper and the j processed product.

$H_{it}^{\min}$、$H_{\mathrm{it}}^{\max}$：The maximum and the minimum treatment capacity of the scrap copper in the t reprocess center.

D_{jk}：The quantity demanded of the j processed product in the k demanding node.

$d_{tk}^{'}$：The distance between the t reprocess center and the k demanding nodes.

$c_{tk}^{'}$：The transportation cost of the reprocessing product sent from the t reprocess center to the k demanding node.

e_{jk}：The price subsidy for the price of the j processed product in the k demanding node.

f_{jt}：The conversion cost of the reprocessed product

$p_{jk}^{'}$：The price of the j processed product (the electrolytic copper and the copper alloy) in the k demanding node.

B. Modeling

According to the above parameter specification, a Linear Programming model is built.

Max (1)

$$\sum_{j=1}^{2}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(p_{jk}^{'}+e_{jk}\right)y_{jtk}-\sum_{i=1}^{3}\sum_{r=1}^{R}\sum_{t=1}^{T}\left(c_{rt}d_{rt}+p_{ir}\right)x_{irt}-\sum_{j=1}^{2}\sum_{t=1}^{T}\sum_{k=1}^{K}\left(f_{jt}+c_{tk}^{'}d_{tk}^{'}\right)y_{jtk}$$

s.t. $\sum_{t=1}^{T}x_{irt}\le C_{ir}\qquad i=1,2,\cdots,I;r=1,2,\cdots,R$ (2)

$$\sum_{r=1}^{R}\sum_{t=1}^{T}q_{1,1}x_{1rt}=\sum_{t=1}^{T}\sum_{k=1}^{K}y_{1tk} \quad (3)$$

$$\sum_{i=2}^{3}\sum_{r=1}^{R}\sum_{t=1}^{T}q_{i2}x_{irt}=\sum_{t=1}^{T}\sum_{k=1}^{K}y_{2tk} \quad (4)$$

$$H_{t}^{\min}\le\sum_{r=1}^{R}\sum_{i=1}^{I}x_{irt}\le H_{t}^{\max}\qquad t=1,2,\cdots,T \quad (5)$$

$$\sum_{t=1}^{T}y_{jtk}=D_{jk}\qquad k=1,2,\cdots,K;j=1,2,\cdots,J \quad (6)$$

$$x_{irt}\ge 0;y_{jtk}\ge 0 \quad (7)$$

Target function (1) stands for maximizing the profit of the whole logistics network. Constraint equation (2) stands for a constraint on the quantity of the scrap copper in the recovery nodes. Constraint equation (3) stands for a material balance constraint on the conversion rate of the

copper alloy. Constraint equation (4) stands for a material balance constraint on the conversion rate of the electrolytic copper. Constraint equation (5) stands for a constraint on handling ability of the reprocess center. Constraint equation (6) stands for a constraint on logistics. Constraint equation (7) stands for nonnegative number.

C. *Model solution*

The above optimal design model of logistics network for scrap copper recycling is a mixed integer linear programming. The optimal design of the logistics network will be a NP-hard problem, if the logistics network includes many variables and constraint equations. The NP-hard problem is solved by Heuristic Approach or Swarm Intelligence Algorithm.

However, the software LINGO 8.0 can be used to solve the model, if the quantity of the nodes in the scrap copper reverse logistics network is relatively small,. The handling capacity and the freight volume between every node of the reverse logistics network are obtained by applying LINGO 8.0.

IV. EXAMPLE

A Scrap Non-ferrous Metals Recycling Logistics Company A takes responsibility to recycling the scrap non-ferrous metals, then after reprocessing, the useful product will be sent into demanding nodes. Under the guidance of government's Policy, this year A company opens two yard for recycling scrap copper($R1, R2$), two reprocess centers ($T1, T2$) and three demanding nodes($K1, K2, K3$). A company plans to recycle $1^{\#}$, $2^{\#}$ and $3^{\#}$ scrap copper($I1, I2, I3$), reprocess these scrap copper into the electrolytic copper and copper alloy($J1, J2$), and sent them into the three different demanding nodes.

Suppose the conversion cost of the copper alloy and the electrolytic copper (f_{jt}) is 1400 yuan/ton and 1800 yuan/ton respectively. The transportation cost per unit of the product ($c_{rt}; c_{tk}$) is 1.2 yuan/ ton · kilometer. The price subsidy per unit of the copper alloy and the electrolytic copper (e_{jk}) is 570 yuan/ton and 700yuan/ton. The production capacity of the scrap copper in the reprocess center $T1$ is [3600, 4000], and the production capacity of the scrap copper in the reprocess center $T2$ is [3400, 3800]. The quarter is calculated to units. Other data refers to the following tables.

Table 1. The price of the processed product in the demanding nodes p'_{jk} (yuan/ton)

Demanding node	Processed product	
	$J1$	$J2$
$K1$	68740	69600
$K2$	68430	69300
$K3$	68950	69795

Table 2. The distance between the reprocess centers and the recovery nodes d_{rt} (kilometer)

Reprocess center	Recovery node	
	$R1$	$R2$
$T1$	108	120
$T2$	107	136

Table 3. The distance between the recovery nodes and the demanding nodes d'_{tk} (kilometer)

Reprocess center	Demanding node		
	$K1$	$K2$	$K3$
$T1$	108	120	135
$T2$	116	129	107

Table 4. The price of the scrap copper in the recovery nodes p_{ir} (yuan/ton)

Recovery node	The scrap copper		
	$I1$	$I2$	$I3$
$R1$	58800	54860	41200
$R2$	59000	54900	40800

Table 5. The maximum of the scrap copper in the recovery nodes C_{ir} (ton)

Recovery node	The scrap copper		
	$I1$	$I2$	$I3$
$R1$	1700	1180	900
$R2$	1800	960	1000

Table 6. The conversion rate between the scrap copper and the processed product q_{ij}

Processed product	The scrap copper		
	$I1$	$I2$	$I3$
$J1$	96%	—	—
$J2$	—	92%	86%

Table 7. The quantity demanded of the processed product in the demanding nodes D_{jk} (ton)

The demanding nodes	Processed product	
	$J1$	$J2$
$K1$	1200	900
$K2$	1000	1100
$K3$	1100	1200

Lingo8.0 is used to solve this design model of logistics network for scrap copper recycling. The profit of the company is 67548320 RMB. The optimal value (see table 8 and 9) shows that company A recycled 31 tons of the $1^{\#}$ scrap copper，33 tons of the $2^{\#}$ scrap copper and 26 tons of the $3^{\#}$ scrap copper respectively in this quarter. The volume of the copper alloy sent to the demanding nodes

($K1$, $K2$, $K3$) are 1200 tons, 1000 tons and 1100 tons respectively, while the volume of the electrolytic copper sent to demanding nodes is 900 tons , 1100 tons and 1200 tons respectively. According to the above result, the production capacity of the scrap copper in the reprocess center $T1$ is 3600 tons, the production capacity rate is 90%; the production capacity of the scrap copper in the reprocess center is 3439.67 tons, the production capacity rate is 90.5%.

V. CONCLUSION

With a view to the lack of the copper resources in the international market and the lower investment, lower energy consumption, lower cost and no secondary pollution in the scrap copper recycling, the comprehensive utilization recovery of scrap copper can realize the durative development of resources and copper industry. Based on the analysis of the scrap copper' state and features, this paper did a research on the design of the logistics network for scrap copper recycling, and constructed an optimal design model which is characterized by various kinds , multi-section and ability limit on logistics network for the scrap copper recycling. Finally, the maneuverability of the model is verified by an example.

REFERENCES

[1] Yarui Zhang, Pin Peng, "The research on the cureent situation and suggestion of Chinese copper scrap recycling", Logistics Sci-Tech, May,2011,pp.8-10

[2] Spengler T, Ploog M, Schmter M, "Integrated planning of acquisition disassembly and bulk recycling: a case study on electronic scrap recovery," OR Spectrum, 2003, pp. 413-442.

[3] Liqing Da, Zuping Huang, Qin Zhang, "Current and Future Studies on Structure of the logistics System: A Review," Chinese Journal of ManarernenL Science, Vol. 12, No. 1, Feb.2004, pp. 131-138.

[4] Zhiqiang Lu, Nathalie Bostel, "A facility location model for logistics including reverse flows: The case of remanufacturing activities," Computers and Operations Research, Feb. 2007, pp.299-323.

[5] Zujun Ma, Yin Dai, "Optimization Model for Reverse Logistics Network Design for Product Recovery," Journal of Industrial Engineering / Engineering Management, Vol. l9, No. 4, 2005, pp. 114-117.

Table 8. The share of recycling product sent from recovery nodes to reprocess centers (ton)

x_{irt}	x_{111}	x_{112}	x_{121}	x_{122}	x_{211}	x_{212}	x_{221}	x_{222}	x_{311}	x_{312}	x_{321}	x_{322}
Value	0	1700	1737.5	0	340.33	839.67	522.17	0	0	900	1000	0

Table 9. The share of processed product sent from reprocess centers to demanding nodes (ton)

y_{jtk}	y_{111}	y_{112}	y_{113}	y_{121}	y_{122}	y_{123}	y_{211}	y_{212}	y_{213}	y_{221}	y_{222}	y_{223}
Value	1200	1000	0	0	0	1100	900	1100	0	0	0	1200

Research on the Price Forecast without Complete Data based on Web Mining

Quanyin Zhu, Sunqun Cao, Jin Ding and Zhengyin Han
Faculty of Computer Engineering
Huaiyin Institute of Technology
Huaian, Jiangsu Province, China
hyitzqy@126.com

***Abstract*—The commodities price evolution is a very interesting data for shopkeepers selling online. By using the Web mining can get more and more data in everywhere such as e-supermarkets and e-commerce. This paper shows a case study for the price extracting of mobile phone selling online. Extracting the commodities price to get the price trend using the uncompleted data is researched. The forecast algorithm is described in detailed. The functions coded by PHP are designed for implantation the price extracting. Experiment demonstrates its performance and proves price evolution is meaningful and useful for the shopkeepers selling online.**

Keywords- price evolution; mobile phone; selling online; Web mining

I. INTRODUCTION

Forecasting is the process to make the statements about events whose actual outcomes (typically) that has not yet been observed. A normal sample data might be used to estimate for some variable of interest at some specified future date. Prediction is a similar way, but need more general term. Both of them might refer to formal statistical methods employing time series, cross-sectional or longitudinal data, or alternatively to less formal judgmental methods[1]. Predicts theory and method can be applied widely distributed in all kinds of areas of natural and social aspects. According to the areas covered, the different research objectives and tasks, forecasting can be classified with different areas of forecasting such as weather forecasting, scientific forecasting, military forecasting, technology forecasting, economic forecasting, and social prediction[2].

As we know, the Web mining can be used to discover the news [3], analyze email communications [4], and require skill sets for computing Jobs [5] and even to latent topics from web sites of terrorists or extremists [6] and so on. When shopkeepers want to know more information about the customer and recommend the new products to them using e-mails, but the most interested data is the commodities price. So, our group researched interesting on the commodities price evolution using Web mining for the shopkeeper's selling online.

II. THE COMMON METHOD

A. *Single errors of predicted value: e_t*

Let's make the actual value is: Y_t, the predicted value is: $\hat{Y}_t$, then the *t* period predicted value is:

$$e_t = Y_t - \hat{Y}_t \qquad (t = 1,2,\ldots,n) \tag{1}$$

$e_t > 0$ is predicted value less than the actual value, $e_t < 0$ is predicted value greater than the actual value, $e_t = 0$ is predicted value is accuracy.

B. *Relative errors of single predicted value $\tilde{e}_t$:*

$$\tilde{e}_t = \frac{e_t}{Y_t} = \frac{Y_t - \hat{Y}_t}{Y_t} \qquad (t = 1,2,\ldots n) \tag{2}$$

Relative errors using percentage to indicated. For example: $\tilde{e}_t$ =1%, means the predicted values on the low side of 1%, or we can say predict accuracy is 99%。

C. *Mean Absolute Errors(MAE)*

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t| = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right| \tag{3}$$

D. *Mean Absolute Percentage Errors(MAPE)*

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\frac{|e_t|}{Y_t} = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|Y_t - \hat{Y}_t\right|}{Y_t} \tag{4}$$

E. *Mean Square Error(MSE)* [7]

$$MSE = \frac{1}{n}\sum_{t=1}^{n}e_t^2 = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2 \tag{5}$$

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.80

III. THE SPECIFIC FORECASTING ALGORITHM

Due to historical data accumulation is very shorter. Generally, in considering the specific forecasting model, you need to be resolved how to deal with seasonal factors.

If a tend of time series effected by seasonal variation, just like the Figure 1, It can be used Tend Ratio Method.

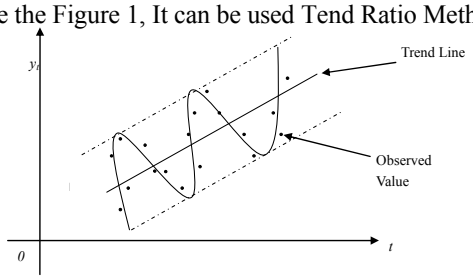


Figure 1. Linear tend seasonal time series

Set up tend line equation

$$T_t = \hat{a} + \hat{b}t \tag{6}$$

According to tend line equation, computing $T_1, T_2, \cdots, T_n$ in each phase.

Reject tend:

$$\widetilde{S}_t = \frac{y_t}{T_t} \qquad (t = 1, 2, \cdots, n) \tag{7}$$

Preliminary estimates seasonal index. compute the average of the common season $\widetilde{S}_t$, remove random disturbance. Set the average value as the seasonal index preliminary estimates:

$$\overline{S}_i = \frac{\widetilde{S}_i + \widetilde{S}_{i+L} + \widetilde{S}_{i+2L} + \cdots + \widetilde{S}_{i+(m-1)L}}{m} \qquad (i = 1,2,\cdots,L) \tag{8}$$

Final estimate the seasonal index. In one phase, the sum of seasonal index value should be equal to L, just:

$$\sum_{i=1}^{L} \overline{S}_i = L \tag{9}$$

But this kind of method gets the seasonal index preliminary estimates value needs adjust. The adjust method is to find out in one phase, each seasonal index preliminary estimates as regulation factor, like:

$$S = \frac{1}{L}\sum_{i=1}^{L} \overline{S}_i \tag{10}$$

Then using each phase seasonal index preliminary estimates $\overline{S}_i$ divide regulation factor S, get the seasonal index value: [2]

$$S_i = \frac{\overline{S}_i}{S} \qquad (i - 1,2,\ldots,L) \tag{11}$$

In order to eliminate seasonal factors, we need to use the mobile phone daily price divided by the seasonal factor to obtain the removal of seasonal value. The prices of mobile phone Web mining data show as Table 1 (Data from 27 March, 2011 to 22 April, 2011).

TABLE I. PRICES OF MOBILE PHONE SALES

Mobile Phone	Price /Days	Price / Days	Price /Days	Price /Days	Seasonal Factor Expression	Seasonal Factor			
						1	*2*	*3*	*4*
Apple:Iphone	5880/24	4999/3			$\alpha = \frac{\overline{Y}_1}{\overline{Y}_2}$ $\overline{Y_1}$:weekly mobile phone average prices; $\overline{Y_2}$ monthly average prices	1.0169	1.0169	1.0169	0.9407
Nokia:E63	1268/20	1218/7				1.0103	1.0103	1.0046	0.9705
Nokia:5233	1062/18	1021/9				1.0130	1.0130	0.9963	0.9739
Sony Ericsson:J20	1399/3	1350/13	1300/5	1240/6		1.0373	1.0214	0.9944	0.9382
Motorola:E11	900/2	899/10	875/15			1.0153	1.0072	0.9879	0.9879
The average value of the seasonal factor						1.0186	1.0138	1.0000	0.9622

TABLE II. THE ERROR ANALYSIS

Mobile Phone	1th WAV	2th WAV	3th WAV	4th WAV	MAV	PE	5th FAV	ERROR
Sumsung S5230c	952.09	922.53	894.14	848.29	904.26	y=-0.01031*x*x+0.062316*x-0.0546	875.17	3.08%
Nokia 5250	983.39	964.63	946.14	976.57	967.68	y=0.025134*x*x-0.0563*x+0.030603	1177.41	17.16%
MOTO XT301	1134.29	1109.14	1066	1066	1093.86	y=0.027948*x*x-0.0444*x+0.049367	1414.86	32.73%

WAG :The weekly each mobile phone price. MAG: The monthly average value of each mobile phone price. PE: The Parabola Equation of the seasonal factor linear difference. FAV: The forecast average value of each mobile phone price. ERROR: The error value of each mobile phone price.

we can assume the equation of the average values is:

$$y = ax^2 + bx + c \tag{12}$$

y: the next average value of seasonal factor.

x: the value of weeks.

In order to simplify the computing, let's x=1; x=2; x=3 into the equation.

We can get the Ternary Once Equations:

$$\begin{cases} y_1 = a + b + c \\ y_2 = 4a + 2b + c \\ y_3 = 9a + 3b + c \end{cases} \quad \text{that is} \quad \begin{cases} a = (y_1 - 2y_2 + y_3)/2 \\ b = (-3y_1 + 4y_2 - 3y_3)/2 \\ c = 3y_1 - 3y_2 + y_3 \end{cases} \tag{13}$$

Because of y_1, y_2, y_3 are known values. So we can get the value of a, b and c . we can express the equations (12) as:

$$y = \frac{(y_1 - 2y_2 + y_3)}{2}x^2 + \frac{(-3y_1 + 4y_2 - 3y_3)}{2}x + (3y_1 - 3y_2 + y_3) \quad (14)$$

Then if $x = 4$, we can get the y=*-0.07692*.

So, we can use y=*-0.07692* to work out the next period seasonal factor:0.88532.

Then each mobile price can use the monthly average prices multiply 0.88532.

Of course, in the actual program, this factor is changed but not static.

Table 2 shows three different kinds of mobile phone forecasting price and the error analysis

From Table 2, we can clearly find out that the error of this model is woeful. This phenomenon is caused by the seasonal factor selected of linear difference analysis methods used parabola which is needed to be improved.

IV. MOVING AVERAGE FORECAST ALGORITHM

Let x_1, x_2, …, x_t ,… is a time series observations, where x_t as t period time series forecast value. Defined $f_{t,1}$ as observation x_t assign the forecast cycle of cycle t+1. The method of the moving average is：

$f_{t,1}$ = the last observation average values = average value of x_t, x_{t-1}, …, x_{t-N+1}. N is appointed parameter[8].

N—The item of moving average(or called step size).

The value N is concern the degree of forecast accuracy. Then, how to select the value of N? Generally, when the trend of time series is stabilize, the value of N is fetched great, else the value of N is fetched lesser. However, this principle is not convenient to use. In the actual forecast, we can take the trial method that means select several different values of N, computing the error value, the error value is short, it means the value of N is coincidence the forecast model [2].

We should attention that the value N get greater, thus the forecast curve will be more smoothly, so it will lost a lot of price information[9].

We extract two kinds of mobile phone to analysis the price trend. The average price of them show as Figure 2 to 3 respectively.

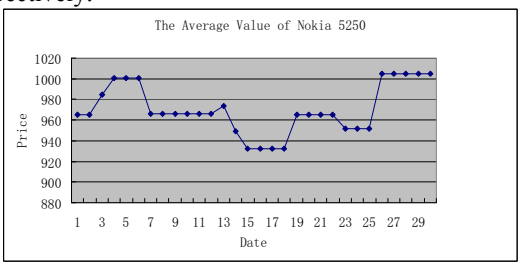


Figure 2. The average price of Nokia 5230

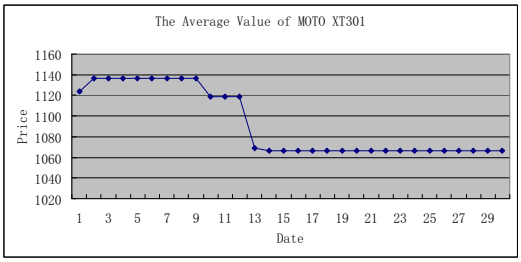


Figure 3. The average price of MOTO XT301

From the Figure 2 and 3, we can find that the price of mobile phone is changing frequently in long period of time, but for a short time, the price is stable correspondingly.

In order to convenient for compare the results, we set N=3 and N=5 to simulation only one month data to determine the price evolution model.

$$\hat{y}_{t+1} = \frac{y_t + y_{t-1} + y_{t-2}}{3} \quad (N=3) \quad (15)$$

$$\hat{y}_{t+1} = \frac{y_t + y_{t-1} + y_{t-2} + y_{t-3} + y_{t-4}}{5} \quad (N=5) \quad (16)$$

Using the formula (15) and (16), we can get the different step size N effect the forecast results which are show as Figure 4 and 5 respectively.

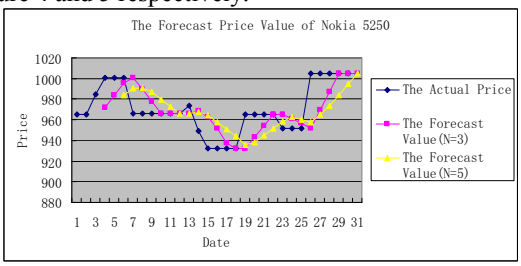


Figure 4. Different step size N effect the forecast value for Nokia 5230

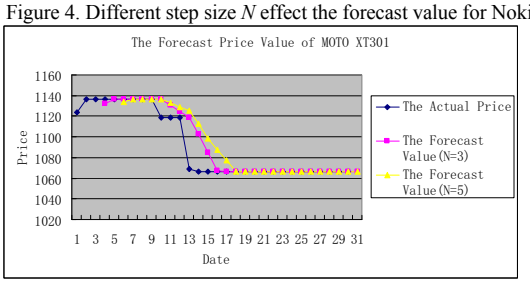


Figure 5. Different step size N effect the forecast value for MOTO XT301

From the Figure 4 and 5, we can find that N=5 is volatility obvious than the N=3 on the moving forecast, it means the forecast of N=5 is slow response to the forecast value.

V. CONCLUSION AND FUTURE WORK

To the mobile phone price forecast, we select N=3 and N=5 as the step size. We use the MSE method to calculate it, and get the results show as Table 3:

For N=3：

$$MSE = \frac{1}{28}\sum_{t=4}^{31}(y_t - \hat{y}_t)^2 \quad (17)$$

For N=5：

$$MSE = \frac{1}{26}\sum_{t=6}^{31}(y_t - \hat{y}_t)^2 \quad (18)$$

TABLE III. THE MSE OF NOKIA 5230 AND MOTO XT301

Mobile Phone	MSE	
	N=3	*N=5*
Nokia 5230	403.0	489.3
MOTO XT301	165.5	292.8

Table 4 shows as two kinds of mobile phone forecasting price and its error analysis among four months.

TABLE IV. FOUR MONTHS FORECASTING PRICE AND ERRORS

Mobile Phone	1			2			3			4		
	actual	*forecast*	*error*	*actual*	*forecast*	*error*	*actual*	*forecast*	*error*	*actual*	*forecast*	*error*
Nokia 5230	965	965	0.00%	952	965	1.37%	952	960.67	0.91%	952	956.33	0.46%
MOTO XT301	1066	1066	0.00%	1066	1066	0.00%	1066	1066	0.00%	1066	1066	0.00%

The four months price forecasting can be achieved more then 98% and it only depend on incompletely extracting data. So it can be touch very useful application for the shopkeeper's selling online.

The market of mobile phone change very fast. So the price extracting and forecasting are pay more attention by the shopkeepers. Our future interesting work is on the decision support system for the mobile phone selling online.

ACKNOWLEDGMENT

This work is supported by the Industry-University Collaboration Project of Huaian City, China (HAC201002); the fund of Huaian Industry Science and Technology, China(HAG2010064, HAG08009).

REFERENCES

[1] http://www.dictionary30.com/meaning/Forecasting.

[2] Zhang Gui-xi and Ma Li-ping, "An Introduction to Forecast and Decision," Beijing: Capital, Economic and Trade University Press, 2006

[3] Cheng-Zhong Xu; Ibrahim, T.I.; A keyword-based semantic prefetching approach in Internet news services. Knowledge and Data Engineering, IEEE Transactions on Volume: 16 , Issue: 5. 2004 , Page(s): 601 – 611

[4] Grobelnik, M.; Mladenic, D.; Fortuna, B.; Semantic Technology for Capturing Communication Inside an Organization. Internet Computing, IEEE. Volume: 13 , Issue: 4. 2009 , Page(s): 59 – 67

[5] Litecky, C.; Aken, A.; Ahmad, A.; Nelson, H.J.; Mining for Computing Jobs. Software, IEEE. Volume: 27 , Issue: 1. 2010 , Page(s): 78 – 85

[6] Hsinchun Chen; IEDs in the dark web: Lexicon expansion and genre classification. ISI 2009. Page(s): 173 – 175

[7] http://baike.baidu.com/view/58062.htm

[8] Wayne L.Winston, "Operations Research Introduction to Probability Models," Tsinghua University Press, 2006.

[9] Han Bo-tang, "Management Operations Research", Higher Education Press, 2010.

Research on ecosystem evolution of Mobile E-Commerce

Shiying Yu
School of Information Management
Wuhan University
Wuhan, China
shyyu2065@yahoo.com.cn

Lu Xiong
School of Information Management
Wuhan University
Wuhan, China
xionglu198921@126.com

***Abstract*—Due to the rapid increase of mobile users and in-depth development of mobile communication technology, mobile e-commerce has been rapid development. A hundred billion of industry ecosystem is becoming visible. It is one of the hottest application for transaction for today's retail industry. In this paper, we mainly describes some important issues in mobile commerce ecosystem. At first, we explain what is mobile e-commerce ecosystem, then made an analysis on its participants, elaborates the hierarchical relationship beyond the participants and finally analysis the growth model of mobile e-commerce in the view of ecological angle that launch of its Evolution Path. The purpose is to understand how Mobile E-commerce evolve.**

Keywords-mobile e-commerce ecosystem; ecosystem; mobile e-commerce

I. INTRODUCTION

With the development of 3G network technology and the delivery of 3G licenses certified, mobile e-commerce began to enter its golden period of development. It broke the traditional e-commerce model to adjust and reform the structure and content of e-commerce. What's more, mobile e-commerce ecosystem have benefited from faster transaction time, increased spending and enhanced customer loyalty while mobile consumers enjoy the convenience of use, speed of transaction and increased security[1]. The development of mobile e-commerce ecosystem related to the entire e-commerce development, it is critical to establish healthy mobile e-commerce system.

Prior research mostly focuses on the analysis of the value chain, technology applications or e-commerce ecosystem ,while there is a scarcity of research related to mobile e-commerce ecosystem. It is one of the purpose to write this article.

II. OVERVIEW *OF MOBILE* E-COMMERCE ECOSYSTEM

A. The concept of Mobile E-commerce and its ecosystem

Mobile e-commerce is an activity realized through wireless communication device combined with Internet technology, mobile communication technology, and other information processing technologies that make people pay online, provide integrated financial services, trade and related business activities at any place and any time[2]. Mobile commerce is the outcome of the further development of e-commerce. Advances in various technologies have made mobile e-commerce possible. 3G network development has brought greater opportunities for mobile e-commerce that has become a technology sign of the times. Furthermore, Mobile e-commerce extends a series of related businesses, that take shape the mobile commerce ecosystem.

Mobile commerce ecosystem refers to that in the eco-system of mobile e-commerce environment, the businesses of the mobile e-commerce activities as a complete economic entity to provide nutrients and conditions to others in this system. It is a co-dependency and co-evolution imitation of natural ecosystems system. Each member of the system closely linked with each other to form an organic whole, and jointly promote the development of national and global economy.

B. Mobile E-commerce ecosystem participants

According to the concept of the ecosystem, we can draw further analysis of the formation and development of mobile e-commerce. It is in a business environment that is consist of but is not limited to relevant government departments, content providers, network platform provider, logistics companies, financial institutions, mobile network operators, mobile device providers, wireless end users and other ingredients who collectively build up the architecture for this ecosystem[3][4]. As is shown in the figure 1, the Government provided basic industry development environment to the entire ecological system, the company in the ecological chain cooperate with each other which is called symbiotic relationship, similar enterprises compete with each other, these jointly support and promote the development of mobile commerce ecosystem.

In the entire ecosystem, policy environment is changing, industry and enterprises paying more attention to mobile e-commerce, the participants in the ecosystem are positively thinking their position and making more refined division of labor, all these promote the formation of mobile commerce ecosystem in a virtuous circle way.

1) The relevant government departments: As an emerging industry, mobile e-commerce is still in the growth stage. It is inevitable that many things are in the process of exploration, during the process there are many problems. Relevant government departments act as a cushion market regulation in the mobile e-commerce transaction process. Its attitude (promotes or limits) to implement the development of mobile e-commerce, the

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.78

policy guidance, funding tilt and other actions have a direct impact on the entire mobile e-commerce ecosystem environment.

2) Content providers: Generally speaking, content providers can be divided into two categories--one is commercial content providers, the other is information content provider. Commercial content providers are online businesses in the traditional e-commerce, they provide users variety of commodity information and transaction services through the mobile terminal equipment. Information content provider refers to the news source of information consuption in the network environment, or businesses who providing digital content such as text, video,music and so on.

3) Network platform providers: Corresponds to the content providers, network platform providers mainly can be divided into two categories, including the Internet trading platform provider and information platform provider. They build a communication bridge between consumers and service providers and play an extremely crucial role in the entire ecosystem.

4) Logistics enterprises: Traditional logistics enterprises in e-commerce still supply for mobile e-commerce, which are responsible for the transfer of goods to deliver products or services directly to users. Following two-dimensional codes with RFID technology developing, logistics enterprises make logistics activities more effective and precise.

5) Finanace institutions: They mainly refers to banking, securities and third party payment companies, that mainly provide capital flow services for mobile e-commerce transactions to promote the development of mobile e-commerce, They provide us security and trust, effectively promote the trading formation.

6) Mobile network operators: Mobile network operators who deploy ,operate and manage the infrastructure of access points, and networking equipment has been in the state that is called "Raised up by Xiao He and Cast down by Xiao He " in China[5]. Previously mobile e-commerce has already begun the layout, but due to lack of 3G network application popularized, the market has been in the initial state, have no further improvement, China entered the era of 3G applications, mobile e-commerce have begun to truly enter into the practice. Network operators provide a good environmernt for the entire eco-system, which is the leader to cultivate new market.

7) Mobile terminal equipment providers: They mainly refer to manufacturers who provide users with a variety of mobile communications equipment. They are responsible for carrying mobile network operators and provide a hardware foundation for good functioning. The development of mobile e-commerce, to a certain extent, also depends on the strength of the network carrying capacity. Their continuous optimization of the carrying capacity of the network and software applications for mobile e-commerce promotes the development of mobile commerce.

8) Wireless users: Wireless uses can conduct mobile e-commerce in several different ways. It could be through infrastructure-based wireless networks such as cellular, PCS, and GSM networks or could be through on and hoc wireless network that can connect directly via satellite-based systems[6]. Most mobile users are directly transferred from the mobile phone users. With the PC-side e-commerce users penetrate into the mobile terminal side, mobile e-commerce users' consumption structure will be similar to traditional e-commerce users. They choose what they prefer to through portal website, then pay for the bill through third-party department, finally logistics enterprises receive the order and distribute products to their customer.

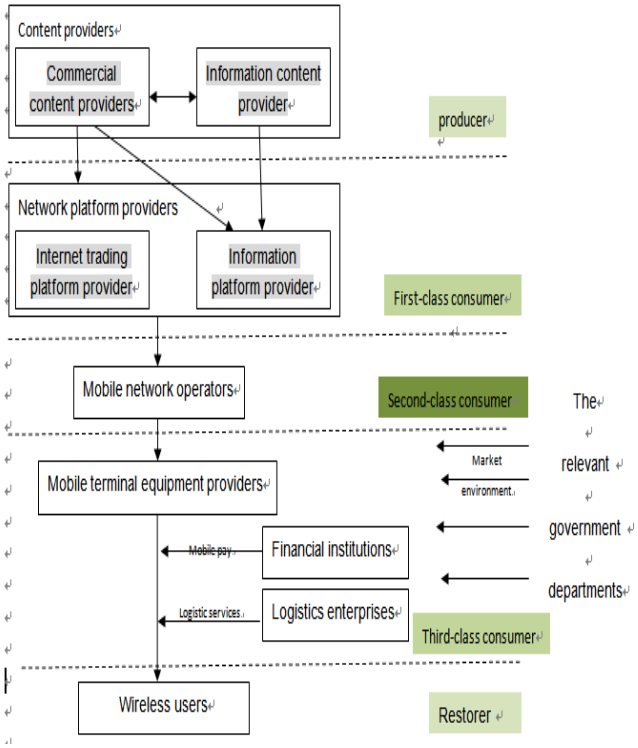


Figure 1. Mobile e-commerce eco-system relationship diagram

C. Ecosystem evolution environment of mobile e-commerce

The development of mobile e-commerce is not only impact by inner business environment, but also by the outside business environment from different levels. Evolving of mobile communication technology and users behavior's changing are directly affecting the entire ecosystem development path. At present, China's credit system is not perfect, the network consumer consciousness lower, laws and regulations are incomplete, lack of network security those things have seriously hampered the evolution of mobile e-commerce ecosystem. The above issues are likely to affect the widespread use of mobile e-commerce applications. It is possible that some of these participates maybe realize the important of the environment, but it should be noted that by proving more nurture each other to improve the ecosystem environment, entire ecological chain might be significant. It is still need to nurture the participating classes, make them more interaction to provide good service to the entire ecological chain.

III. EVOLUTIONARY PATH AND GROWTH MODEL OF MOBILE COMMERCE ECOSYSTEM

A. Growth Model of ecosystem

In the mobile e-commerce ecosystem, the ecological relationship between the main body is complicated, the basic relationship between them is competition and symbiotic. In 1940s, Lotka and Volterra put forward inter-specific competition equation which has a significant impact on the modern ecological theory, and mobile e-commerce eco-system itself existing is modeled in the form of natural ecosystems, the same applies to their proposed Lotka-Volterra model of inter-specific competition. In this paper we introduce this competition model to the mobile e-commerce. The concept of species populations is mainly used in this model, and each major participates of the mobile e-commerce eco-system can be regarded as a species. In the modeling process, we call it participate class the time being. The following targeted ecological competition and symbiosis relationship to establish the growth model of mobile commerce ecosystem [4].

1) The growth model of the ecosystem under competition relationship

In the mobile commerce ecosystem, there are internal competition between the participating classes, such as terminal equipment that include mobile phone providers, Pocket PC, PDA and other categories, they compete with each other for customers among the common resources. What's more, each category also contains a variety of brands, the competition exist inevitably. In the case of limited resources, this competition relationship will make the resources of participants tilt to the winner. We can make an assume that there are N type businesses in a participate class, where the competition coefficient between the ith and jth categories can be expressed by a_{ij} ; k_i is used to describe environmental capacity that include policy environment, market environment and so on; r_i refers to class growth rate; $1-\frac{\sum_{j=1}^{n} a_{ij} N_j}{k_i}$ means free environmental capacity. We can draw the growth model of several participates under competition, the model is shown in equation (1).

$$\frac{dNi}{dt} = r_i \ N_i(1-\frac{\sum_{j=1}^{n} a_{ij} N_j}{k_i}) \cdot \qquad (1)$$

(Note: Due to the same categories is in a fully competitive relationship, so a_{ii}=1)

In certain circumstances, environmental capacity(market size) is limited, resources title begin to emerge because of the competition between the various categories .we can see from the above model, when such competitors become more, the growth factor Will become smaller and smaller.

2) The growth model of the ecosystem under symbiotic relationship

In the mobile commerce ecosystem, symbiotic relationship exists between those different species, for example, content provider provide more value-added services to mobile network operators to attract more customers, while content providers parasites on the mobile operators to obtain the corresponding profits. They work together to improve the environmental capacity of each other. We can draw the growth model of several participates under symbiotic relationship, the model is shown in equation (2).

$$\frac{dN_i}{dt} = r_i N_i(1-\frac{N_i}{K_i+\sum_{j=1}^{n} b_{ij} N_j}) \cdot \qquad (2)$$

(Note: Due to there is no symbiotic relation that inter-dependence between species is absent ,so $b_{ii}=0$)

In which N means the number of participation of class, the symbiotic coefficient between the ith and jth categories expressed by b_{ij} ; k_i is environmental capacity; r_i is stand for the growth rate of the participating classes; $1-\frac{N_i}{K_i+\sum_{j=1}^{n} b_{ij} N_j}$ is free environmental capacity. In Entire ecosystem, these two relations are exist whether inside or outside the class, and the complex interactions between them promote the formation of ecosystem evolution paths.

B. The ecosystem evolution path

The emerging of evolution path is not in accident, the interaction results between the above two relationship are displayed by the number of above users. The results present in the form of a number of small points, as time going on, the path of evolution emerge when we connect these points. From the above discussion, we can see the evolution path is to present a nonlinear relationship. In addition, we can follow the traditional path of ecosystem evolution, the evolution can be divided into four stages, namely development stage, growth stage, maturity stage, decline stage (Figure 2)[8].

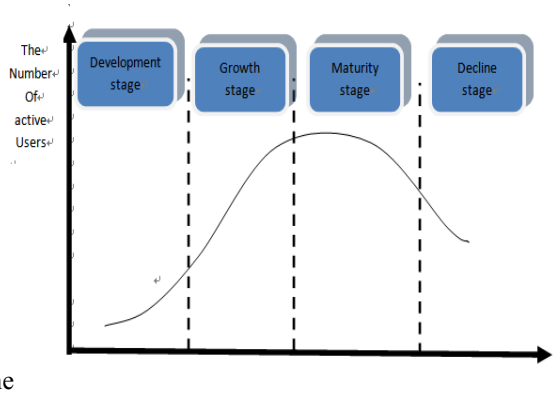


Figure 2. The ecosystem evolution path of mobile e-commerce

Development stage means part of traditional e-commerce customers began receiving the form of mobile e-commerce in conscious, the participants of ecosystem didn't include financial institutions, information value-added service providers and so on. Technology development and operation mode consummation promote the expansion of the mobile e-commerce ecosystem. During the period there have been many representative incidents happen. In 2004, Alibaba.com constructed the first mobile e-commerce platform in china combine with the Internet, then in 2007 Trade Manager on mobile realize online and offline mobile interconnection.

With the continuous development of ecological systems, payment security has become a key factor. Breeding classes in support of the development of the whole class, More and more members join in the class make the system expansion in scale and parasitic class also began to emerge. Competition becomes intense and natural selection begin to function. During this period, Taobao.com and alipay.com join the queue of mobile e-commerce.163.com also started a so called "toocle" service that is a kind of business search engine used in mobile electronic commerce.

The rapid growth of the participating classes make the dependencies and competition between them become increasingly complex, and produce more interest conflict, the ecosystem began to enter the self-adjustment status in other words maturity state. Mobile e-commerce industry chain is forming, which includes the enterprises, businesses, e-commerce providers, telecom operators and other business systems, but also the supporting role of the terminal manufacturers, financial and payment service providers, logistics providers and other types of service providers. Each subject in the system interact with each other through information flow, capital flow and logistics, to play the role of providing interfaces, applications and services.

External environment changing such as policy environment and new models put forward further evolution of the entire ecosystem, the ecosystem began to enter into recession state when the number of active users began to decline, other businesses began to split share of mobile e-commerce. It began to slowly shift into the next stage.

IV. CONCLUTION

This paper introduce ecology view into the mobile commerce ecosystem evolution, but its evolution is still slowly advancing, there are many issues to resolve. For example, insufficient bandwidth is difficult to support non-text information transmission in mobile commerce; online payment, security certification, distribution offline and other issues is especially prominent in the system. Ecosystem of mobile e-commerce still have a long way to self-improvement, it's in the rapid formation and development period at present. There are many things we need to investigate on ecological system of mobile e-commerce in future studies. We can be verified through some examples to validate the above models so as to better guidance ecosystem research of mobile commerce.

ACKNOWLEDGE

We would like to express our sincere thanks and warmest appreciation to those industry experts for sharing their valuable thoughts and incredible insights.

REFERENCES

[1] P.M.Lai, K.B.Chuah, "Developing an Analytical Framework for Mobile Payments Adoption in Retailing: A Supply-Side Perspective," proc. IEEE Symp. 2010 International Conference on Management of e-Commerce and e-Government, IEEE press. 2010, pp.365-371, DOI :10.1109/ICMeCG.2010.79

[2] Baidu Encyclopedia, "Mobile e-commerce" [EB / OL]. Http://baike.baidu.com/view/41363.htm.2011-3-30 .(In Chinese)

[3] Wang Na, "A study on ecological system model of Mobile e-commerce value net", Beijing University of Posts and Telecommunications pp. 2009-3. (In Chinese)

[4] Zhang Jing, Huo Yumei, "An analysis on the business ecosystem of mobile game," Beijing University of Posts and Telecommunications (Social Science Edition), Aug. 2008, pp.16-19 (In Chinese)

[5] David Wright, "The Role of Wireless Access Interconnection in Mobile e-Commerce Industry Evolution, " Eighth World Congress on the Management of eBusiness(WCMeB 2007), IEEE press. 2007

[6] Upkar Varshney, "Mobile Commerce: Framework, Applications and Networking Support," Mobile Networks and Applications 7, 2002,pp.185-198.

[7] Sheng Zhen, "Stydy On the growth of population in Taobao.com ecosystem," The second session Conference of ecological network operators and e-commerce Proceedings. 2009.(In Chinese)

[8] Hu Gangfeng, Lu Xianghua, Lihua, "E-commerce ecosystem and its evolution path. Economic Management ," Jun.2009, pp.110-116. (In Chinese)

A Typical Study of a Cooperation between an E-commerce Company and a Higher Vocational College

Song Zhijie
School of economic management
Yanshan University
Qinhuangdao，China
songzhj@ysu.edu.cn

Wang Guohong
School of economic management
Yanshan University
Qinhuangdao，China
wgh8281@163.com

Wang Nan
Economic department
Enviromental management college of China
Qinhuangdao，China
wangnan@163.com

***Abstract*- How does the higher vocational liberal arts major，such as Marketing, establish work-integrated learning practice system with enterprise on the basis of school-enterprise win-win? It is both the necessary condition that cultivates currently the high-skilled talents to satisfy the needs of society, but also the difficulty in the process of many institutions' professional construction. This article proposes the innovation idea that Marketing major constructs the work-integrated learning project by analyzing of the school-enterprise cooperation project between Environmental Management college of China and Beijing 2688 e-commerce Company.**

Keywords-work-integrated learning ; order-based training ; school-enterprise cooperation ; Marketing, practice teaching

TABLE I. I. BACKGROUND

Following the deepening of market economy system reform in our country and the intensification of market competition, the needs of the enterprise for Marketing major talents not only increase on the number but also improve on the quality requirements. Compare higher vocational talents who major in Marketing with undergraduates and higher academic degree graduates, the requirement of the enterprise for basic theory is relatively loose, but for practical operation capacity is higher and finer.

How to closely combine the talent needs of those top enterprise, construct work-integrated learning practical teaching links and achieve the order-based training on higher vocational marketing major, thus finally to achieve sustainable development of Marketing major. These have been the confusion and bottleneck which is faced not only by higher vocational Marketing major but also by other higher vocational liberal arts majors.

According to the cultivation concept of work-integrated learning and following the premise condition on basic fundamentalist of market economy, Environmental Management College of China and Beijing 2688 E-commerce Company launched a three-year school-enterprise cooperation which introduced real market project as a practical teaching content for marketing major based on market conditions in the campus. So far, 80% graduates participated in this practice project. They entered to this company to work, highly welcomed by the company, and 50% among them has promoted as the regional manager. Meanwhile the recruit status of the marketing major continue to rise, so the cultivation mode of work-integrated learning get through passageway of recruitment, cultivation and employment for marketing major.

TABLE II. II. PROJECT INTRODUCTION

Beijing 2688 E-commerce Company is professionally engaged in the B2C e-commerce and realizes the C.O.D. network distribution mode by developing particular distribution mode of "procurement service". Marketing major of Environmental Management College of China and Beijing 2688

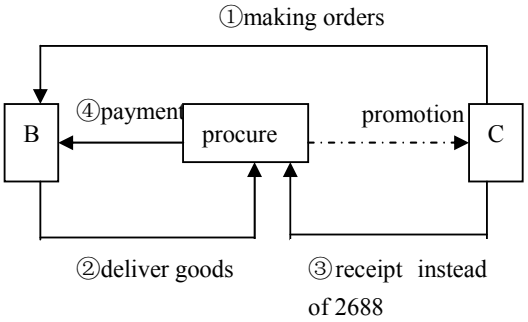
978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.17

e-commerce Company (hereinafter referred to as 2688) have signed "agreement of procurement service in the campus market" in 2007, the main contents of the agreement are as follows:

Party A (2688) opens qualification permission for party B (Environmental Management College of China) that act as purchasing agency of the campus' market, and provide the corresponding marketing publicity and material support. Party B develops to acts as purchasing agent based on www.2688.com network platform in the market permission which is opened by the party A. And the concrete content includes platform promotion, goods scheduling of online order, payment credit, etc. Party B allocates 40% of the sale of goods as profits in accordance with the provision of contract.

B2C E-commerce workflow of 2688 "procurement service" is shown as follow, and in it B is 2688 e-commerce platform, C is the end customers. The end customer registers firstly through Internet platform and browse, and after these operations will submit the shopping orders and then 2688 platform arrange to send the goods after verifying to the nearest place of procurement service to the customers which is the system matched automatically or to the place of procurement service that selected by the customers. The procurement service place is responsible for providing delivery and collects payment on behalf of the company.

I. TABLE1， 2688 CORP. WORKFLOW



The whole process of purchasing the company firstly receive the payment in advance from purchasing agency, the consumers just need to pay the money to the purchasing agency after accepting the goods. The work that the purchasing agency should do is not only accepting the goods, delivering the goods and collecting the payment but also coordinating the returns of goods, exchanging the goods and consulting the disputes with 2688's customers, etc. Meanwhile in order to ascend achievements and increase the income, the purchasing agency needs to actively explore the market, promote sales promotion to the customers, try to increase sales achievements and set up the market image of the website.

III. PRACTICE SYSTEM STRUCTURE

A. . The practical Teaching Link

Under the framework of School-enterprise cooperation agreement, marketing major of Environmental Management College of China (hereinafter referred to as EMCC) constructed practical teaching link for 1.5 semesters which take the 2688 market project as the main line. The practice content covers the entire core contents about the marketing course, such as market environment analysis, market research, etc. in addition, can effectively combine with customer management, market survey prediction course, etc. The specific time content are shown as follow:

Tabel2. PRACTICE CONTENTS

Practice link	***Main contents***	***Check form***
Marketing environmental analysis	Micro and macro environment analysis of campus market on 2688online shopping	Analysis report
Market research and forecast	investigation and forecast on campus market potential of 2688 online shopping platform, investigation on customers' behavior, customers' market information	Investigation report

	feedback, etc.	
Promotional design and implementation	Formulation and implementation on long-term sales promotion plan, Formulation and implementation on festival and holiday	Scheme; implementation effect
Channel design and implementation	Design and implementation on Channel Development Scheme	Design; implementation effect
Customer service management	Establishment of customer database, Customer Maintenance, New customer development	satisfaction ; the number of new costumers
Project operation and management	Management of accounts, commodity and practitioners, etc.	Daily order

The whole practice link begins at the seco The whole process of purchasing the company firstly receive the payment in advance from purchasing agency, the consumers just need to pay the money to the purchasing agency after accepting the goods. The work that the purchasing agency should do is not only accepting the goods, delivering the goods and collecting the payment but also coordinating the returns of goods, exchanging the goods and consulting the disputes with 2688's customers, etc. Meanwhile in order to ascend achievements and increase the income, the purchasing agency needs to actively explore the market, promote sales promotion to the customers, try to increase sales achievements and set up the market image of the website.

B. The Post *Practice* Link

After market practice in the campus for 1.5 semesters, students will enter 2688 company to practice for six months to a year. The post of practice is mainly market promotion, channel management and customer service positions.

The post practice Relative to the campus practice helps students rise from the purchasing agency, namely enterprise channel distributor to enterprise channel management level, and the main work content is communicating with channel distributor, like the purchasing agency, and first customer. Preliminary campus practice laid a steady foundation for the post practice. Enterprise said that students are generally familiar with company business, and approve the company philosophy, and students are engaged in work quickly.

In post practice link, because the students haven't graduated yet, they accept the company's management and the professional teachers' guidance on employment and specialty. The post practice link not only realizes the further promotion of students' practical ability, but also helps students to understand the company further and communicate with the company deeper, and lay the foundation for the realization of next order-based training employment.

C. The employment realization link

Through the campus practice and post practice, the students are not only familiar with 2688 business mode, they also cultivate loyalty for enterprise, meanwhile it can effectively combine relevant basic theoretical knowledge and practical work through long-term actual practice, and the students will be logically and preferentially employed.

The introduction of this project help the students majored in marketing of environmental Management

College of China start from the real market practice in the campus, and then go to the enterprise to practice as the transition, finally realizes the employment. It forms a precise and complete process of order-based training and it provides a good model for domestic other similar major on the construction of work-integrated learning system and formation of order-based training mode.

□.ENLIGHTENMENT

A. The school-enterprise win-win, practice link should be introduced into the real project

School-enterprise cooperation will be built as far as possible on the basis of school-enterprise win-win, and put the introduction of real project as the support of practice teaching based on the campus market or other available market. Work-integrated learning blindly relies on the enterprise to provide internship positions which will lay the burden upon the enterprise and reduce enterprises' enthusiasm to participate in the education; meanwhile the students will lack of interest in the project without real pressure, and ultimately affect the sustainable development of the school-enterprise cooperation.

It should seek for the entry point of win-win in the nature of enterprise pursues profits and the objective of talent training in the higher vocational education. By developing operation of a real market project which is guided by professional teachers and managed by the enterprise personnel in the campus market, it can effectively reduce the cost, optimize the allocation of resources, and improve the participant enthusiasm of many parties.

B. Real situation, joint management and full responsibility for one's own profits and losses

Real project can give students the real situation, only in completely real teaching situation, students can have the most direct and rational cognition to the market, and can truly realize effect mechanism of the market discipline. Management of practice process insists common management of enterprise and specialty, and the teachers as expert guide participate in practice guidance, the students participate in practice as "company employee" should accept the company management. It will for students construct the campus practice link which not only has the teachers' guidance but also do a real and practical work.

In budget aspect, it also imitate the real business that students independently check and take full responsibility for one's own profits and losses in order to supervise students to completely stand in the real environment when making decisions. Students should rationally consider the cost, risk factors, profit, and other elements and not easy to make general students' fault that only considers effect and regardless of the cost.

C. Step by step, the theory study integrates into practice

Practice link is to better melting and understanding theory knowledge, and the theory will be better used in future real work. Therefore in the practical process professional teachers, in particular, should pay attention to students who blindly pursue profits and neglect the fusion with theoretical knowledge in this period.

How to design practical task which goes on step by step and combines closely with professional knowledge and students can be constantly reminded practice target in the practice process from the initial understanding the market and generating profits to planning and developing promotional activities, customer management, etc.

4. Adapt to the environment, marketing develops the orientation training

The training of higher vocational marketing major shall adapt to the needs of different industries and enterprises in combination with the characteristics of regional economy. We can carry out orientation training mode, and establish school-enterprise win-win cooperation training mode of work-integrated learning with some enterprises based on actual situation.

REFERENCES

[1] Zhou Jiansong, Guo Fuchun. Develop order-based training and realize "zero transition" between students graduated and mount guard [J]. China higher education, 2005, (19): 39-40.

[2]. Liu Xiaoqin. Discuss the basic conditions of talent training mode on "order-based" [J].2nd seminars on countrywide higher vocational principal joint conference, 2004.

[3] Wang Guohong, Fu Yixin, Zhang Yuqing. Exploration on establishment the stimulant company in training room in higher

vocational marketing major [J]. Vocational and Technical Education, 2009, (5): 69-70.

[4]Liu Hua, Wang Guohong, Li Keguo. Exploration on talents training mode operation mechanism of work-integrated learning under the employment guidance. [J]. Adult Education, 2009, (9): 26-29.

[5] Chen Jiefang. Education theory on cooperation and its practice in China - education mode study combining learning and working. [M]. Shanghai: Shanghai jiaotong University Press, 2006.

[6] Huang Aijin. University management faces reform of the new economic era [J]. Higher Education Theory and Practice, 2001, (6).

[7], Wang Guohong, Xu Hui, Wang Nan. Implementation path of order-based training mode in Higher vocational college –take marketing major as an example [J]. Journal of Environmental Management College of China, 2009, (4).

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Research on the Relationships among Expectation, Perceived Service Quality and Customer Satisfaction

XU Xian-ying
Economic and Management School, Shenyang
Aeronautical University
Shenyang, P. R. China, 110136
e-mail: xuxianying234@sohu.com

***Abstract*—In order to reveal the effecting mechanism produced by expectation to service quality and customer satisfaction, we made the studies on the relationships between expectation, perceived service quality and customer satisfaction. Based on the attributes division of Kano's model, this paper built the structure model about the relationships of must-be attributes expectation, one-dimensional attributes expectation, attractive attributes expectation, perceived service quality and customer satisfaction, and made the hypothesis. This paper tested the hypothesis by choosing the restaurants as the empirical study background, and the results showed that different attributes expectation has different impacts on perceived service quality and customer satisfaction.**

Keywords-Expectation; Perceived service quality; Customer satisfaction; Kano Model

I. INTRODUCTION

Expectation, perceived service quality and customer satisfaction is the important variables with mutual interactive in marketing practice for service enterprise. Among them, perceived service quality and customer satisfaction are the base for the enterprise to retain customers and gain profits. Expectation impacts customer comprehensive evaluations (perceived service quality and customer satisfaction), and is influenced by business marketing strategies. In literatures, the effects to service quality perception or customer satisfaction by expectation are different, even conflict. Such as positive influence [1, 2], negative influence [3, 4] and have no significant influence [5]. All these different results, provide the paradox theory direction for enterprise marketing communication, the dilemma is that whether to deliver high expectation or low expectation in the marketing strategy chosen.

Maybe the reasons for the different results were related to the different understanding of definitions about expectation, perceived service quality and customer satisfaction, or influenced by the different criteria when the interviewees answer the questionnaires, or because the differences in the research situations. But all these reasons can not explain the mechanism from expectation to customer comprehensive evaluation. The dilemma in theory and practice call for further thinking: what is the effect mechanism by expectation to customer comprehensive evaluation, and the reasons for the different results. The answers to these problems will be benefit to the implement of scientific method to manage expectation and improve the perceived service quality and customer satisfaction effectively.

II. LITERATURE REVIEW

A. Positive influence from expectation to customer comprehensive evaluations

In perceived performance model, expectation has positive influence to customer comprehensive evaluation, it is the anchor effect. Assimilation theory in society psychology can give the reasons: the difference between expectation and perception will incur the dissonance to customer and negative emotions. So in order to avoid the negative feelings, when the difference between expectation and perception was not big, customers prefer to adjust the perceived performance to expectation directions and degree. Through this mechanism, expectation increases customer comprehensive evaluations.

In empirical studies, Ladhari concluded that expectation had positive influence to customer perceived service quality [1]. Anderson and Sulliven also supported that high expectation will increase customer service evaluations [2]. The comprehensive evaluation was high when expectation was high and perception was high, and the comprehensive evaluation is low when expectation was low, even perception was high.

B. Negative influence from expectation to customer comprehensive evaluations

Supporting theories for the negative influence were Oliver's expectation disconfirmation [6] and Parasurama，Zeithaml and Berry's gap model [7]. All these theories claimed that before purchasing, the customers formed their expectations about the product based on the past experience and advertisement, and then they perceived the performance of product when using, finally compared the expectations with perceived performance, and judged the gaps between them. In the above process, expectation played as the standard role. It's the gaps between expectation and perceived performance decide customer comprehensive evaluations. So, if the expectation is higher the positive gaps will be smaller, and the comprehensive evaluations will be lower.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.81

C. *Expectation has no significant influence to customer comprehensive evaluations*

In the empirical studies of Burke, they got the results about no significant influence from expectation to customer comprehensive evaluations [5].

According to the above researches conclusions, we know that the integrate expectation has different influence to customer comprehensive evaluations. And because the formality was different, expectation produced different influence dependent on the industries. Expectation is a multi-dimensional construct had been approved in literatures [8, 9]. To consumers, they had different familiar degree to different attributes even in one enterprise. Further researches were needed to test whether expectation to attributes with different familiar degree will have different influence to customer comprehensive evaluations.

III. CONCEPT MODEL AND HYPOTHESISES

From the theory definition of expectation and the formation process of customer expectation, we got that the operational concept of expectation including two dimensional features: importance and possibility, they are two independent aspects for customer expectation. For example, "provide immediate service" expectation, reflects "the importance of immediate service to meet the customer's needs", such as two persons who have enough time will think "immediate service" is low important attribute relatively than who have no time, and "the forecasting possibilities for the realization of immediate service by customers".

In these structure models about expectation, the measurement to expectation always adopts the "predicting expectation". Under this definition, expectation understood as the forecasting possibilities of attributes performance. And the "importance" dimension of expectation was ignored in the literatures. This paper considered all these two dimensions to explore the influence from expectation to customer comprehensive evaluations.

Using two factors theory, Kano gave three attributes: must-be attributes, one-dimensional attributes and attractive attributes. As experience accumulating, three kinds of attributes will be changed from attractive attributes to one-dimensional attributes and must-be attributes. With this changing, the formalities to three kinds attributes were from low to high, they represent the attributes which are difficult to guess for customers, which can be thought by the customers and which are familiar to customers respectively.

So three kinds attributes by Kano model are different in customer's familiar degree and importance to meet customer's need. Therefore, we referred to the Kano division table to divide the importance dimension of expectation and got three kinds of expectations as figure 1 shows.

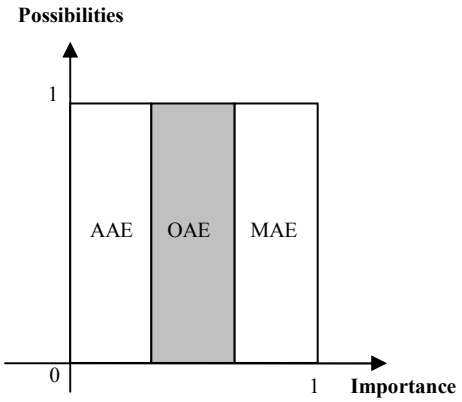


Figure 1. Expectation groups based on the important dimension

Attentions: (1) MAE means must-be attributes expectation; OAE means one -dimensional attributes expectation; AAE means attractive attributes expectation;
(2) MAE, OAE and AAE are the same meanings in Figure 2 and F

Next, we will test the relationships among three kinds of attributes expectations, perceived service quality and customer satisfaction. The concept model referred to figure2.

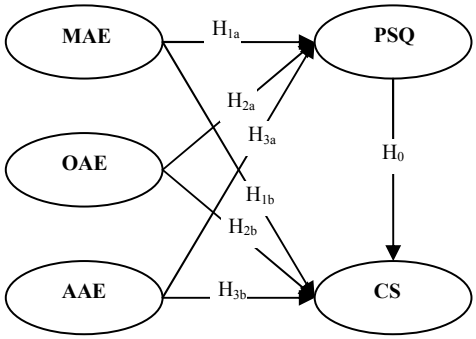


Figure 2. Concept model of relationships about expectation, perceived service quality and customer satisfaction

Attentions: (1) PSQ means perceived service quality; CS means customer satisfaction.
(2) PSQ and CS are he same meanings in Figure 3.

A. *Hypothesises on the relationships among must-be attributes expectation, perceived service quality and customer satisfaction*

Must-be attributes are these features which will be showed in any enterprise of the same industry, they are the core elements to meet customer needs, and do not changed with the different enterprise. So the customers are familiar with these must-be attributes, for they always have enough industry experience for these attributes. They knew what they need clearly and compared with the performance they perceived. In this process, expectation played as standard role, and had negative influence to comprehensive

evaluation. Therefore, the following hypothesis has been made:

H1a: must-be attributes expectation has negative influence to perceived service quality.

H1b: must-be attributes expectation has negative influence to customer satisfaction

B. Hypothesises on the relationships among one-dimensional attributes expectation, perceived service quality and customer satisfaction

One-dimensional attributes are the qualitative attributes, their performance will impact the customers choice, that is the customers will choose one service refer to one-dimensional attributes performance. The performance of one-dimensional attributes was always influenced by some situational factors with uncertain. For customers, one-dimensional attributes are important with relative low familiars. That means, although customers gave the forecasting possibilities of one-dimensional attributes performance, they are not sure for the realization. Referring to the pursuing to positive affect, and devoid for the negative affect. When consuming the service, customers will pay attention to the information which will prove their expectation. From this way, expectation has positive influence to comprehensive evaluations. So put forward the hypothesis:

H2a: one-dimensional attributes expectation has positive influence to perceived service quality.

H2b: one-dimensional attributes expectation has positive influence to customer satisfaction

C. Hypothesises on the relationships among attractive attributes expectation, perceived service quality and customer satisfaction

Attractive attributes are expressed by customer fewer, sometimes, even the customer themselves do not know their latent wants. Maybe attractive attributes are not very important to meet customer needs, but their realization can give surprise to customers. So they are low importance and low familiar to customers.

We made an experiment for the influence from attractive attributes expectation to customer comprehensive evaluations. Tell the interviewees that there is one restaurant which is very special, for example, thinking about the customers with glasses, they will provide the glasses cloth, they also serve free internet service, free clean shoes etc. Then measured the interviewees' forecasting for the realization of these attributes, and we got two groups: high forecasting group and low forecasting group. Next we declared the situation that the service provider provide, and measured the final evaluation. At last, we found customers in high forecasting group also have high evaluations, while low forecasting customers' evaluations are very common. To these phenomenon, we made deep interviews, and found that customers in the low forecasting group thought the corresponding attributes were fancies, so when we tell the service process, they did not pay attention to the extra service. The result was that although the company has done many things, the customers did not feel them. While customers in the high forecasting group had good evaluations, because they experiencing the additional service, and thought the experience add their values, and increased the perceived performance. The following hypothesis was put forward,

H3a: attractive attributes expectation has positive influence to perceived service quality.

H3b: attractive attributes expectation has positive influence to customer satisfaction

Finally, based on most literatures' conclusions, hypothesize perceived service quality influence customer satisfaction positively, that is,

H0: perceived service quality has positive influence to customer satisfaction.

IV. VARIABLES MEASUREMENT AND DATA COLLECTIONS

We chose restaurants as the experiment background. To the measurement of perceived service quality and customer satisfaction, we referred to the works of Dabholkar [10], Spreng and Mackoy [11]. The measurement of expectation was divided by two steps. First through literatures and interviews, we got the expectation content for restaurants service. Then use Kano table to divide the content into three kinds attributes. In the questionnaires, we only measured the possibilities dimension of expectation.

V. DATA ANALYSIS AND RESULTS DISCUSSIONS

A. Data analysis

The division of expectations referred to Kano table, and each kind of expectation was constituted by many attributes expectation. In data analysis, we think the attributes in one group have the same importance, and use the average of all forecasting value in the same group as the value of each kind expectation. Adopt Amos 5.0 to make the path analysis. Results refer to figure 3.

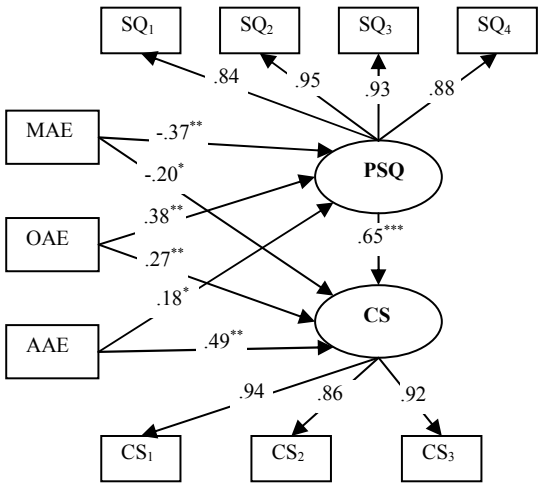


Figure 3. Path analysis results

Attention: χ^2 (31) =47.360, P =0.030, RMSEA=0.042, AGFI=0.902, CFI=0.965, IFI=0.928, TLI=0.921, NFI=0.932.

From figure 3, we know hypothesis 1b and 3a are significant under 0.05 levels; and hypothesis 1a, 2a, 2b and 3b are significant under 0.01 levels. The fit indexes also showed that the structural model was supported.

B. Results discussions

1) Must-be attributes expectation has negative influence to perceived service quality and customer satisfaction

The path coefficient to perceived service quality and customer satisfaction from must-be attributes expectation is -0.37, -0.20, which means the higher forecasting of must-be attributes performance, the higher possibilities of negative evaluations. So in the orientation of marketing, please do not make the customers to form high expectation for must-be attributes.

2) One-dimensional attributes expectation and attractive attributes expectation has positive influence to perceived service quality and customer satisfaction

The influence to perceived service quality and customer satisfaction from one-dimensional attributes expectation is 0.38, 0.27, and from attractive attributes expectation is 0.18, 0.49. These results suggested the enterprise should give customers an expectation space for one-dimensional attributes and attractive attributes. Such as, tell customers the features of enterprise, and how to use these attractive attributes, and the comfort ability after using.

3) The influence from integrate expectation to customer comprehensive evaluation is uncertain

In different industry, customer integrate expectation was formed by different quantity of must-be attributes expectation, one-dimensional attributes expectation and attractive attributes expectation. And because the must-be attributes expectation has negative influence to customer comprehensive evaluations; while the influence from one-dimensional attributes expectation and attractive attributes expectation was positive. So the influence from integrate expectation to perceived service quality and customer satisfaction was uncertain.

VI. CONTRIBUTIONS AND LIMITATIONS

This paper first put forward that expectation to attributes with different importance to customer needs has different influence to customer comprehensive evaluations. And explained why the results about the relationships among expectation, perceived service quality and customer satisfaction are different. And give clear directions to the enterprise to propagate based on the attributes features.

At the same time, there are some limitations. First, we use students as the simple samples to divide the expectation groups. And test the structure model only in the restaurants industry, the suitability for other industries should be tested further.

[1] R. Ladhari, "Alternative measures of service quality: a review," Managing Service Quality, vol. 18, 1, 2008, pp.65-86.

[2] E. W. Anderson and M. W. Sullivan, "The Antecedents and Consequences of Customer Satisfaction for Firms," Marketing Science, vol. 12, 2, 1993, pp.125-143.

[3] J. E. Bigne, A. S.Mattila and L. Andreu, "The Impact of Experiential Consumption Cognitions and Emotions on Behavioral Intentions," Journal of Services Marketing, vol. 22, 4, 2008, pp.303-315.

[4] L. M. Caro and J. A. M. Garcia, "Cognitive-Affective Model of Consumer Satisfaction. An exploratory Study Within the Framework of a Sporting Event," Journal of Business Research, vol. 60, 1, 2007, pp.108-114.

[5] K. G. Burke, S. E. Kovar and P. J. Prenshaw, "Understanding the Satisfaction Process for New Assurance Services: the Role of Attitudes, Expectations, Disconfirmation and Performance," Advances in Accounting, vol. 20, 1, 2003, pp.43-75.

[6] R. L. Oliver, "A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions," Journal of marketing research, vol. 17, November, 1980, pp.460-469.

[7] A. Parasuraman, V. A. Zeithaml and L. L. Berry, "SERVQUAL, a Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality," Journal of Retailing, vol. 64, 1, 1988, pp.12-40.

[8] W. J. J. Baker, "An Exploratory Study of a Multi-Expectation Framework for Services," Journal of Services Marketing, vol. 14,5, 2000, pp.411-431.

[9] M. Laroche, M. Kalamas, S. Cheikhrouhou and A. Cezard, "An Assessment of the Dimensionality of Should and Will Service Expectations," ASAC, 1, 2004, pp.361-374.

[10] P. A. Dabholkar, "The Convergence of Customer Satisfaction and Service Quality Evaluations with Increasing Customer Patronage," Journal of Consumer Satisfaction, Dissatisfaction and Complaining Behavior, vol. 8,1995, pp. 32–43.

[11] R. A. Spreng and D. M. Robert, "An Empirical Examination of a Model of Perceived Service Quality and Satisfaction," Journal of Retailing, vol. 72, 2, 1996, pp.201–214.

Effects of China's Outward FDI on domestic Employment

Yu Chao
Management Science and Engineering
Business school, Hohai University
No.1Xikang Road Nanjing, China
chaoyu@hhu.edu.cn

Zhou Ye
Management Science and Engineering
Business school, Hohai University
No.1Xikang Road Nanjing, China
zhouye325@hhu.edu.cn

***Abstract*—With the improvement of international competitiveness of Chinese economy and the implementation of going global strategy, China's outward FDI has developed quickly. This paper explores the effects of China's outward FDI on domestic employment by examining the data of 25 provinces from 2003 to 2009, and combines the Cobb-Douglas production function and system GMM estimation. The result shows that preliminary labor demand and current wage income have substitution effect on domestic employment, GDP and preliminary wage income have significantly positive effect on domestic employment, China's outward FDI can promote domestic employment, but the effect is not significant.**

Keywords: outward FDI; employment; system GMM

I. INTRODUCTION

As a developing country, China has large population and rich resources. How to alleviate employment pressure is a long-term problem we have to face. Depending on labor cost advantage and huge market potential, China has achieved great success in attracting foreign investment and expanding foreign trade Since reform and opening up. The influx of foreign capital and increasing trade can not only fill the financing gap in the process of economic growth but also promote the development of productivity and the upgrading of industrial structure. With further development of economy, a series of economic and social problems appear. First, the rapid increase of foreign exchange reserves has exacerbated inflationary pressure. Second, trade frictions have increased. The Commerce Department has pointed out that the trade frictions China faces are not only at traditional competitive industries, but also at high and new technology industries; not only from developed economies, but also from some developing economies now. Third, the degree of dependence on oversea supply has become higher, which means our economy is more vulnerable to the changes of uncertain international trade environment. Since 2008, influenced by the global financial crisis, China's foreign trade volume and foreign capital inflow have decreased, employment pressure increased. Under such circumstances, to invest abroad has become a good choice for many enterprises. As more and more enterprises choose to invest abroad, the impact of outward FDI on domestic employment is taken into account by more and more people.

Since reform and opening up, China has enjoyed rapid economic growth, but the development of outward FDI is relatively slow. Until 2001, when government increases support for enterprises engaged in overseas investment and economic cooperation, China's outward FDI begins to show a trend of rapid development. As an increasing number of firms choose to move production abroad, debate on the impact of such strategies on domestic employment has greatly intensified. The impact of outward FDI on the home economy has attracted researchers' attention for a couple of decades.

II. RELATED LITERATURE

When analyzing the effect of outward FDI on home country employment, it is most important to distinguish the differences between horizontal FDI and vertical FDI. Vertical FDI is made by enterprises that fragment their production into stages, typically on the basis of exploiting lower

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.40

factor prices or reducing transaction costs (Ekholm and Markusen 2002; Kokko 2006). It is reasonable to allocate the stages with high labor intensity to countries with low levels of labor cost and the stages requiring high and new technology to high income countries. With vertical FDI, there is a complementarity between outward FDI and domestic employment, because when one of the activities expands, it will cause the expansion of the other activity. Horizontal FDI is made by enterprises that seek to exploit their existing advantages and replicate roughly the same activities in many locations (Braconier and Ekholm, 2001; Massom and Vahter, 2009). With horizontal FDI, there is a substitution between outward FDI and domestic employment, because the major trigger of moving abroad is the intention to reap benefits of the market opportunities abroad and use the economies of scale effect.

The question whether outward FDI substitutes or complements domestic employment has been discussed in many empirical studies, which can be divided into two major groups. The first group finds that there is a substitution relationship between outward FDI and domestic employment. Substitution occurs between countries with comparable factor endowments (Braconier and Ekholm 2001, Cuyvers, 2005; Hanson, 2005). The second group of empirical studies has concluded that outward FDI has a complementary effect on domestic employment. Outward FDI could increase an enterprise's competitiveness, promote its use of economies of scale, and reduce its cost, which can lead to an increase in domestic employment (Lipsey, 2000; Mariptti, 2003; Becker, 2005; Masso, Varblane and Vahter, 2009).

Though the above scholars draw different conclusions by analyzing the relations between outward FDI and domestic employment, the impact of outward FDI on non-multinational enterprises has been neglected. However, the effects of outward FDI may well be extended to other, non-multinational enterprises. On the one hand, local suppliers could suffer as they are replaced by foreign suppliers. On the other hand, to invest abroad may create new demand for intermediate inputs from domestic producers. So to study the impact on all the enterprises no matter whether they have invested abroad or not will be more accurate (Markusen, 2004; Koings and Murphy, 2006; Federico and Minerva, 2008). Since the analysis of the employment effect of outward FDI has produced mixed results, we will use Cobb-Douglas production model to analyze the relations between China's outward FDI and domestic employment in the following.

III. THEORETICAL MODEL

In this article, the improvement of Cobb-Douglas production model is used as follows.

$$Q_{it} = A_i^{\gamma} K_{it}^{\alpha} L_{it}^{\beta}$$

$$i = 1, \cdots, N \quad t = 1, \cdots, T \qquad (1)$$

Where Q_{it} stands for GDP, K_{it} and L_{it} denote the capital input and labor input respectively, α, β denote the output elasticity coefficients of capital and labor respectively. A_i is the efficiency index of output growth and γ is the share of factors which affect domestic output. The marginal products of labor and capital are wage ω and cost c respectively. Assume the market is perfectly competitive, we use the profit-maximizing equilibrium condition and get the following.

$$Q_{it} = A_i^{\gamma} \left[\frac{\alpha L_{it}}{\beta} * \frac{\omega_{it}}{c} \right]^{\alpha} L_{it}^{\beta} \qquad (2)$$

Take logarithms we can get

$$LnL_{it} = -\frac{\alpha}{\alpha+\beta}(Ln\alpha - Ln\beta) -$$

$$\frac{\alpha}{\alpha+\beta} Ln(\omega_{it}/c) + \frac{1}{\alpha+\beta} LnQ_{it} - \frac{\gamma}{\alpha+\beta} LnA_i$$

(3)

Let $\psi_0 = -(\alpha Ln\alpha - \alpha Ln\beta)/(\alpha+\beta)$,

$\psi_1 = -\alpha/(\alpha+\beta)$, $\psi_2 = 1/(\alpha+\beta)$,

$\psi_3 = -\gamma/(\alpha+\beta)$. Then (2) can be written as the following

$$LnL_{it} = \psi_0 + \psi_1 Ln(\omega_{it}/c) + \psi_2 LnQ_{it} + \psi_3 LnA_i$$

(4)

According to the basic labor demand equation, outward FDI affects domestic labor demand through three channels. First, outward FDI can increase or reduce domestic output directly; second, the change of outward FDI has effect on domestic production efficiency; third, outward FDI can affect the price elasticity of production factors. The second channel will be systematically analyzed in this article, because of the following two aspects. On the one hand, the continuous development of foreign investment produces reverse technology spillover effect (or crowding out effect) on domestic enterprises, which will influence the improvement of productivity; on the other hand, the development of outward FDI has effect on the import and export trade, no matter whether the effect is substitution or complementary. The efficiency index A can be expressed as follows.

$$A_i^{\gamma} = e^{\lambda T_i} FDIS_{it}^{\delta} \quad (5)$$

Where λ , $\delta > 0$, $FDIS$ is defined as the ration between the flow of outward FDI and domestic output, T_i is the timing variables, then the demand equation becomes

$$LnL_{it} = \psi_{0i} + \psi_{1i} Ln(\omega_{it}/c_{it}) + \psi_{2i} LnQ_{it} + \lambda_i T_i + \delta_i FDIS_{it}$$

(6)

Taking account of the dynamic adjustment process of labor demand, the structural change of outward FDI and the time delayed impact on domestic employment, we would add lagged variables into the function. Because the accurate measurement of capital price is very difficult and timing variables such as inflation, macro-policy adjustment will not change with industrial character, the impact of these variables will be neglected. Now the demand function becomes

$$LnL_{it} = \beta_0 + \sum_j \beta_{1j} Ln\omega_{it-j} + \sum_j \beta_{2j} LnQ_{it-j} + \sum_j \beta_{3j} LnFDIS_{it-j} + \sum_j \beta_{4j} LnL_{it-j} + \lambda_i + v_i + u_{it}$$

(7)

Where λ_i and v_i denote fixed effect and dummy variables respectively. Random error u_{it} is independent identically distributed.

IV. EMPIRICAL RESEARCH

Taking account of the availability of data, we choose the panel data of 25 provinces from 2003 to 2009. The data of outward FDI comes from "China's Outward Foreign Direct Investment Statistical Bulletin". To make the results more accurate, we will convert figures in dollar terms into RMB according to the exchange rate. The data of exchange rate comes from People's Bank of China. Other data comes from "CEI Net" and "China Statistical Yearbook". Considering that preliminary outward FDI, wages and domestic output also have effects on current labor demand and the time span of our research is not large, we will use first-lagged variables. The model is as follows.

$$LnL_{it} = \beta_0 + \beta_{10} Ln\omega_{it} + \beta_{11} Ln\omega_{it-1} + \beta_{20} LnQ_{it} + \beta_{21} LnQ_{it-1} + \beta_{30} LnFDIS_{it} + \beta_{31} LnFDIS_{it-1} + \beta_{41} LnL_{it-1} + \lambda_i + v_i + u_{it}$$

(8)

Because explanatory variables are related to random errors (explanatory variables contain the lag of explained variables in the models) and there are perhaps endogeneity of the variables, the result will be biased if we use random-effect

models or fixed-effect models. Differential GMM estimation can effectively overcome the problems of endogeneity and heteroscedasticity , but it can cause the lack of some sample information and reduce the effectiveness of instrumental variables. We will use System-GMM estimation (Arellano and Bover, 1995; Blunder and Bond, 1998). System GMM estimation includes the information of differential and horizontal equations, so it is more efficient than differential GMM estimation. We will select the Sargan statistic to test the effectiveness of instrumental variables, and use AR (1) and AR (2) to test the serial correlation of residual term. System GMM estimation could be classified into one-step GMM estimation and two-step GMM estimation. The latter will be selected, because it is less susceptible to the interruption of heteroskedasticity.

TABLE1 Results of System GMM

variables	coefficient	std. error	t-value	Prob.
LnL_{it-1}	-0.857	0.081	-8.258^{***}	0.000
$Ln\omega_{it}$	-0.069	0.061	-2.454^{**}	0.047
$Ln\omega_{it-1}$	0.195	0.076	3.864^{**}	0.031
LnQ_{it}	0.564	0.061	4.007^{***}	0.001
LnQ_{it-1}	0.866	0.013	8.918^{***}	0.000
$LnFDIS_{it}$	0.026	0.028	0.943	0.232
$LnFDIS_{it-1}$	0.015	0.031	0.501	0.568
AR(1)	-2.361^{**}		0.024	
AR(2)	0.537		0.789	
Sargen test _P	0.834			

Note: The table reports the results of System GMM regressions using Stata 10.0. The dependent variable is domestic R&D investment. Instrumental variables are two-period lagged L, Q, FDIS and ω. *,**,*,denote significant at 1, 5, 10 percent level respectively.**

In the models, the Sargan statistic is not significant, which means the choice of instrumental variables is effective. The test of AR (1) rejects the null hypothesis, but the test of AR (2) accepts the null hypothesis, which means the residual sequence is not correlated. The model shows that the last issue of domestic employment has significantly positive effects on current labor demand. Both domestic output and the last issue of wage income can effectively improve domestic employment, but current wage income has significant substitute effect on domestic employment. So the continuous improvement of productivity and the long-time growth of GDP can effectively promote domestic employment. Though the increase of wage income can reduce current employment opportunities, it will help to increase domestic employment rate in the future.

From table 1, we can find the impact of outward FDI on domestic employment is not significant. Possible reasons are as following. First, as a developing country, China has large enough market. Enterprises are easy to form development strategies dominated by domestic market, which leads to less pressure and motivation to expand overseas market by invest abroad. So many enterprises concentrate on domestic market, and pay less attention on external market. Second, domestic enterprises are unable to make full use of the location advantages of host countries and lack of Ownership Advantage and Internalization Advantage, which has caused low technical content of outward investment. To reflect a country's outward FDI level, UNCTAD has defined the performance index of outward FDI--OND. OND can be defined as the following.

$$OND_i = \frac{FDI_i / FDI_W}{GDP_i / GDP_W} \tag{9}$$

Where FDI_i refers to the FDI outflows of country i, and FDI_W stands for the FDI outflows of the world. The average OND of China is below 0.3 from 2003 to 2009, which is under the average level of the world. On this point, there is a still a big gap between China and industrial countries, though we have made rapid progress in the development of outward FDI.

V. CONCLUSIONS

This paper examines the effect of outward

FDI on domestic employment. Empirical research shows that though China's outward FDI has a positive effect on domestic employment, the effect is not significant. Domestic enterprises are lack of Ownership Advantage and Internalization Advantage, and unable to make full use of domestic and international markets and resources to enhance the economic growth efficiency. With the acceleration of global economic integration and rapid development of information technology, economic relations between countries are getting closer and closer. The strategy of "going out" will become more and more important on domestic employment.

Based on the above analysis, we provide the following suggestions. First, speed up the investment in developed countries. When investing in developed countries, domestic enterprises can make full use of the location advantage of larger markets and economies of scale to promote the upgrade of industrial structure. Second, optimize the industrial structure of outward FDI. To enhance technical content of outward FDI, the improvement of R&D capacity and international competitiveness of domestic enterprises becomes necessary. Third, to domestic enterprises, when investing abroad, never neglect the consolidation of domestic market. Only in this way, can the sustainable growth of outward FDI be guaranteed.

REFERENCES

[1] Arellaon, M. and O. Bover, "Another Look at the Instrumental Variable Estimation of Error-Components Models," Journal of Economics, Vol. 68, pp. 29-51, 1995.

[2] Becker, S., K. Ekholm, R. Jäckle, and M. A. Muendler, "Location Choice and Employment Decisions: A Comparison of German and Swedish Multinationals," Review of World Economics, Weltwirtschaftliches Archiv , vol. 141, pp. 693–731, 2005.

[3] Blundell, R. and S. Bond, "Initial Conditions and Moment Conditions in Dynamic Panel Data Models," Journal of Econometrics, vol. 87, pp. 115-143, 1998.

[4] Braconier, H. and K. Ekholm, "Foreign Direct Investment in Central and Eastern Europe: Employment Effects in the EU," CEPR Working Paper No. 3052, 2001.

[5] Cuyvers, L., Dumont, M., Rayp, G. and K. Stevens, "Home Employment Effects of EU Firms' Activities in Central and Eastern European Countries," Open Economies Review, vol. 16, pp. 153-174, 2005.

[6] Ekholm, K. and J. Markusen, "Foreign Direct Investment and EU – CEE Integration," background paper for the conference Danish and International Economic Policy, University of Copenhagen Copenhagen, May 2002.

[7] Federico, S. and A. Gaetano (2008), "Outward FDI and Local Employment Growth in Italy," Review of World Economics, vol.144, pp. 295-324, 2008.

[8] Hansson, P., "Skill Upgrading and Production Transfer within Swedish Multinationals in the 1990s," Scaninavian Journal of Economics, vol. 107, pp. 673–692, 2005.

[9] Kokko, A., "The Home Country Effects of FDI in Developed Economies", The European Institute of Japanese Studies, EIJS Working Paper No. 225, 2006.

[10] Konings, J. and A. Murphy, "Do Multinational Enterprises Relocate Employment to Low Wage Regions? Evidence from European Multinationals," Review of World Economics, vol. 142, pp. 267-286, 2003.

[11] Konings, J., and A. Murphy, "Do Multinational Enterprises Relocate Employment to Low-Wage Regions? Evidence from European Multinationals," Review of World Economic, Weltwirtschaftliches Archiv, vol.142, pp. 267–286, 2006.

[12] Lipsey, R. E., Ramstetter E. and M. Blomstrom. "Outward FDI and Parent Exports and Employment: Japan, the United States and Sweden," Global Economic Quartly, Vol. 4, pp. 285-302, 2000.

[13] Mariotti, S. Mutinelli M. and L. Piscitello, "Home Country Employment and Foreign Direct Investment: Evidence from the Italian Case," vol. 27, pp. 419-431, 2003.

[14] Markusen, J.R., "Multinational firms and the theory of international trade," Cambridge, MA: MIT Press, 2004.

[15] Masso, J., and P. Vahter, "Technological innovation and productivity in late-transition Estonia: econometric evidence from innovation surveys," The European Journal of Development Research, vol. 61, pp. 240-261, 2009.

[16] Masso, J., Varblane, U. and P. Vahter, "The Impact of Outward FDI on Home-Country Employment In a Low Cost Transition Economy," Tartu University Press. ISSN 1406-5967, 2009.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Inventory-based Resource Management in Cloud Computing

Qinlong Jiang, Weibing Feng, Junjie Peng,
Fangfang Han, Qing Li, Wu Zhang
School of computer engineering and science
Shanghai university
Shanghai, China
e-mail: qinlong_jiang@shu.edu.cn

Haitao Su
Beijing Benz-DaimlerChrysler Automotive Co., Ltd.
Shanghai, China

***Abstract*—As a new application model, cloud computing enables users to distribute virtual resource in cloud data center as demand. However, the performance declines as the increase of users. This paper focuses on the network and I/O performance decline due to the virtual machine image and virtual disk image densely deployment in Cloud computing management platform. The method and model for rapid deployment and resource share are proposed on basis of shared storage and inventory theory which is widely used in commodity market. Experiment shows the method can effectively alleviate the performance decline caused by users' dense image requests in cloud computing management platform.**

Keywords-inventory; resource management; cloud computing; virtualization

I. INTRODUCTION

Cloud computing is a new technology development of from grid computing distributed computing, utility computing and virtualization. It provides services through the huge amount of computing, storage, network, and software resources in cloud data center to users. Users are able to use the virtual resources located in remote data center on demand through simple terminals and Internet. And the cloud data center guarantees to provide resources users apply for [1].

Cloud computing is featured by rapid deployment, instant response to users' requests, dynamic scalability and pay-as-you-go, etc. These enable users to get the convenient and economical cloud services. However, as the users are enjoying the convenience, cloud computing service providers will meet the challenges as requests and applications increased. The decline of performance of CPU, memory, and I/O overload will influence the quality of services. These are all the bottlenecks for high quality cloud service [2].

Cloud computing management platform is the foundation of cloud computing. It provides users with on-demand and rapid deployment virtual machine by virtualization technology [3]. Its performance matters a lot to the overall service. Many researchers studied the performance optimization on CPU, memory, I/O when the VMs are on. [4][5][6] studied load balancing, trying to deploy the VMs in the physical servers uniformly. While as usual, before the start of VMs, it takes some time to deploy the cloned VM images from the image template repository to specified location of the VM running related path. As the increase of VMs' requests, the clones of these images will strive for network and I/O resources. This will result in the increase of deploying time.

Take the open source cloud computing management platform OpenNebula as example. It implements two kinds of image storage structure: Network File System (NFS) and none-NFS [7].

As to NFS, it shares each node's VM running related images storage path by NFS server. It is good for the images management and share and is the precondition of virtual machine migration in OpenNebula. VMs clone the VM image template into the shared path from the image template repository before start. In None-NFS structure, the related images are stored in the local disks. They are not shared by other nodes.VM image template are cloned from the template repository to the locations where the reference locations for VM's running before its start.

No matter using NFS or None-NFS, when the requests become frequent, the deployment of VM image becomes quite frequent and dense. In NFS, it strives for the network and I/O resource between the template repository and VM running shared path. In none-NFS, it influences the network bandwidth of template repository and each cluster node. NFS gets the advantages in collective management while none-NFS is inefficient in the dense transmission of images. Take the above in consideration, using NFS as basic structure, an inventory based resource sharing and rapid deployment method and resource management model are proposed. By using our model and method, cloud data center gets well prepared before the users' application for the images. So users can use it just by locating the prepared image. This will alleviate the performance decline problem above stated.

The remainder of this paper is organized as follows: in Section Ⅱ, we present the design of image inventory. In Section Ⅲ, we describe inventory model of cloud resource. Section Ⅳ shows the experiment and analysis. Finally the paper concludes with a summary of our findings and highlighting our plans for the immediate future.

II. THE DESIGN OF IMAGE INVENTORY

A. Architecture of Image Inventory

Fig. 1 is the architecture of image inventory based on NFS and inventory concept. Similar to the NFS structure in OpenNebula, the storage paths of running VMs in each

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.54

cluster node are shared through NFS. To solve the performance decline when the requests for images are dense and frequent, we add an image inventory to the OpenNebula shared storage structure. That is done before the requests. We pre-store the images into the inventory. When the request comes, all it has to do is just locating the image in the inventory. (Note that, the VM running path is same as the image inventory path.)

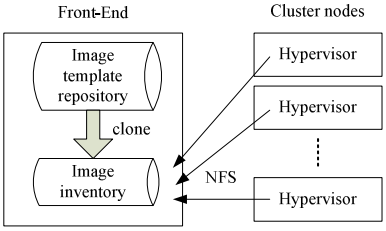


Fig. 1. Architecture of image inventory

B. Definitions and Concepts

1) Image: image is the storage supporter of virtual machines. In this paper, we define two kinds of images: OS and DATABLOCK. OS is images with pre-installation of operation systems. Each VM has to have only one OS image. DATABLOCK is for the virtual disk of the VM to storage users' data. Each VM can get no or multiple copies of this image. DATABLOCK can be set to different sizes as demand (5 GB, 10 GB, etc.).

2) Image template repository: it is used to store image template. We usually store the OS images pre-installed with different OSs(ubuntu10.10、redhat9、window 7, etc.) into image template repository. When users request for the deployment of VMs, according image will be cloned from image template repository to the location of running VMs. Here, the comparatively shorter image cloning time accelerates the VM deployment process compared to the long system installation time. Similarly, DATABLOCK is also stored in the image template repository.

3) Image inventory: is an available space in NFS server which is used for image resource pool. It is the place for storing user needed images. When users request for image resource, just locating the corresponding image in the image inventory. Prepared the images users may ask for in a short while in the image inventory can reduce the consumed time, can alleviate the performance decline problem mentioned above. Note that the image inventory path is the same as the VM running path in NFS. In order to distinguish the images, we identify them into different states.

4) Shortage queue: is a FIFO queue. When users request image is null in the inventory, it is pushed into this queue. When users request image is not in the inventory, it is pushed into this queue according to chronological order.

5) States of the images in inventory: there are three kinds of states of images in the inventory: unallocated, allocated unused and allocated used. The newly cloned image from the image template repository is unallocated. When image is allocated to users, it is the second state. When users use the applied image, it turns to be the last state.

C. Inventory Control Module

The inventory control module is the core of the image inventory design.

First and foremost, inventory control module is responsible for the periodical supply of images from the image template repository to the image inventory. At the beginning of a period, this module gets the log of image usage and shortage. Then calculate the proper inventory amount based on the logs and inventory model this paper puts forward. Last, it controls the cloning of images from template repository to inventory.

Secondly, inventory control module processes users' requests for images. When users apply for images, this module will check whether the image is in the inventory. If it is in the inventory, locate the image; or push the applying information into the shortage queue.

Then, read the top of shortage queue when the queue is not empty. And clone the images into inventory. Delete the top data of the queue after cloning. The module processes the shortage by the order of application which leases the performance decline.

Lastly, the module will call back the relevant images when users cancel the request to save storage space.

D. Inventory Operation Process

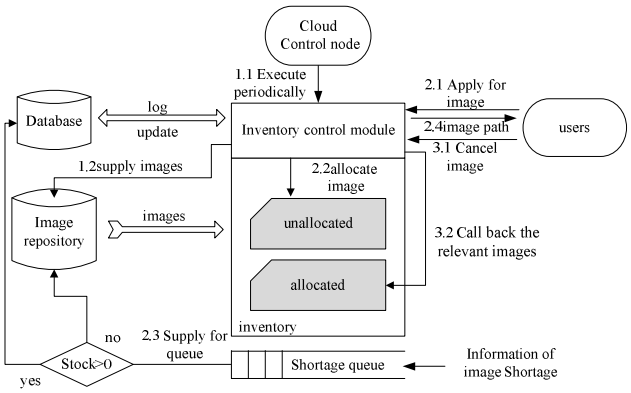


Fig.2 Flowchart of inventory control

- Step 1.1: add control function in the cloud control node to run inventory control module periodically. The module gets users' request for each kind of images and historic using log from database. It uses the proposed model to analyze the demanding amount of images.
- Step 1.2: clone the demanding amount of images to inventory. Store the new image path into database and set the new image state as unallocated.
- Step 2.1: users ask for images to the inventory control module. The module checks out the inventory log in database. If there is unallocated image, go to step 2.2; otherwise, put this requested

image information (including its name, type and user ID) into the shortage queue for the analysis of the model. Then go to step 2.3 to replenish.

- Step 2.2: allocate demanding images to users and set their state into allocated unused.
- Step 2.3: replenish the shortage queue. Get the head the queue, query the inventory according to the image name and type in the top of the queue. If the image exists in the inventory, allocate to the user, set the state to allocated unused; otherwise, clone it, and store the path in inventory into database, allocate to user, and set state to allocated unused.
- Step 2.4: after step 2.2 or 2.4, return image path to user.
- Step 3.1: users cancel all the possessed images, including images in all states.
- Step 3.2: inventory control module withdraw images users have cancelled. Then these images are available for other users. The callback of allocated used image is beneficial to save inventory space and allocated unused can lease workload causes by cloning images.

III. INVENTORY MODEL OF CLOUD RESOURCE

This sector depicts the inventory model for the inventory control module's periodical replenishment combining the classical inventory theory. This model is responsible for the prediction and analysis of the images needed by the replenishment process.

In this model, S depicts the maximum inventory volume of NFS server. For convenience, we set V as the same volume of each image. k= {1, 2,…K} represents the different types of images in the inventory. Vector q= ($q_1, q_2, \ldots q_K$) is the demanding image amount in each period. Vector r= ($r_1, r_2, \ldots, r_K$) is left image amount after last period.

Single exponential smoothing [9] is used to calculate the demanding amount F_t^k of k in the next period t.

However, if the predicted amount exceeds the real quantity, surplus images will be stored. This will cause the waste of NFS space and may cause the shortage storage of other images. Otherwise, some request cannot be found in inventory as they are put into the shortage queue. This will take more time for the request. To avoid above problems, the needed amount of image k ranges among these two thresholds: $Q_{\min}^k, Q_{\max}^k$.

According to the commercial market analysis, we know image use accord with the normal distribution in each period. According to the theory of normal distribution, we set $Q_{\min}^k = \bar{x}_k$ and $Q_{\max}^k = \bar{x}_k + SS_k$. Here, $\bar{x}_k$ is the mean using quantity of image k each period and SS_k is the secure inventory of image k.

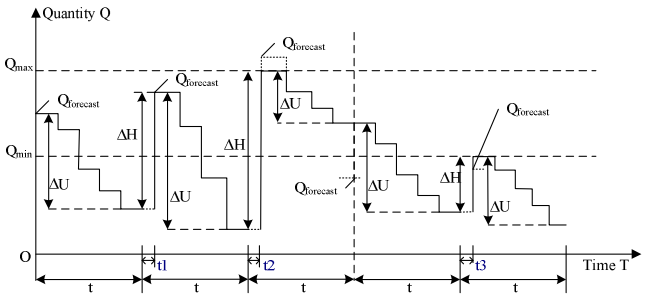


Fig.3 image inventory model

Based on the analysis, Fig.3 depicts the image inventory model. In Fig.3, abscissa values represent the time period. The ordinate represents image inventory quantity during a period. And t is periodical time unit (it depends on circumstances, usually set as 1 day or hour). Here, t1, t2, t3 are the time for replenishment. ΔU is the inventory quantity during a period. ΔH is the actual replenishing amount next period. $Q_{forecast}$ is the predicted amount of next period demanding inventory. $Q_{\min}$ is minimum demanding image amount next period. $Q_{\max}$ is the maximum demanding image amount next period. In Fig.3, in each period, inventory quantity declines as time goes by. At the initial stage of a period, after the replenishment, the new inventory quantity is between $Q_{\min}$ and $Q_{\max}$.

After the analysis of this model, we propose the demanding quantity of image k in next period as

$$q_k = \begin{cases} r_k, Q_{\min}^k < r_k < Q_{\max}^k, F_t^k < r_k \\ F_t^k, Q_{\min}^k < F_t^k < Q_{\max}^k \\ Q_{\min}^k, F_t^k < Q_{\min}^k \\ Q_{\max}^k, Q_{\max}^k < F_t^k \end{cases}$$

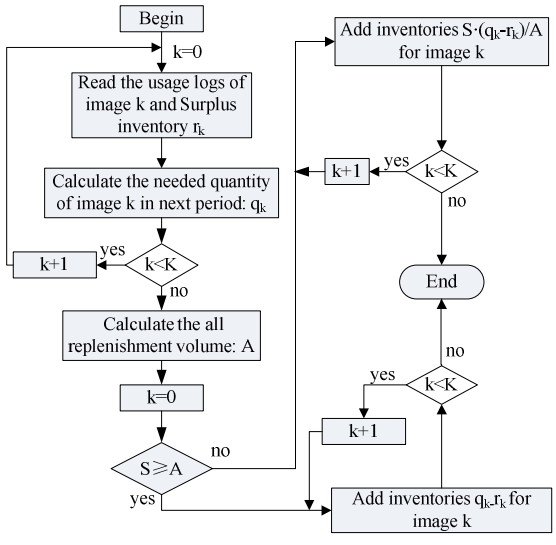


Fig.4 Computation flow chart of inventory model

By computation of the equation, all the replenishment volume is $A = V \bullet \sum_{k=1}^{K} (q_k - r_k)$. The inventory model computation flow is given in Fig.4.

IV. Experiment and Analysis

In order to verify the efficiency and validation of proposed model and method, this section will conduct the comparison and analysis of the original and modified image storage structure.

We use OpenNebula as experimental platform, modifying the shared storage structure when requests are dense. Simulations are adopted to imitate OpenNebula's allocation of images. The mean spending time is computed of different types of images request interval in some period. First, we choose NFS server whose CPU frequency is 2.66GHz, disk volume is 500GB, disk I/O reading speed is 80MB/sec, memory is 2GB.Then install CentOS 5.5(64 bits). In the NFS server, assign an image template repository path and store a 5GB image as template and set the VM running storage path. Run the simulation program, then the image request arrives, simulation program will clone the corresponding image from the template repository to VM running path, and record the time for requesting the image. After computation, we get three groups of data as in Table.1 10min, 20min, and 30min. they are the mean request spending time of the time duration 10mins, 20mins and 30mins. As can be seen from the table, at some specific period, the consumed time increases with the increase of the requests. The same image request in different period, as the shorten of time, the mean request spending time also increases. It can be seen that the original OpenNebula shared storage structure has severe performance decline problem when the image requests are dense.

To verify whether the method this paper proposed can effectively alleviate the problem or not, we did a simulation. We choose the same PC as NFS server, also assign the image template repository path, store a 5GB image template and set the image inventory path. We clone 50 images to the inventory and record each images' path to MYSQL database in advance. When the request arrives, it can use the image by reading the image path in database. Through computation, we also get the 3 groups data: 10'min, 20'min, and 30'min. From the table, we can see that by using inventory, the rapid response can be achieved. The time for image request is not affected greatly by the increase of requests.

TABLE.1 TABLE OF EXPERMENT RESULT

	Traditional NFS Response(s)			Inventory used Response(s)		
request	30min	20min	10min	30'min	20'min	10'min
4	185	262	858	0.0131	0.0131	0.0135
8	199	331	1544	0.0130	0.0131	0.0135
12	387	748	2047	0.0132	0.0132	0.0137
16	662	1627	2679	0.0134	0.0131	0.0139
20	1372	1960	3223	0.0134	0.0133	0.0137
24	2564	2985	3844	0.0134	0.0136	0.0140
28	3391	3482	4606	0.0135	0.0135	0.0141
32	4076	4228	5454	0.0135	0.0136	0.0140
36	4923	5431	6285	0.0134	0.0138	0.0140
40	5533	5948	7331	0.0136	0.0136	0.0144
44	6155	6480	8430	0.0135	0.0137	0.0149
48	7009	7765	9955	0.0135	0.0140	0.0148

From the experiment, the conclusion that proposed method can respond rapidly to the request and alleviate the performance decline problem effectively can be drawn.

V. Conclusion and Future Work

This paper focuses on the study of I/O and network performance decline problem of VM image and virtual disk dense deployment. A resource sharing and rapid deployment method based on the widely used inventory concept in commercial market is proposed and the corresponding model is constructed. The experiment result shows the method can effectively alleviate the above stated problem. In future, we will work on promoting the stability of the inventory model, and consummating it to meet higher market demands and test.

Acknowledgment

This work was supported by the Key Project of Science and Technology Commission of Shanghai Municipality (No. 10510500600), Shanghai Leading Academic Discipline Project (No. J50103), the Key Laboratory of Computer System and Architecture (Institute of Computing Technology, Chinese Academy of Sciences), the Ph.D. Programs Fund of Ministry of Education of China (No. 200802800007) and Innovation Program of Shanghai Municipal Education Commission under grant (No.11YZ09).

References

[1] Foster, I., Yong Zhao, Raicu, I., Lu, S., et al., Cloud Computing and Grid Computing 360-Degree Compared, Grid Computing Environments Workshop, 2008, 1-10

[2] Michael Armbrust, Armando Fox, et al., A view of cloud computing, Communications of the ACM, 2010, Volume 53 Issue 4, 50-58

[3] Paul Barham, Boris Dragovic, Keir Fraser, et al., Xen and the art of virtualization, ACM SIGOPS Operating Systems Review - SOSP '03, 2003, Volume 37 Issue 5, 164 - 177

[4] Norman W. Paton, Marcelo A. T. de Arag˜ao, Kevin Lee, et al., Optimizing Utility in Cloud Computing through Autonomic Workload Execution, Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2009, 51-58

[5] Barath Raghavan, Kashi Vishwanath, Sriram Ramabhadran, et al., Cloud Control with Distributed Rate Limiting, ACM SIGCOMM Computer Communication Review, 2007, Volume 37 Issue 4, 337-348

[6] Ping Wang, Wei Huang, Carlos A. Varela, Impact of Virtual Machine Granularity on Cloud Computing Workloads Performance, Grid Computing (GRID), 2010 11th IEEE/ACM International Conference, 2010, 393-400

[7] OpenNebula, http://opennebula.org/documentation:archives:rel2.0:img_guide

[8] Ronald S.Tibben- Lembke, Yehuda Bassok, An inventory model for delayed customization: A hybrid approach. European Journal of Operational Research 165, 2005, 748-764

[9] Robert P.Trueblood and John N.lovett.JR, Data Mining and Statistical Analysis using SQL, Springer, Springer-Verlag, 2001

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

A QOS BASED LOAD BALANCING FRAMEWORK FOR LARGE SCALE ELASTIC DISTRIBUTED SYSTEMS

V.H. Nguyen, S. Khaddaj, A. Hoppe, Eric Oppong
Faculty of Computing, Information Systems and Mathematics, Kingston University,
London, Kingston upon Thames, KT1 2EE, UK
v.nguyen@kingston.ac.uk, s.khaddaj@kingston.ac.uk
a.hoppe@kingston.ac.uk, k0212653@kingston.ac.uk

***Abstract* - The emergence of grid and cloud computing require load balancers to deal with potential problems, such as high level of scalability and heterogeneity of computing resources. In this paper, we present a generic load balancing framework which separates allocating process and migrating process while preserving a guaranteed level of service. Based on this framework, an intelligent load balancer that is aware of multiple Quality of Services and directed by users' requirements is proposed. The load balancer aims to deal with elastic heterogeneous distributed computing environments.**

Keywords: Quality of Service, Scheduling, Distributed Architectures, Intelligent Decision Making

1. INTRODUCTION

Recent years have witnessed the emergence of new distributed computing systems such as grid computing and cloud computing. These systems are no longer limited by their geographical locations due to the rapid growth of the Internet. They are composed of a large number of resources and provide high performance computing capacity to a variety of applications with different resource demands. However, multiple Quality of Services (QoS) awareness and heterogeneity become important requirements of these systems [2][13].

Guaranteeing multiple QoS is a complex issue in large-scale heterogeneous distributed computing systems. QoS parameters are contained in Service Level Agreements between users and resource providers to ensure that user requirements can be satisfied. However, measuring, monitoring and reporting on QoS aspects of resources is based on users' experiences or users' ability to consume resources. Resources provided by different providers give different level of QoS which may not be measured in advance. The development of virtualization resources complicates the problem since two "identical" resources may result in different performances [6]. In addition, resources provision and the precision of the prediction of dynamical changes of resources are still challenging tasks particularly in term of load balancing and scheduling.

A majority of load balancers and schedulers are system-centric and hence do not meet a wide range of user requirements. In conventional systems, the decreasing makespans and increasing throughput are two common objectives of which eleven heuristics algorithms are compared by Braun in [3]. Other QoS parameters such as reliability have been also considered [5]. Reference [10] proposed a scheduling algorithm which minimizes the total cost of services while keeping user requirements.

To address the above issues, an intelligent load balancer framework, which optimizes the resources from providers while maximizing the satisfaction of diverse requirements of users, is proposed in this paper. The load balancer is developed to adapt the dynamically scalable characteristic of the large-scale, elastic distributed computing system. It also aims to autonomously balance the potential conflicts between requirements of the tasks as well as of users.

We start by discussing a framework of an intelligent load balancer that is driven by user needs. This follows by proposing a load balancer for large scale systems which is aware of multiple quality of services. A new novel model of resource management in load balancer is proposed. A number of simulations of a distributed computing system with different load balancers are conducted. Finally, we conclude with some suggestions for future work.

2. A GENERIC INTELLIGENT LOAD BALANCER DRIVEN BY USER REQUIREMENTS

2.1 A Generic Intelligent Load Balancer framework

Intelligence, by its definition [11], is the integration of perception, reason, emotion and behaviour in a sensing, perceiving, knowing, caring, planning and acting system that can succeed in achieving its ultimate goals in the world. An intelligent load balancer must be able to sense the resources, to interact with users, to analyse the system as the unified, solid collection of resources, and to adjust its decision making process to act in respond to the change of the systems.

Inspired by above concept, we divide the components of an intelligent load balancer into two parts (Figure 1). The first part is information collection, which provides the values of resource requirements of tasks, of resource attributes, and of the objectives of providers, users and vendors. These values are the references of the decision making processes which allocate resources and give the negotiation conditions of requested tasks and of requisite resources.

In conventional systems, the decision maker in a load balancer has two basic functionalities: allocating resources for requested tasks and migrating tasks from overloaded resources or mal functional resources to other resources. The objectives and the requirements of these functionalities are different. The first one can be either system-centric or user-centric or both in trying to optimize both user expectation and resource utilization. However, the second one aims to keep system balance amid the

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.12

dynamical changes of the resources, and hence is more system-centric.

The first component in the decision maker, Directing Load Balancer, allocates resources for the requested tasks. This is often considered as a global load balancer or scheduler. The input of the Directing Load Balancer is the requested task; and the output is the resource which is going to execute the task. In the negotiation supporting system, the output of the Directing Load Balancer is also the negotiation condition in case there is no resource satisfying the requirements. In multiple QoS awareness system, the decision of a Directing Load Balancer refers to observed values of attributes of resources and objectives of users, providers and system. This may not be the best optimized solution since task scheduling is known as an NP-hard problem. Moreover, its perceived information of resources is depended on the collected information which may be insufficient or inaccurate. In large scale distributed computing system, the directing load balancer has to solve a dilemma problem: On one hand it needs to use all possible QoS values; on the other hand, it has to reduce the number of these values due to the constraint of the load balancing time.

To optimize the workload distribution of the system and to adapt the dynamically change of the resources, the system needs a second component, a Migrating Load Balancer, which equilibrates workloads among the resources during runtime. The input of the Migrating Load Balancer is the overloaded resources and tasks in these resources; and the output is the destination resource to which the tasks are going be migrated. One of the references of the Migrating Load Balancer is the information given by Migrating Network Database. Previous studies shows that job migration is generally expensive in most distributed computing systems [5]. The cost of task migration is not only the time to transfer data among resources in the network but also the impact of network communication as well as potential violation in QoS constraints. The transfer cost is increased if the network topology distance is large. The migrated tasks are hence allocated at the resources which are near to the original resource in the network. Therefore, a Migrating Load Balancer plays a role as a local load balancer that requires information of local resources.

The advantages of cloud computing and virtualization technology motivate a new functionality of decision maker, which answer the questions when and how to shrink and expand resources on demand. Initial researches on this decision making process was studied [1][14]. Cloud systems such as EC2 of Amazon1 allow users to automatically add the requisite amount of resource. This is the base of resource federated systems, which use the third component, the Shrinking/Expanding Decision maker, to strengthen its capacity dynamically. New resources are added if current resources cannot meet the workload of the system. They are then removed when the workload of the systems decreases.

[1] http://aws.amazon.com/elasticloadbalancing/ (Last visit: October, 2010)

In Information Collection part, users, providers and venders interact with system through Human Controll Interface component. The provided knowledge are interpreted by QoS interpreter. These are resources requirements and references of objectives of the decision maker. The information are weighted by the role of the person who supplies them. For example, users provide the weight of each requirement in relative to the others [12]. Users, providers and vendors have different objectives, evaluations, and requirements of the resources. They feed and direct the system with their knowledge and expertise.

A QoS intepreter, which intepretes user abstract requirements into concrete QoS values, contains two components. The first component is Semantic Engine that tranfers requirements of user and express them in a machine processable language. For example, the QoS values of resource performance and the cost are related to each others. In user requirements, this relations are not explicitly expressed. In general, users submit upper values of the cost. The Semantic Engine will translate user requirements and the relations into QoS concreted values. In this example, the appropriate value of performance is calculated. The second component is QoS Profiling that associate the resource requirements to the tasks. This component is widely used in workflow scheduling [17]. In this case, the requirements of an application is transformed to the requirements of its component tasks under the QoS profiling.

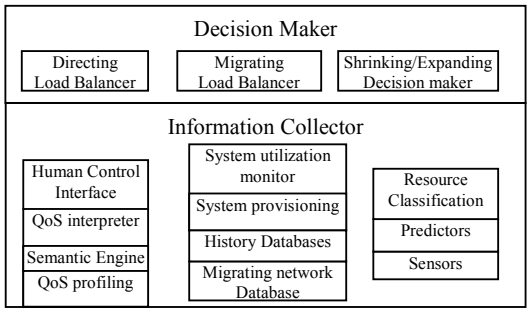


Figure 1. An Intelligent Load Balancing Framework based on Quality of Service.

From system side, QoS values of resources are determined not only by sensor but also by predictor or provisioning engine. The attributes of resources in heterogeneous distributed systems can be achieved by three methods: pre-estimating, sensing and provisioning. Although some resources attributes can be given by benchmarks, others dynamically change during runtime. For example, in resource-sharing systems, the performance of a resource depends on its workload. Besides, resource attributes such as reliability or availability cannot be obtained by only sensing. Provisioning those attributes is an essential requirement for enforcing QoSs. To collect attributes of resources, sensors are integrated to the resources. The collected information is processed to evaluate the capacity of the resources under the standard quality of service metrics. The accuracy of the information is limited by two factors. First, the load balancer may not have privilege authority

to gather necessary information. Second, to avoid the overhead of the system, the sensors can only send back the information periodically. This delay may cause significant impact to the precision of information. Resource provisioning, which consists of learning and predicting processes, is based on experience and knowledge domain of the load balancer. There are three approaches: simulation, analytical modelling, and historical statistical learning. In simulation, the set of tasks and resources is enumerated before its actual execution. In analytical modelling, QoS values are estimated by models such as fuzzy map and neuron network. In historical statistical learning, QoS values are estimated based on the history of the tasks and resources as well as the history of users and providers.

In distributed system management the system utilization monitor is an essential component. It gives not only the information of capacity of systems but also the overview of usage of resources. This information is particularly important in Shrinking/Expanding Decision maker. Although the system provisioning is now widely studied, but it is commonly believed that a good system provisioning engine improves significantly the decision maker.

Theoretically, history databases contains history of tasks, history of resources, history of users, history of providers and history of the whole system. The history of task and of users is used to predict the resource requirements of tasks at system levels. The history of resources is used to predict the QoS values of resources. Some of QoS values are availability and reliability. One example of the history of system is the trace of arriving tasks [8], which have complicated patents. Therefore, in practice, it is very important to select the characteristics to store in history database as well as the provisioning or predicting engine.

2.2 Load balancer based on multiple Quality of Services

As a first step in the construction of the framework we are going to study the decision making process in Directing Load Balancer, which is aware of multiple quality of services. We assume that the sensors, predictor and QoS interpreter are able to provide the load balancer a vector of QoS values (QoS1,..., QoSn) as shown in Figure 2. The problem is how the load balancer uses the information to allocate resources and to give the negotiating condition in the case when there are no resources satisfying requirements.

In large scale distributed system, the number of QoS parameters and the number of resources are the most important factors affecting the performance of load balancers and schedulers. There are several solutions based on the deployed models and structures of the management system. Tree based model and agent model are the most popular models due to their extendibility. Tree based models, or hierarchical structures, are widely used in centralized or hybrid systems, for example in multi-site grid computing systems [15]. Agent based models are applied in decentralized system where the load balancer has less permission to control of the resources [4]. We are going to introduce a new model that dynamically categorizes the resources based on multiple quality of services.

For each QoS parameter, we classify the resources into different levels. At each level, resources may have the QoS values in a continuous interval. For example, the reliability of a resource can be high, medium and low, where high reliability means that the reliability of the resource is greater or equal to ninety percent; medium reliability means that the reliability of the resource is from seventy percent to ninety percent; and low reliability means that the reliability of the resource is less than seventy percent.

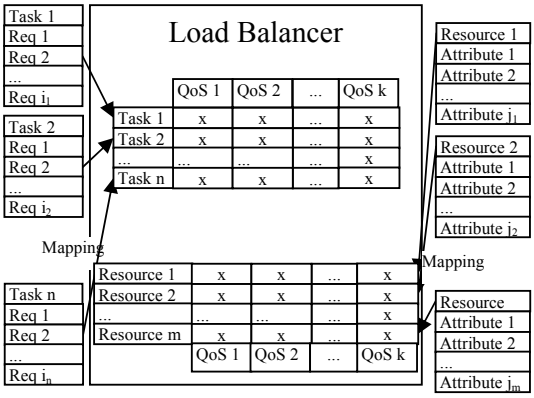


Figure 2. Reference Information inside load balancer based on QoS.

The classification of the resources is based on QoS due to two reasons. The first reason is that users often require the resources in abstract level and given upper and lower values of the requirements. The second reason is that the values of QoSs may vary within a range due to the dynamicity of the resources as well as the accuracy limitation of sensors and predictors.

In general, users are not able to specify their requirements of resources in details. Based on their knowledge and motivations, they submit the requirements of the tasks. However, the attributes of the resources are implicit and dynamically changed inside the system. Therefore, user requirements are often described in a highly abstract level. They are availability, capacity, performance, scalability, and reliability while the resource attributes are speed of processors, number of processors, memory size, disk size, and network bandwidth and so on, see Figure 3. The influences of the resource attributes on the QoS parameters vary, particularly since attributes of some resources might be incomparable. Different resource providers apply different standards to measure the attribute of their resources. This problem is more complicated in system consisting of both physical resources and virtual resources. For example, current virtualization technology allows splitting one physical server into several logical servers of which the functionalities and capability can be different.

2.2.1 User satisfaction

In a system with soft constraint requirements, we define the measurement of user satisfaction for each task as:

$$u_t = \sum_{i_t} w_{i_t} a_{t,i_t}$$

where wk is the weight of QoS parameter k in the view of user, and ai,,k is the user satisfaction under the QoS k.

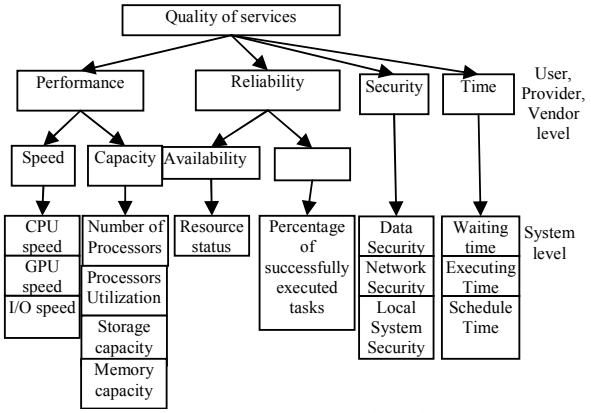


Figure 3. An AHP of QoS in Grid /Cloud Computing.

3. EXPERIMENTS

The aim of the experiments is to demonstrate that the proposed framework is capable of dynamically adapting in response to its environment. During the experiments, a number of simulations were conducted using a distributed computing system. We simulate the system based on discrete-event simulation model. The simulator is written in Java.

3.1 Simulation setting and evaluation measurement

The simulated distributed computing system consists of 100 nodes of six types, which are generated randomly. The resources are classified based on the reliability and performance aspect. We assume that resources of the each type have the same level QoS and that each resource has a buffer containing maximum 10 tasks and can only execute one task at a time, i.e the resources cannot be shared. The reliability of resources does not depend on the task, but the performance of the resources may vary for different tasks. The tasks are independent and non pre-emptive. And each resource contains no queues.

The tasks are created randomly with following parameters: arriving time, executing time on the slowest node. In the performance mapping, the execution time of a task running on medium machine is only one third that of the task running on slow node. Similarly, the execution time of a task running on fast node is only one third that of the task running on medium node.

The tasks are randomly generated. The number of arriving task per second is a Poisson variable which has the average value λ. To change the workload of the whole system, we change the value λ. When λ is increasing, there are more tasks requested, and hence the workload of the system is also increased. The execution time on the slowest node is a uniform variable in the interval [0; 150] (in unit time).

Users input their expectation of the tasks under two objectives: and resource performance, and reliability of the resource. These objectives from different users and tasks have different weights. We assume that tasks are submitted in together with user requirements.

In the first experiments, we consider only single quality of service parameter, the performance. Three load balancers are evaluated. The first load balancer is Round Robin denoted by L1. The second load balancer, L2, is based on the performance of the resources, i.e. it allocates the fastest available resource for the next task. The last load balancer, L3, classifies resources in three logical resource based on the performance of resources. Tasks requested a resource of type i are submitted to the resources of type i. In case, there is no resource of type i, the load balancer will look for a resource of the other types.

The second experiment aims to test the load balancers with soft constraints assumption. In this experiment, users also submit the weights of their requirements, i.e. users have to indentify the important level of each requirement. Ten tests were conducted. Tasks are requested during a period of time, 5000 minutes. In the first subtest, the average number of requested task per minute λ =3; in the second subtest, λ =3.5; and so on.

These load balancers are evaluated under two metrics: resource utilization, and user satisfaction, which indicates the level of requirement violation. When the requirements are satisfied, user satisfaction is one hundred percent. When the requirements are not satisfied, user satisfaction is the weighted average of satisfaction level of all requirements.

These load balancers were tested in the same environments, i.e. the same configuration set of tasks and resources. Each test is repeated 20 times. Thus the result of the experiment is the average of the results of 20 tests.

In the experiment, when the average number of task per minutes is less than 3.5, the number of rejected tasks under load balancer L2 is less than those under load balancers L1, and L3, see Figure 6. In this case, the executing time of task in fast resources is shorter, and hence the L2 are able to execute more tasks. However, when the average number of task per minutes is less than 3.5, the workload is heavier, the resource utilization is larger, and the slow resource is also used more frequently. Heavy task are allocated to the slow resources. This creates negative impact to load balancers L2. This is showed in the graph of resource utilization, see Figure 7. In case the average number of task per minutes is less than 3.5, the resource utilization with load balancer S1 is less than 90%. Concerning user satisfaction, since the load balancers S3 is designed to keep users requirements; its satisfaction is 100%. The other load balancers do not take user requirement into its decision, and hence degrade the user satisfaction, see Figure 8. Since user requirements is the performance level and the load balancer S2 always user the fastest available resources, it performs well in case the workload is low. In the other cases, it significantly decreases the user satisfaction.

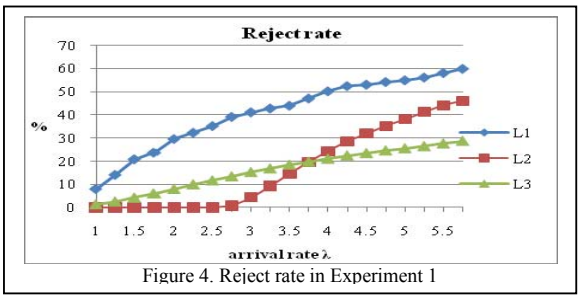


Figure 4. Reject rate in Experiment 1

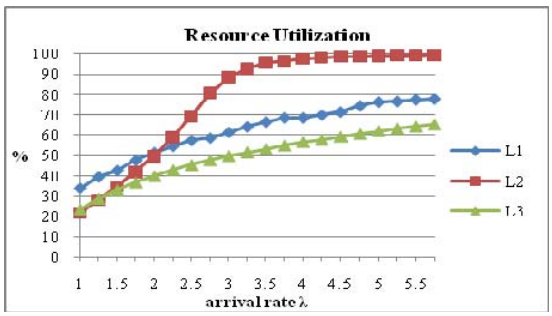


Figure 5. Resource Utilization in Experiment 1.

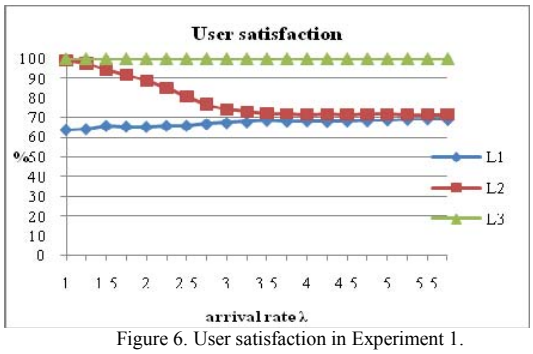


Figure 6. User satisfaction in Experiment 1.

4. CONCLUSION AND FUTURE WORK

The main contribution of this paper is to propose a generic intelligent load balancer framework, which is awareness of multiple quality of service, in large scale distributed computing system. One important characteristic of this framework is that there is a flexibility to integrate various load balancers with different objectives into a system.

Distributed computing systems are nowadays not only designed for integration a small group of resources to execute some class of tasks but also for a pool of resources that execute a large range of task requirements. Those resources are heterogeneous, dynamical and pluggable. Load balancer needs to be more intelligent to adapt the environment change. Therefore, we analyze the logic scheme of an intelligent decider, which is the foundation of intelligent load balancers.

We categorize the resources based on multiple QoSs, and model the environment based on the type of resources. With this model, we propose a load balancer that is aware of the change of utilization of the pool of resources and autonomously adjust its decision to adapt the environment. The advantage of this framework is scalability, extendibility and ability to harmonizing user/task requirements to reduce less violation of these requirements in case the system is overloaded.

For the future work, integrating the decision maker into a system which supports enforcing and negotiating is an interesting problem. Many new problems arise such as competition among users for scarce resources, system maintenance, and task duplication. One of the challenges in load balancing is how it is adapted to different resource-sharing policies. We are also studying on the adaptability of shrinking and expanding mechanism when it is integrated with different sensors and provisioning engines.

REFERENCES

[1] Andrzejak, A., Kondo, D., Yi, S., 2010. *"Decision Model for Cloud Computing under SLA Constraints,"* mascots, pp.257-266, 18th Annual IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems.

[2] Buyya, R., Yeo, C.S., Venugopal, S., Broberg, J., Brandic, I., 2009. *Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility,* Future Generation Computer Systems, Volume 25, Number 6, pp: 599-616.

[3] Braun , T. D., Siegel, H. J., Beck, N., Boloni, L.L. Maheswaran, M., Reuther, A.I., Robertson, J.P., Theys, M. P., Yao, B., Hensgen, D., Freund, R. F., 2001. *A Comparison of Eleven Static Heuristics for Mapping a Class of Independent Tasks onto Heterogeneous Distributed Computing Systems,* Journal of Parallel and Distributed Computing, Volume 61 , Issue 6, Pages: 810 – 837.

[4] Cao, B.Q., Li, B., Xia, Q.M., 2009. *A Service-Oriented Qos-Assured and Multi-Agent Cloud Computing Architecture.* Cloud Computing, Springer Berlin / Heidelberg.

[5] Dasgupta, G.B., Dasgupta, K., Purohit, A., Viswanathan, B., 2006. *QoS-GRAF: A Framework for QoS based Grid Resource Allocation with Failure provisioning.* New Haven, IWQoS. 14th IEEE International Workshop.

[6] Dejun, J., Pierre, G., Chi, C.H., 2009, *"EC2 performance analysis for resource provisioning of service-oriented applications."* In Proceedings of the 3rd Workshop on Non-Functional Properties and SLA Management in Service-Oriented Computing.

[7] Kang, W., Grimshaw, A, 2007, *Failure Prediction in Computational Grids,* pp.275-282, 40th Annual Simulation Symposium (ANSS'07).

[8] Li, H., 2007. *Performance Evaluation in Grid Computing: A Modeling and Prediction Perspective,* Seventh IEEE International Symposium on Cluster Computing and the Grid.

[9] Liang, Q., Wang, Y., 2010. *The Representation and Computation of QoS Preference with Its Applications in Grid Computing Environments,* Annals of Telecommunications, Volume 1 / 1946 - Volume 65.

[10] Menasc, D. A., Dubey, V., 2007. *Utility-based QoS Brokering in Service Oriented Architectures,* IEEE International Conference on Web Services.

[11] Meystel, A.M., Albus, J.S., 2001. *"Intelligent Systems: Architecture, Design, Control."* s.l. : Wiley-Interscience. ISBN:0471193747.

[12] Nguyen, V.H., Khaddaj, S., 2010. *"A Brokerage Framework for Intelligent Resource Allocation in Distributed Systems"*, DCABES & IEEE Computer Society, CPS Publishing, Hong Kong.

[13] Rings, T. G. Caryer, J. Gallop, J. Grabowski, T. Kovacikova, S. Schulz, I. Stokes-Rees, *Grid and Cloud Computing: Opportunities for Integration with the Next Generation Network.* J Grid Computing, Volume 7. ISSN 1570-7873.

[14] Sarathy, V., Narayan, P., Mikkilineni, R., 2010. *Next Generation Cloud Computing Architecture: Enabling Real-Time Dynamism for Shared Distributed Physical Infrastructure,* WETICE '10 Proceedings of the 2010 19th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises.

[15] Toporkov, V. 2009. *Application-Level and Job-Flow Scheduling: An Approach for Achieving Quality of Service in Distributed Computing,* Proceeding PaCT Proceedings of the 10th International Conference on Parallel Computing Technologies.

[16] Xhafa F., Abraham, A., 2009. *A Compendium of Heuristic Methods for Scheduling in Computational Grids.* Springer Berlin / Heidelberg, Vol. 5788.

[17] Wieczorek, M., Hoheisel, A., Prodan, R., 2008. *"Taxonomies of the Multi-criteria Grid Workflow Scheduling Problem"*, Grid Middleware and Services, Springer US.

Qualitative Behavior of a Rational Difference Equation $x_{n+1}=\frac{ax_n^2}{cx_n+bx_{n-1}}$

XIAO Qian
Department of Basic Courses
Hebei Finance University
Baoding, China
xiaoxiao_xq168@163.com

SHI Qi-hong
Department of Basic Courses
Hebei Finance University
Baoding, China
shiqh03@163.com

Abstract **This paper is concerned with the following rational difference equation $x_{n+1}=ax_n^2/\left(cx_n+bx_{n-1}\right)$, with the initial conditions $x_{-1},x_0\in(0,+\infty)$, and $a,b,c\in R^+$. Locally asymptotically stability, global attractivity and boundedness character of the equilibrium point of the equation are investigated. Moreover, simulation is shown to support the results.**

Keywords: Global stability; Attractivity; Boundedness; Numerical simulation

I. INTRODUCTION

Difference equations are applied in the field of biology, engineer, physics and so on[1]. The study of properties of rational difference equations have been an area of intense interest in recent years. There has been a lot of work dealing with the qualitative behavior of rational difference equation. For example, Agarwal et al. [2] investigated the global stability, periodicity character and gave the solution of some special cases of the difference equation

$$x_{n+1}=a+\frac{dx_{n-l}x_{n-k}}{b-cx_{n-5}}.$$

Saleh et al. [3,4] studied the difference equation

$$y_{n+1}=A+\frac{y_n}{y_{n-k}}.$$

In [5] C. Wang et al. dealt with the asymptotic behavior of equilibrium point for the rational difference equation

$$x_{n+1}=\frac{\sum_{i=1}^{l}A_{s_i}x_{n-s_i}}{B+C\prod_{j=1}^{k}x_{n-t_j}}.$$

In [6] Elabbasy et al. has got the global stability, periodicity character and gave the solution of special case of the following recursive sequence

$$x_{n+1}=ax_n-\frac{bx_n}{cx_n-dx_{n-1}}.$$

In this paper we consider the qualitative behavior of rational difference equation

$$x_{n+1}=\frac{ax_n^2}{cx_n+bx_{n-1}},\ n=0,1,\cdots \tag{1}$$

with initial data $x_{-1},x_0\in(0,+\infty)$, and $a,b,c\in R^+$.

II. PRELIMINARIES AND NOTATION

Let us introduce some basic definitions and some theorems that we need in what follows.

Lemma 1 Let I be some interval of real numbers and

$$f:I^{k+1}\to I$$

be a continuously differentiable function. Then for every set of initial conditions $x_{-k},x_{-k+1},\cdots,x_0\in I$, the difference equation

$$x_{n+1}=f(x_n,x_{n-1},\cdots,x_{n-k}),\ n=0,1,\cdots \tag{2}$$

has a unique solution $\{x_n\}_{n=-k}^{\infty}$.

Definition 1(Equilibrium point) A point $\overline{x}\in I$ is called an equilibrium point of (2), if

$$\overline{x}=f(\overline{x},\overline{x},\cdots,\overline{x})$$

Definition2 (Stability)（1）The equilibrium point $\overline{x}$ of (2) is locally stable if for every $\varepsilon>0$, there exists $\delta>0$, such that for any initial data $x_{-k},x_{-k+1},\cdots,x_0\in I$, with

$$\left|x_{-k}-\overline{x}\right|+\left|x_{-k+1}-\overline{x}\right|+\cdots+\left|x_0-\overline{x}\right|<\delta$$

we have $\left|x_n-\overline{x}\right|<\varepsilon$, for all $n\ge -k$.

(2) The equilibrium point $\overline{x}$ of (2) is locally asymptotically stable if $\overline{x}$ is locally stable solution of (2), and there exists $\gamma>0$, such that for all $x_{-k},x_{-k+1},\cdots,x_0\in I$, with

$$\left|x_{-k}-\overline{x}\right|+\left|x_{-k+1}-\overline{x}\right|+\cdots+\left|x_0-\overline{x}\right|<\gamma$$

we have

$$\lim_{n\to\infty}x_n=\overline{x}.$$

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.71

(3) The equilibrium point $\bar{x}$ of (2) is global attractor if for all $x_{-k}, x_{-k+1}, \cdots, x_0 \in I$, we have $\lim_{n\to\infty} x_n = \bar{x}$.

(4) The equilibrium point $\bar{x}$ of (2) is globally asymptotically stable if $\bar{x}$ is locally stable and $\bar{x}$ is also global attractor of (2).

(5) The equilibrium point $\bar{x}$ of (2) is unstable if $\bar{x}$ is not locally stable.

Definition 3 The linearized equation of (2) about the equilibrium $\bar{x}$ is the linear difference equation

$$y_{n+1} = \sum_{i=0}^{k} \frac{\partial f(\bar{x}, \bar{x}, \cdots, \bar{x})}{\partial x_{n-i}} y_{n-i} \tag{3}$$

Lemma 2[2] Assume that $p_1, p_2, \cdots, p_k \in R$ and $k \in \{1, 2, \cdots\}$, then

$$\sum_{i=1}^{k} |p_i| < 1$$

is a sufficient condition for the asymptotic stability of the difference equation

$$x_{n+k} + p_1 x_{n+k-1} + \cdots + p_k x_n = 0,\ n = 0, 1, \cdots \tag{4}$$

Lemma 3[1] Let $g : [p,q]^2 \to [p,q]$ be a continuous function, where p and q are real numbers with $p < q$ and consider the following equation

$$x_{n+1} = g(x_n, x_{n-1}),\ n = 0, 1, \cdots \tag{5}$$

Suppose that g satisfies the following conditions:

(1) $g(x,y)$ is non-decreasing in $x \in [p,q]$ for each fixed $y \in [p,q]$, and $g(x,y)$ is non-increasing in $y \in [p,q]$ for each fixed $x \in [p,q]$.

(2) If (m,M) is a solution of system

$$M = g(M,m) \text{ and } m = g(m,M)$$

then $M = m$.

Then there exists exactly one equilibrium $\bar{x}$ of (5), and every solution of (5) converges to $\bar{x}$.

III. THE MAIN RESULTS AND THEIR PROOFS

In this section we investigate the local stability character of the equilibrium point of (1).

If $c + d \neq a$, then the unique equilibrium point of equation (1) is $\bar{x} = 0$.

Let $f : (0,\infty)^2 \to (0,\infty)$ be a function defined by

$$f(u,v) = \frac{au^2}{cu + bv} \tag{6}$$

Therefore it follows that

$$f_u(u,v) = \frac{acu^2 + 2abuv}{(cu+bv)^2},\ f_v(u,v) = -\frac{au^2 b}{(cu+bv)^2}.$$

Theorem 1 Assume that $ac + 3ab < (c+b)^2$, then the equilibrium point $\bar{x} = 0$ of (1) is locally asymptotically stable.

Proof. When $\bar{x} = 0$,

$$f_u(\bar{x},\bar{x}) = \frac{ac+2ab}{(c+b)^2},\quad f_v(\bar{x},\bar{x}) = -\frac{ab}{(c+b)^2}$$

The linearized equation of (1) about $\bar{x} = 0$ is

$$y_{n+1} - \frac{ac+2ab}{(c+b)^2} y_n + \frac{ab}{(c+b)^2} y_{n-1} = 0 \tag{7}$$

It follows by Lemma 2, (7) is asymptotically stable, if

$$\left|\frac{ac+2ab}{(c+b)^2}\right| + \left|\frac{ab}{(c+b)^2}\right| < 1$$

or

$$\frac{ac+3ab}{(c+b)^2} < 1$$

so

$$ac + 3ab < (c+b)^2$$

This completes the proof.

Theorem 2 The equilibrium point $\bar{x} = 0$ of (1) is global attractor if $a \neq c$.

Proof. Let p, q are real numbers and assume that $g : [p,q]^2 \to [p,q]$ is a function defined by $g(u,v) = \frac{au^2}{cu+bv}$, then we can easily see that the function $g(u,v)$ is increasing in u and is decreasing in v.

Suppose that (m,M) is a solution of system

$$M = g(M,m) \text{ and } m = g(m,M).$$

Then from (1)

$$M = \frac{aM^2}{cM+bm},\ m = \frac{am^2}{cm+bM}.$$

Therefore

$$(a-c)M^2 = bMm, \tag{8}$$

$$(a-c)m^2 = bMm. \tag{9}$$

Subtracting (9) from (8) gives

$$(a-c)(M^2 - m^2) = 0.$$

Since $a \neq c$, it follows that

$$M = m.$$

Lemma 3 suggests that $\overline{x}$ is global attractor of (1) and then the proof is completed.

Theorem 3 For initial data x_{-1}, x_0, if $a \leq c$, that every solution x_n of (1) is bounded, and $x_n \leq M$, here $M = \max\{x_{-1}, x_0\}$.

Proof. Let $\{x_n\}_{n=-1}^{\infty}$ be a solution sequence of (1). It follows from (1) that

$$x_{n+1} = \frac{ax_n^2}{cx_n + bx_{n-1}} \leq \frac{ax_n^2}{cx_n} = \frac{ax_n}{c}$$

Then $x_{n+1} \leq x_n$ for all $n \geq 0$.

Then the sequence $\{x_n\}_{n=0}^{\infty}$ is decreasing and bounded derived from $M = \max\{x_{-1}, x_o\}$.

Corollary 1 For any initial data x_{-1}, x_0, if $a \neq c$ and $ac + 3ab < (c+b)^2$, then the equilibrium point $\overline{x} = 0$ of (1) is globally asymptotically stable.

As a special case of (1), if $a = c$, indeed , the structure suggests the following proposition:

Corollary 2 Assume that $a = c$, that every solution x_n of (1) is non-increasing and bounded, then the equilibrium point $\overline{x} = 0$ of (1) is globally asymptotically stable.

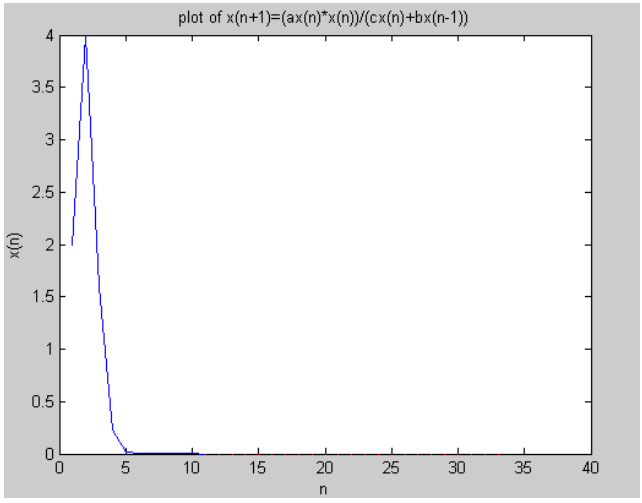
IV. NUMERICAL SIMULATION

In this section, we give some numerical simulations to support our theoretical analysis. For example, we consider the equation

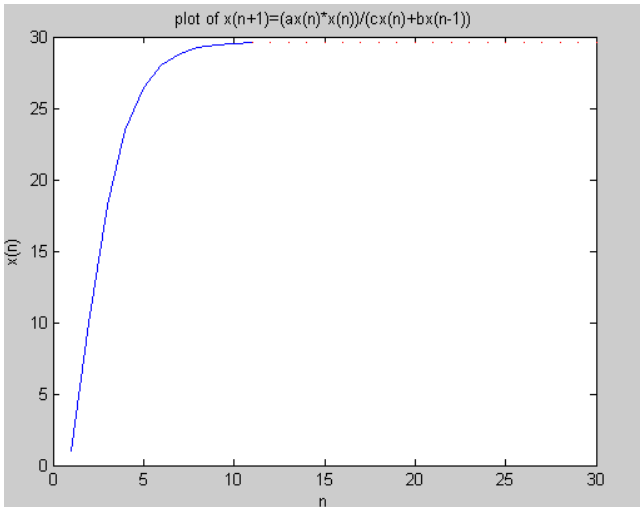
$$x_{n+1} = \frac{3x_n^2}{4x_n + 7x_{n-1}} \tag{10}$$

$$x_{n+1} = \frac{2x_n^2}{x_n + x_{n-1}} \tag{11}$$

We can present the numerical solutions of (10) and (11) which are shown, respectively in Figure 1 and 2. Figure 1 shows the asymptotic behavior and boundedness of the solution to (10) with initial data $x_0 = 2$, $x_1 = 4$, Figure 2 shows the dispersed behavior of the solution to (11) with initial data $x_0 = 1$, $x_1 = 10$.



Figue 1. This figure shows the solution of $x_{n+1} = \frac{3x_n^2}{4x_n + 7x_{n-1}}$, where $x_0 = 2$, $x_1 = 4$



Figue 2. This figure shows the solution of $x_{n+1} = \frac{2x_n^2}{x_n + x_{n-1}}$, where $x_0 = 1$, $x_1 = 10$

REFERENCES

[1] L. Berezansky, E. Braverman, E. Liz, Suffcient conditions for the global stability of nonautonomous higher order difference equations, J. Difference Equ.Appl. 11 (9) 2005, pp. 785–798.

[2] R.P. Agarwal, E.M. Elsayed, Periodicity and stability of solutions of higher order rational difference equation, Adv. Stud. Contemp. Math. 17 (2), 2008 , pp.181-201.

[3] M. Saleh, M. Aloqeili, On the difference equation $y_{n+1} = A + y_n / y_{n-k}$ with $A < 0$, Appl. Math. Comput. 176 (1) 2006, pp. 359-363.

[4] M. Saleh, M. Aloqeili, On the difference equation $x_{n+1} = A + x_n / x_{n-k}$, Appl. Math. Comput. 171, 2005, pp. 862-869.

[5] C. Wang, Q. shi, S.Wang, Asymptotic behavior of equilibrium point for a family of rational difference equation. Advances in Difference Equations. 2010. Article ID 505906.

[6] E.M. Elabbasy, H. El-Metwally, E.M. Elsayed, On the difference equation $x_{n+1} = ax_n - bx_n / \left(cx_n - dx_{n-1} \right)$, Adv. Difference Equ. 2006, pp. 1-10. Article ID 82579.

[7] E.M.E. Zayed. M.A. El-Moneam, On the rational recursive sequence $x_{n+1} = \left(\alpha x_n + \beta x_{n-1} + \gamma x_{n-2} + \delta x_{n-3} \right) / \left(Ax_n + Bx_{n-1} + Cx_{n-2} + Dx_{n-3} \right)$, Comm. Appl. Nonlinear Anal. 12, 2005, pp. 15-28.

[8] R. Memarbashi, Sufficient conditions for the exponential stability of nonautonomous difference equations, Applied Mathematics Letters, 3 (21) (2008), pp. 232–235.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

A Simple Security Model based on Cloud Reference Model [1]

Xiaoli Li
Nantong University
Nantong, China
nuonuoli1218@163.com

Jinhua Chen
Nantong University
Nantong, China
chenjh1981@ntu.edu.cn

Min Luo
Wuhan University
Wuhan, China
jsjgfzx@whu.edu.cn

***Abstract*—Cloud computing is Internet-based computing, whereby shared resources, software and information, are provided to computers and devices on-demand. That is a kind of computing paradigm and means service - oriented architecture. but it also makes security problems more complicate and more important for users than before. In this paper, we define cloud and its characteristics firstly, and then propose a security model based on Cloud Computing Model (SM_CRM). Finally, the experimental results show that the SM_CRM has the more stable performance when facing the attack threat.**

Keywords- cloud computing, Cloud Reference Model, security model

I. INTRODUCTION

Since being proposed in 2007, cloud computing has been a hot researching area of computer network technology at home and abroad. At present, some giant industry companies can offer the cloud services, such as IBM, Google, Amazon, Microsoft and so on. Amazon's EC2 and Google's Google App Engine are typical cloud service. Through the Internet, cloud computing links a large number of resources, then software and IT infrastructure as a resource for outside services. Following the distributed processing, parallel processing and grid computing, Cloud computing becomes a new computing model.

However, security and privacy issues present a strong barrier for users to adapt into Cloud Computing systems. According to an IDC survey in August 2008, which is conducted of 244 IT executives/CIOs and their line-of-business (LOB) colleagues about their companies' use of and views about IT Cloud Services, security is regarded as the top challenge of nine[1]. Fig. 1 shows the nine challenges in detail. Security is the top one concern, say users of Cloud Computing worry about their businesses' information and critical IT resources in the Cloud Computing system which are vulnerable to be attacked.

In this paper, we will briefly introduce Cloud computing, and then introduce the Cloud Reference Model which proposed by CSA. Finally, we propose a security model based on the Cloud Reference Model.

II. CLOUD COMPUTING

A. Definition

Cloud computing is a new computing model, the large

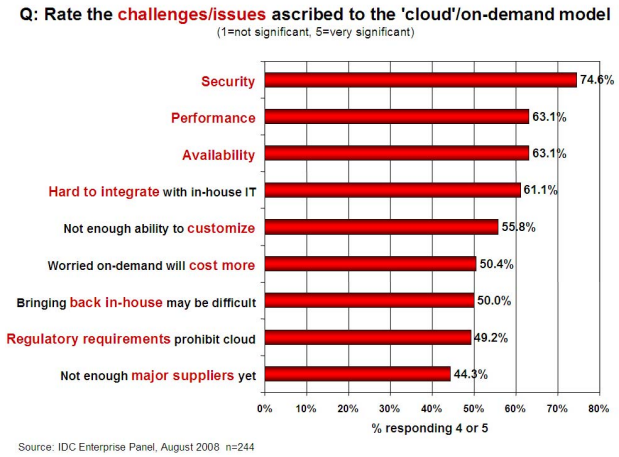


Figure 1. Rate the Challenges/Issues Ascribed to the Cloud On-demand Model

computing was run in the various computing resource on network. Based on user requirements, it can dynamically allocate, deploy, redeploy and cancel the cloud services. Its goal is to make the "computing power" as the water and electricity to supply for user, so that to make it easier for users to use the cloud services [2].

According to the National Institute of Standards and Technology, the definition for "Cloud Computing" is this incomprehensible piece of nonsense clearly written to be as confusing as possible:

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

B. Essential characteristics of Cloud Computing

Cloud services exhibit five essential characteristics[3] that demonstrate their relation to, and differences from, traditional computing approaches:

- **On-demand self-service.** A consumer can unilaterally provision computing capabilities such as server time and network storage as needed automatically, without requiring human interaction with a service provider.
- **Broad network access.** Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs) as

1. Supported by the National Natural Science Foundation of China under Grant Nos. 90718006,90718005, 60743003

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.15

well as other traditional or cloud-based software services.

- **Resource pooling.** The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a degree of location independence in that the customer generally has no control or knowledge over the exact location of the provided resources, but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter). Examples of resources include storage, processing, memory, network bandwidth, and virtual machines. Even private clouds tend to pool resources between different parts of the same organization.
- **Rapid elasticity.** Capabilities can be rapidly and elastically provisioned — in some cases automatically — to quickly scale out; and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.
- **Measured service.** Cloud systems automatically control and optimize resource usage by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, or active user accounts). Resource usage can be monitored, controlled, and reported — providing transparency for both the provider and consumer of the service.

C. *Cloud service model*

There are three kinds of cloud services model, namely, Software as a Service (SaaS), Platform as a Service (PaaS) and Cloud Infrastructure as a Service (laaS), each with different benefits and limitations.

- SaaS is software that is deployed over the internet and/or is deployed to run behind a firewall in your local area network or personal computer. This is a "pay-as-yougo" model and was initially widely deployed for sales force automation and Customer Relationship Management (CRM).
- PaaS, another SaaS, this kind of cloud computing provide development environment as a service. You can use the middleman's equipment to develop your own program and deliver it to the users through Internet and servers.
- IaaS delivers a platform virtualization environment as a service. Rather than purchasing servers, software, data center space or network equipment, clients instead buy those resources as a fully outsourced service.

III. SECURITY ISSUE AND POLICY IN CLOUD COMPUTING ENVIRONMENT

Cloud computing is a new computing model, regardless of the system's architecture or service's deployment is different from the traditional computing model. Therefore traditional security policies are not able to respond to the emergence of new cloud computing security issues [4].

A. *Security issue in cloud computing environment*

Upon strategically deciding on the appropriate cloud delivery and deployment models to explore, security officers should be aware of the current Cloud computing concerns experienced in the Cloud environment. Gartner has conducted an investigation regarding the information security issues that should be considered when dealing with Cloud computing.

The following list contains several security issues highlighted by Gartner that organisations and key decision makers, as a prerequisite, should unpack with Cloud computing vendors [5]:

- Privileged access: Who has specialised/privileged access to data? Who decides about the hiring and management of such administrators?
- Regulatory compliance: Is the cloud vendor willing to undergo external audits and/or security certifications?
- Data location: Does the cloud vendor allow for any control over the location of data?
- Data segregation: Is encryption available at all stages, and were these encryption schemes designed and tested by experienced professionals?
- Recovery: What happens to data in the case of a disaster, and does the vendor offer complete restoration, and, if so, how long does that process take?
- Investigative Support: Does the vendor have the ability to investigate any inappropriate or illegal activity?
- Long-term viability: What happens to data if the cloud vendor goes out of business, is clients' data returned and in what format?

B. *Security policy in cloud computing environment*

In order to solve these problems, the security policy should include the following points[6]:

- Divided into multiple security domains in the cloud computing environment, different security domain operation must be mutual authentication, each security domain internal should have main map between global and local.
- Ensure that the user's connection and communications security with the SSL, VPN, PPTP, etc. Using license and allowing there are multiple authorizations among user, service owner and agents, to ensure user access to data securely.
- User data security assurance: according to the different user's requirements, different data storage protection should be provided. At the same time, the efficiency of data storage should be improving.
- Using a series of measure to solve the user dynamic requirements, including a complete single sign-on authentication, proxy, collaborative certification, and certification between security domains.
- Establishment of third-party monitoring mechanism to ensure that operation of cloud computing environment is safe and stable.
- The computing requested by service requestor, should carry out the safety tests, it can check whether they

contain malicious requests to undermine the security rules.

IV. CLOUD SECURITY MODEL

For understanding cloud computing security risks, we should understanding the relationships and dependencies in the Cloud Reference Model which proposed by Cloud Security Alliance(CSA) firstly.

A. Cloud Reference Model

In the CRM (Cloud Reference Model), IaaS is the foundation of all cloud services, with PaaS building upon IaaS, and SaaS in turn building upon PaaS as described in the Cloud Reference Model as shown in figure 2[3]. The interrelationship and logical boundaries of three cloud service delivery models which Mentioned above were depicted in the CRM.

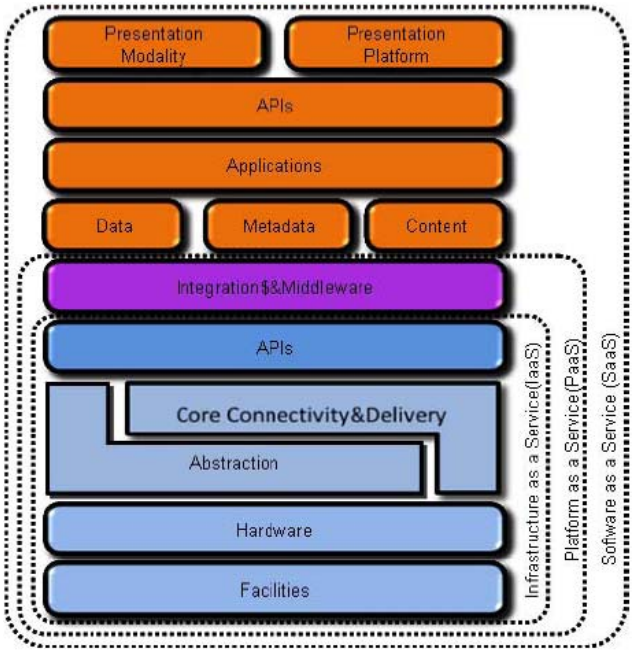


Figure 2. CRM (the Cloud Reference Model)

It is important to note that commercial cloud providers may not neatly fit into the layered service models. Nevertheless, the reference model is important for relating real-world services to an architectural framework and understanding the resources and services requiring security analysis.

IaaS includes the entire infrastructure resource stack from the facilities to the hardware platforms that reside in them. It incorporates the capability to abstract resources (or not), as well as deliver physical and logical connectivity to those resources. Ultimately, IaaS provides a set of APIs which allow management and other forms of interaction with the infrastructure by consumers.

PaaS sits atop IaaS and adds an additional layer of integration with application development frameworks; middleware capabilities; and functions such as database, messaging, and queuing; which allow developers to build applications upon to the platform; and whose programming languages and tools are supported by the stack.

SaaS in turn is built upon the underlying IaaS and PaaS stacks; and provides a self-contained operating environment used to deliver the entire user experience including the content, its presentation, the application(s), and management capabilities.

B. Model Related

1) Physical Security

In the Cloud Reference Model, physical security include the choice of physical location, physical access control, anti-theft and anti-sabotage, anti-lightning, fire, water and moisture, anti-static, temperature and humidity control, power supply, electromagnetic protection, and so on.

2) Computing and Storage Security

Here we mainly consider the authentication, security token, access control, trusted computing[7] and trusted path, security audit, residual information protection, intrusion prevention, malicious code protection, resource control.

3) Networking Security

Network security includes the structural safety, access control, security audit, border integrity checking, intrusion prevention, malicious code protection, network equipment protection, etc.

4) Middleware Security

In Middleware Security, we mainly concern virtual systems, resource manage, transaction monitor, security management, database access.

5) Data Security

In the Cloud Reference Model, Data Security concentrally performance in authentication, access control, data integrity, data confidentiality, backup and recovery, establishment of private “cloud”, data encryption, data isolation, manage and monitor, and so on.

6) Platform Security

Platform security includes platform vulnerability, platform reliability, platforms availability and platform integrity. Corresponding security measures is platform upgrade and Parley-X protection.

7) API Security

API Security mainly performance in loopholes to avoid the API itself, establish a comprehensive security system, security testing for network application software.

8) Web services security

Web services security[8] threats includes the threat of communication and access control. Related technologies mainly include WS-Security, XKMS, SAML, SOAP, XMLDigital Signature, XML Encryption, etc.

In addition to considering security technology, we must also consider the safety management[9], which includes safety management system, security management institutions,

personnel security management, system construction and management.

C. SM_CRM (the Security Model based on CRM)

An analysis of the cloud security problems and the current state of security of the best known clouds shows that these problems do not have any real comprehensive solution and existing cloud security is in its infancy. There is a need for an approach to cloud security that is holistic, adaptable, and reflects client requirements. Refer to the CRM, We propose a simple security model (SM_CRM) as shown in Figure 3[10].

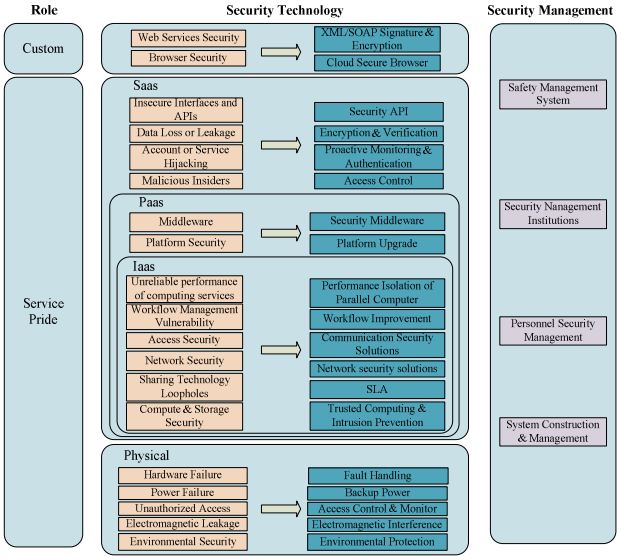


Figure 3. SM_CRM

Actually, cloud security is not only related to technology and management which is mentioned above, but also related to legal and commercial. Specifically, divided into the following aspects.

1) Security and Privacy

It need to assess the implementation process of the following security features:

- authentication
- access control
- encryption/decryption algorithms
- service availability
- application
- security
- data protection
- security incident reporting
- personnel and physical security

2) Governance

In the use of cloud computing model, users give up or reduce the number of decision-making and management which affecting security issues at the same time.

For example,

- external penetration testing is often not allowed
- the full effectiveness of the security log
- fore-nsics process in security incident
- the specific physical storage location and security Configuration of important data
- supply chain security

3) Compliance

Can CSP meet the relevant policies, such as requirements of data privacy protection? Can CSP allow users or third parties to Audit it, for verifying the completeness and effectiveness of its safety policy?

4) Legal

Physical storage of important data may span different countries and regions. And different countries have different justice systems, which will bring the potential legal risks. For example, in different countries judicial interpretations of the responsibility for data loss, data protection of intellectual property and data's disclosure policy may not be the same.

5) Business

Each service mode of CSP is very different. Once you select a CSP, you can be locked by him. This poses a potential security risk in business. For example, in future, once you are not satisfied with the security practices of a CSP, or the CSP exit cloud computing, the user's cost of switching to other CSP is very large.

V. EXPERIMENTS

AAA online shopping system, which is developed by our school, is based on the cloud computing environment. In our experiment, we attack the system before/after adding the SM_CRM security model in the system. By comparing the results, we found that after using SM_CRM security model, the probability of system under attack is reduced.

In the experiment, we used the five most common attacks to attack the system, five attacks are: SQL injection attack, Buffer Overflow Vulnerability, Dos attack, Man-in-the-Middle attack and Electromagnetic attack. Table I is the data comparison in the experiment.

TABLE I. THE DATA COMPARISOM

Security Classification	Example of Attack	Protective Measures in SM_CRM	Attack Number	Attacked Rate	
				No using SM_CRM	Using SM_CRM
Web services security	SQL injection attack	Digital Signature	20	0.9	0.15
			50	0.88	0.22
			100	0.92	0.18
API Security	Buffer Overflow Vulnerability	Patch or Upgrade	20	0.8	0.2
			50	0.86	0.16
			100	0.87	0.09
Data Security	Dos attack	Firewall	20	0.85	0.1
			50	0.9	0.16
			100	0.83	0.12
Network Security	Man-in-the-Middle attack	MAC and IP bundling	20	0.85	0.25
			50	0.92	0.14
			100	0.76	0.08
Physical Security	Electromagnetic attack	Electromagnetic Interference	20	0.75	0.2
			50	0.82	0.18
			100	0.74	0.21

As seen from the above table, the attacked rate of cloud environment, which using SM_CRM security model is much less than the one that no using security model. The experimental result shows that our security model is very effective.

VI. CONCLUSION

All in all, Cloud Computing represents one of the most significant shifts in information technology many of us are likely to see in our lifetimes. It is critical to provide some reliable security technology on cloud computing platform to

prevent security attacks, as well as the destruction of infrastructure and services.

In this paper, SM_CRM security model is proposed, which based on the CRM. And the availability of the model were verified by experiment, that can provide the basis for the deeper research on security deployment of cloud computing.

Cloud computing has the potential to become a frontrunner in promoting a secure, virtual and economically viable IT solution and future work and progress lies in its depth study of the safety program on the base of existing security model.

REFERENCES

[1] IDC, "It cloud services user survey, pt.2: Top benefits & challenges," *http://blogs.idc.com/ie/?p=210*, 2008.

[2] Wikipedia, http://en.wikipedia.org/wiki /Cloud_Computning.

[3] Cloud Security Alliance. Security Guidance for Critical Areas of Focus in Cloud Computing. http://www.cloudsecurityalliance.org/guidance/csaguide.pdf

[4] Brodkin J, 2008, 'Gartner: Seven cloud-computing security risks', *Infoworld*, viewed 13 March 2009, http://www.infoworld.com/d/security-central/gartner-seven-cloudcomputing- security-risks-853?page=0,1.

[5] Francesco M.A and Gianni F. "An approach to a cloud Computing network", IEEE, August 2008, pp113-118.

[6] Xue Jing and Zhang Jian-jun. "A Brief Survey on the Security Model of Cloud Computing". 2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science, 10 –12 August, pp475-478.

[7] Zhidong Shen, Li Li, Fei Yan, Xiaoping Wu. "Cloud Computing System Based on Trusted Computing Platform". 2010 International Conference on Intelligent Computation Technology and Automation. 11-12 May 2010. pp942-945.

[8] Meiko Jensen, J¨org Schwenk, Nils Gruschka, Luigi Lo Iacono. "On Technical Security Issues in Cloud Computing". 2009 IEEE International Conference on Cloud Computing. 21-25 Sept. 2009. pp109-116.

[9] Zhiyun Guo, Meina Song, Junde Song. "A Governance Model for Cloud Computing". Management and Service Science (MASS), 2010 International Conference on. 24-26 Aug. 2010.

[10] Sang-Ho Na, Jun-Young Park and Eui-Nam Huh. "Personal Cloud Computing Security Framework". Services Computing Conference (APSCC), 2010 IEEE Asia-Pacific. 6-10 Dec. 2010. pp671-675.

The Design of a Private Cloud Infrastructure Based on XEN

Xinyu Miao
BeijingUniversityofPostsTelecommunications,100876,China
Miaoxinyu1987@126.com

Jing Han
Beijing University of Posts Telecommunications,100876,China
Babyblue110128@hotmail.com

Abstract—In order to benefit from cloud computing services without risk, a good solution is building a private cloud in premise. This paper analyzes the advantages of a private cloud, through physical machine clusters and XEN virtualization technology, A private cloud topology is designed first, and analyze cache mechanism and load balancing policy of a private cloud infrastructure, then design a XEN-based private cloud infrastructure. This private cloud infrastructure provides unified and simple interface for users.

Keywords-Cloud Computing ; private cloud ; XEN

1.INTRODUCTION

So far, the industry does not have a uniform definition about cloud computing. Everyone has their own understanding of the cloud in the minds. The emergence of cloud computing is the inevitable result of the development of parallel computing technology, software technology and network technology. The emergence of parallel computing is due to the fact that people were not satisfied with the growth rate of CPU speed of Moore's, In Parallel computing era, people strongly pursue high-speed computing and use expensive servers, this huge investment can't be recycled. In the cloud computing era , we do not pursue the use of expensive servers, cloud computing has the important feature of on-demand services and on-demand payment. The user only requests the calculation and other services from cloud computing centers when needed , when the use is completed the cloud computing center release resources immediately and redistribute to other terminal. Cloud computing will be all hidden in the clouds; the average user is no longer concerned about computer viruses. All this is done by the Center of cloud computing, the average user has the opportunity to enjoy high-performance computing because the cloud center can provide almost unlimited computing power, flexible computing and storage is also the important feature of the cloud computing flexibility.

This work is supported by the National Key project of Scientific and Technical Supporting Programs of China (GrantNos.2008BAH24B04,2008BAH21B03,2009BAH39B03); the National Natural Science Foundation of China(Grant No.61072060); the Program for New Century Excellent Talents in University (No.NECET-08-0738); Engineering Research Center of Information Networks.Ministry of Education.

2.PRIVATE STORAGE CLOUDS AND ITS ADVANTAGES

2.1.What Is a Private Cloud

The private storage cloud is for the public storage clouds. Related to the large amount of data stored in the public cloud, more businesses and government hope to achieve a secure internal data storage platform to further ensure security at the same time, instead of the original solution such as FTP, and building private clouds is able to achieve this goal[1]. This is almost a fully-equipped private cloud, so that private storage clouds only provides storage service and the corresponding quality of service (QoS) for limited users. Users who use storage services do not need to understand the "cloud" and the specific details included, just know the corresponding interface; the remaining works were completed by the "cloud". Users only need to look the private cloud as a black box pool of resources; users do not need to care about how the interior is, how to configure it, what kind of technology is needed and what kind of platform is used.

2.2.The Advantages of Private Cloud

Private clouds can manage massive amounts of data unified. First, decentralization does not guarantee data consistency; Second, the users want to manage their own storage and lead everyone to do repetitive tasks which result in the efficiency of ground and waste of human resources; again, it is difficult for an effective information control, information disclosure, and security will become a prominent issue.

The private cloud is easy to implement centralized backup and disaster recovery and it is easy to extend and upgrade. Because users only know that storage interface, does not know the realization of storage, which is equivalent to the private storage clouds and joins the middle layer between users, so the private cloud changes in the backend, does not affect the front end, and will not affect the users. Space for the expansion of the private storage clouds, maintenance, upgrades bring flexibility, back-end changes have minimal impact.

2.3.Development of Cloud Computing

In addition to Amazon that, many multinational companies such as information technology industry, IBM, Yahoo, Microsoft and Google, are beginning to use the concept of cloud computing to sell their products and services, cloud computing are favorable for the following reasons[3]:

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.91

Price: cloud computing adopts a distributed system which has better performance and low cost than centralized system, the users can even use the cheapest form of cloud PC, and computing performance can indeed exceed the mainframe.

Distributed Application: For example the majority of enterprise applications has itself distributed.

Reliability: a highly fault-tolerant mechanism for distributed systems, relying on server clusters ensures that, even if the single point of the cloud server fails, the task can also be calculated easily to migrate to other servers, to ensure the normal operation of computing.

Scalability: You can add new server in real time to an existing server cluster to increase the computing power of the cloud.

Flexibility: Cloud Computing can be compatible with different hardware vendors, configure the low machines and peripherals in order to make it compatible to obtain high-performance computing.

3.XEN VIRTUAL TECHNICAL OVERVIEW

XEN is based on open-source GPL License virtualization software, which originated in a research project at Cambridge University, and later XEN was independent from the research project into a community-driven open source software projects. XEN supports full virtualization (HVM) and the par virtualization (PV), initially using XEN virtualization technology class the idea of a Linux kernel modified to achieve the CPU and memory virtualization, hardware virtualization (HVM) also have added the development of XEN, XEN has full support for Intel VT and AMD-V hardware technology etc. terms which are related to XEN will be introduced as follow:

XEN Hypervisor is hypervisor, many places can be called VMM, which is located between the hardware and operating system, and it offer virtual environment for the operating system.

There are two virtual domains in the XEN, respectively Domain-0 and Domain-U, One Domain-0 is running in privilege mode, and Domain-U is our virtual machine, which we call VM.

Virtual MMU is used to help Guest Os to complete the address translation, which is to help us complete the virtual memory.

Event channel is mainly used for the Domain and the asynchronous event notification mechanism between the domain and the hypervisor.

Native device driver is used for general operating system drivers, and Domain in XEN, which only runs in privilege mode, can call native device drivers to access the hardware.

Front and rear drive (FE / BE) is often applied in I / O device emulation, through a request to implement transactional data communication between the Domain. Front-end devices which are used to sent access requests through the event channel asynchronous to back-end, and the end-back will call back native driver to call the hardware , and make a feedback on the front end.

4.THE PROCESS OF XEN-BASED PRIVATE CLOUD INFRASTRUCTURE

Just for outside customers, Private clouds only provide a interface to be called. Internal implementation relies on dynamic start and distribution of computer nodes. Designed to be scalable to the size of cloud dynamic clusters, the cloud can be effectively utilize computing resources and the advantages of virtual machine technology. In the private cloud, for each node requires a sound management control mechanisms, particular in cloud stretching and node startup, shutdown, data dump and resource allocation.

4.1.Improved Star Topology Model

Star topology is simple and can facilitate centralized managing, all data must be through central processor, the other nodes link the central processor by physical link directly. The bus-shaped topology, only two computers can communicate at the same time, and network extension is limited, the number of nodes is limited, so that it can't meet the requirements of the private cloud infrastructure flexibility. For the ring topology a node failure will cause physical paralysis and can't achieve security and reliability of protection of private clouds. But this topology has limitations, that central node has single point of failure. Based on consideration of all aspects, improved star topology is the most appropriate framework for private clouds.

The improved star structure of the design is conducive to give full play to the scalability of the cloud features. Private cloud use a central control node connected to multiple sub-nodes and control points of the star node structure. Deputy central node is on when the central node fails, that ensure the continued connectivity of topology. Shown in Figure 3.1

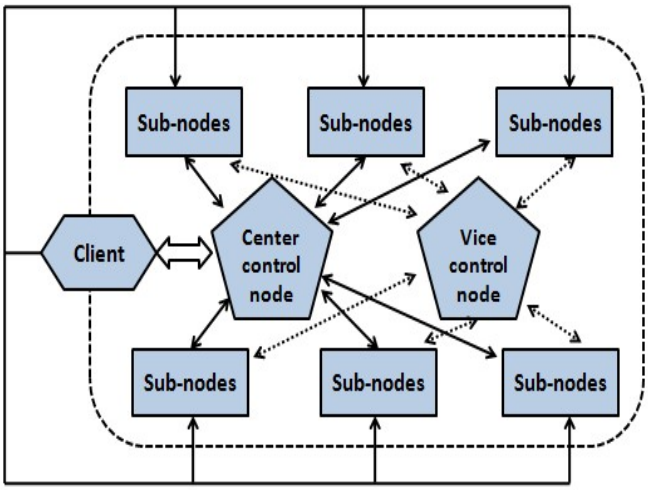


Figure 3-1 Model-based improvedngg star topology design

4.2.Cache Design in the Control Center

The data in the private cloud storage service, the data stored in the sub-node first, but the number of nodes in the cloud and private node resources are limited, so data must be dumped from the node to the data pool for long-term storage. The widespread typical application of cache is that

client data uploads to the private cloud, then the data still remain in the node which has not moved to the data pool, if the service receives the request at this time, the node can directly provide services for the client. Cache design principle is: Each node can runs in the cloud as a Cache, when the node storage capacity does not meet the ceiling or the node does not receive clear instructions to dump the data to the data pool before the server is always to deal with it as Cache.[8] When a client requests to download data, the central node check Cache table, to determine whether the requested data downloaded already exists in a node, if found, shall be deemed that Cache hit, client feedback information to control center, allowing the client to interact directly with the node to complete the data download; If not found, then the positioning data pool and client interact with each other. Figure 3-2 Basic Cache management.

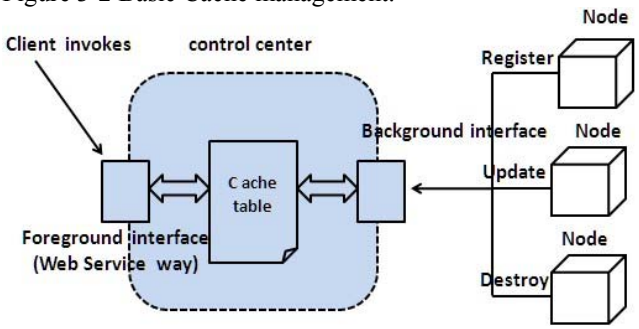


Figure 3-2 Cache Management System of Control Center

As to the maintenance of cache table and the starting of a new code, it must receive the data from the client and download from the data pool, After the data on each node updates, the node through the control center's corresponding service interface interact and then registries and updates the information from the control center on to the Cache table. When a node is destroyed, at the same time node information on the Cache table is also destroyed.

4.3.Strategy of Load Balance in Private Cloud

Each node in the private clouds has different performance and actual operation; it is need to add load balancing mechanism in the control center for the management of each node. Control Center constantly monitors all running node, the node itself is defined performance weights by a performance of hardware, and for its operation of the node because one or several processing is running, the rest of their computing power can be used as another indicators. When control center node distributes the operation, it will follow a certain algorithm, based on the amount of work to run and load, quantify the distribution of the computation of a desired target, and compares to the remaining computing power of each running node and identify the most appropriate nodes to run this job, and distribute the operation to it. When the new jobs' demand of any computing power is greater than the capacity of the node, the control center will launch a new node, expand the carrying capacity of the cloud. Service program on the control node is defined as the maximum node load time.

4.4.S3 Storage Pool

Various sub-node in private cloud infrastructure is mainly used for the mass data calculation, as to mass data storage in the private cloud, the sub-node has only a small amount of Cache, when facing with the destruction or the node in response to the client to upload or download massive amounts of data, We need to dump the data on the node which will be destroyed to the data pool. When customers need to upload huge amount of data, the control center adjusts the manners that respond to the data pool to communicate directly with the client. Data storage pool which can use Amazon's S3 (Amazon Simple Storage Service), of course, before dumping the data to s3 to be encrypted..

S3 provides online storage space, relying on a distributed storage technology, to solve the storage problems of mass data (especially unstructured data), while the Web Service of S3 provides developers with a development interface and allows third-party tools, For example, S3 Backup, Duplicity, S3 Solutions and more.

5.XEN-BASED ARCHITECTURE MODEL OF PRIVATE CLOUD

5.1.A description of each part

Through the analysis of Section III, We build a private storage cloud and service as follows, the data is encrypted and dumped in the data storage pool, and cloud processing center uses a physical machine cluster of the enterprise LAN, XEN-based private cloud infrastructure shown in Figure 3-3:

In the private cloud, the control center use a strong physical machine with strong performance and other sub-nodes build up star distribution，Each sub-node is a XEN-Linux-installed component of the physical cluster of physical machines, the permanent storage of data is provided by the S3 storage pool. Open the SSH service for each physical machine, after each code of private cloud starts, the SSH service can achieve that we visit it without the password. The control center connect to physical machine through SSH, the SSH controller in the control center use jash library to achieve java programming to establish SSH connection to the cluster of physical machines. When the connection is successful, call them to achieve a good shell script writing, and start the virtual machine, then inform the control center.

Control Center manage virtual machines by running the XEN's xm command and use to start the virtual machines, usually employ the way to load the configuration file, the private configuration files used in the cloud is as follows:

```
name="hadoop 52vm"
uuid=" 10689058-1754-4103-ba57-47147a65227d"
manxmen = 512
memory = 512
vcpus= 1
boot loader="/usr/bin/hadoopm"
on_poweroff="destroy"
on—reboot="restart"
on—crash="restart"
```

vfb=[Ittype—vnc,vncunused=l,keymap=en—US”]
disk=[“tap:aio:/xen-vms/hadoop52vm,xvda,W”]
vif=[“bridge=xenbr0”]

The configuration file is generated by the shell program, the value of disk which refers to the file that corresponds to virtual disk files, which is located in the control center NFS shared directory, SSH connection in the physical machine running shell script first copies the file to a local copy, and then generates the configuration file, while uuid is unique identification of a virtual machine, which is randomly generated .

In order to achieve fault tolerance of the private cloud, the

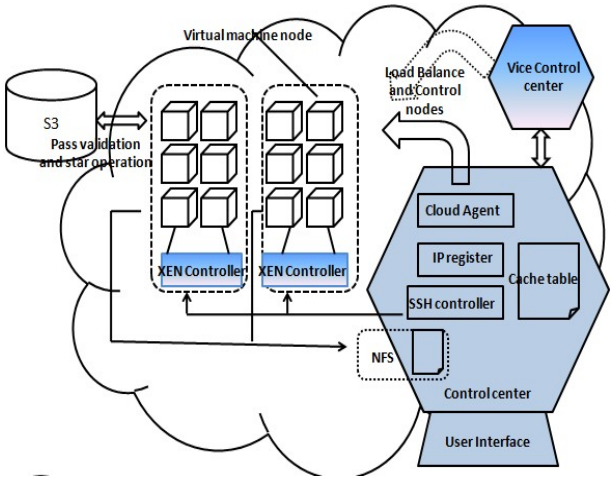


Figure 3-3 Relationship between the components of private cloud

cloud system in the control center of the private uses the NFS file system to implement the simultaneous updates between code content and the central node content and the corresponds directory automatically, to eliminate the trouble that copies a node to a node in order to save the file each time.

Figure 3-3 control center in private cloud infrastructure, whose mission is to receive and process requests from a large number of concurrent users, the users interact with the control center through Web Service interface which is provided by the private cloud platform. The internal nodes interact with each other through cloud agent, when a client makes a request, the control center sends it to cloud agent after receiving and analyzing, cloud agent achieves a Map <String,CloudNote>, key for the nodeID, value for the CloudNote object, When checking all nodes to find the maximum node, if the node has the ability to accommodate to a new job, then calls the node addJob method to add new jobs to the node's job list, or call the new node directly to interact with users handle the job . Cloud monitoring agent runs the thread at the same time, constantly checks the jobs running on each node, if found operating in the client request times out, then deletes operation and releases the node resources. If the central node is down, the vice center node take over its deputy to complete the task.

5.2.Flowchart nodes in the private cloud

Node management should include the start node, initialization, allocate and release a series of control activities. Control Center monitors the operational status of resources in the cloud, through a certain dynamic adjustment algorithm each node running in the cloud. when the cloud would not be suitable to run the new user request because of lacking nodes, Control Center will start a new process node and running services to be initialized, assigned to the cloud cluster, in operation, the expansion of the load capacity of the clouds; When the cloud reaches a certain space of a node limit, the control center to properly handles their data, and releases node, recovery.

When the cloud size is insufficient to meet the request of concurrent users, and the cloud has not enough free resources to run the new processing nodes, the client request will return an error message. Although the "private cloud " is designed to make users feel the physical machine resource constraints, this should be considered in the design of data storage services, node management flowchart is as follows:

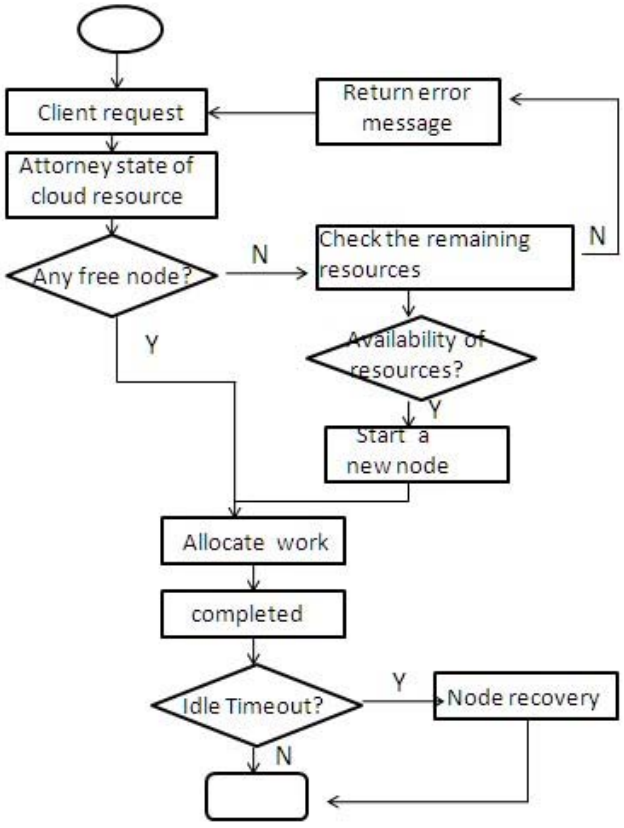


Figure 3-4 Node management process in the private cloud

6.CONCLUSION

This paper studies the XEN-based private cloud architecture, the main problem is how to build a range of computing and storage resources through the "private cloud" approach, which has the characteristics of flexibility. Of

course, framework is just the first step, and how to provide users with cost-effective private cloud services in an inexpensive way, so the private cloud services used in real life really is the ultimate goal, there is a long way to go for cloud services, Believe that private cloud will play an enormous impact on our lives.

REFERENCE

[1] Shuai Zhang; Shufen Zhang; Xuebin Chen; Xiuzhen Huo; "Cloud Computing Research and Development Trend "Future Networks, 2010. ICFN '10.

[2] Jiyi Wu; Lingdi Ping; Xiaoping Ge; Ya Wang; Jianqing Fu; "Cloud Storage as the Infrastructure of Cloud Computing" Intelligent Computing and Cognitive Informatics (ICICCI), 2010

[3] Shufen Zhang; Shuai Zhang; Xuebin Chen; Shangzhuo Wu; "Analysis and Research of Cloud Computing System Instance" Future Networks, 2010. ICFN '10.

[4] Baun, C.; Kunze, M.; "Building a private cloud with Eucalyptus" E-Science Workshops, 2009 5th IEEE International Conference on

[5] Hofmann, Paul; Woods, Dan; "Cloud Computing: The Limits of Public Clouds for Business Applications" Internet Computin

[6] Mikkilineni, R.; Sarathy, V.; "Cloud Computing and the Lessons from the Past"Enabling Technologies: Infrastructures for Collaborative Enterprises, 2009. WETICE '09. 18th IEEE International Workshops on

[7] Yun Chan Cho; Jae Wook Jeon; "Sharing data between processes running on different domains in para-virtualized xen" Control, Automation and Systems, 2007. ICCAS '07.

[8] Zhang Jian; Li Xiaoyong; Guan Haibing; "The Optimization of Xen Network Virtualization" Computer Science and Software Engineering, 2008

[9] Nguyen Anh Quynh; Takefuji, Y.; "A Real-time Integrity Monitor for Xen Virtual Machine" Networking and Services, 2006. ICNS

[10] Fallenbeck, N.; Picht, H.-J.; Smith, M.; Freisleben, B.; "Xen and the Art of Cluster Scheduling" Virtualization Technology in Distributed Computing, 2006. VTDC 2006.

Face Recognition with Single Training Sample Based on Local Feature Fusion

Chen Yang
1Chinese Information Processing Research Center,
Beijing Information Science and Technology University
Beijing, China
yangchen49781932@sina.com

Shuicai Shi
1 Chinese Information Processing Research Center,
Beijing Information Science and Technology University
2 Beijing TRS Information Technology Co., Ltd.
Beijing, China
shi.shuicai@trs.com.cn

Lin Li
1 Chinese Information Processing Research Center,
Beijing Information Science and Technology University
2 Beijing TRS Information Technology Co., Ltd.
Beijing, China
li.lin@trs.com.cn

Xueqiang Lv
1 Chinese Information Processing Research Center,
Beijing Information Science and Technology University
2 Beijing TRS Information Technology Co., Ltd.
Beijing, China
lv.xueqiang@trs.com.cn

***Abstract*—In the condition of single training sample,traditional methods get low recognition accuracy,or even can not be used. In view of this situation, this paper proposes a method to slove this problem.Firstly,face image is decomposed by imagepyramids.Then, each layer image segmentation into sub images with th same size.After that,the feature of each sub image,which got with $(W2DPC)^2A$,gets a weight through the adaptive method.Finally, Euclidean distance is used to classify face images. Experimental results on ORL and Yale show that the presented method can achieve a certain degree of recognition accuracy.**

Keywords-microblog; Face Recognition with Single Training Sample; Principle Component Analysis;Image bolck; $W(2D)^2PCA$;Adaptive weight

I. INTRODUCTION

Face recognition technology has been a very hot topic in mode identification field,which has broad prospect of application in security systems, crime recognition and identity,etc.The difficulty of face recognition lies in the face with rich expressions, illuminations, imaging angles and time factor. Through the research for many years, Face recognition technology has made great progress and many methods[1,2] used for face recognition were proposed. But only one image for each person coule be obtained in some special circumstances,such as Id or passport validation,so single sample face recognition problem appears. The traditional methods of face recognition get low recognition accuracy,or even can not be used in this condition.The research of face recognition with single training sample has great significance and faces a challenge.

II. CORRELATION STUDY

At present, there are several methods on face recognition with single training sample ,as follows: geometry character method, sample expansion method, feature subspace expansion method, general learning framework method, image enhancement method, neural network method, hidden markov model,etc. Principal component analysis[3](PCA) belong to feature subspace expansion method.

In the classical principal component analysis, the image vector,which the image matrix need to be converted into ,has high dimension.Not only is the process of calculation and analysis time-consuming, but also it is difficult to get eigenvalues and eigenvectors of covariance matrix in the high dimension space.And it would loose some structural information of the image in the conversion process. Two-dimensional principal component analysis[4](2DPCA) offers a good method to solve the problem. Because it is feasible to directly computing covariance matrix of a image matrix, Computing time and recognition result have improved.Based on 2DPCA, two direction of two-dimensional principal component analysis[5] ($(2D)^2PCA$) is proposed which could get a lower dimensional matrix coefficient and cut down the time of recognition.

There are no difference between treatment effect of each eigenvector of image characteristics matrix in 2DPCA. This paper presents a method called $(W2DPC)^2A$ which is improved in $(2D)^2PCA$ and W2DPCA mentioned in literature[6].In this paper,firstly,face image is decomposed by image pyramids.After that each level image in image pyramids is segmented into blocks with the same size.Next step is to get the feature and the weight of each block in $(W2DPC)^2A$. Finally,the recognition results are showed using Euclidean distance after feature fusion.

III. 2DPCA

A. Algorithms

Assuming X is a n dimension vector of columns and A is a image matrix whose size is m by n,image projects to

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.43

vector X according to the linear transform formula $Y=AX$,Y a m x 1 dimensional vector.best projection of the projection sample overall spreading matrix is described by trace of the covariance matrix of projection characteristics vector,that projection axis X is determined by the spread of the eigenvector Y.

$$J(X) = tr(S_x) \tag{1}$$

In formula(1), S_x is the covariance matrix of the training sample projection eigenvector and $tr(S_x)$is the trace of matrix S_x.According to formula(1),you can get a projection direction X,which projection to produce a maximal general discrete degrees of projection samples. Covariance matrix S_x can be expressed as:

$$\begin{aligned} S_x &= E(Y-EY)(Y-EY)^T \\ &= E[AX-E(AX)][AX-E(AX)]^T \\ &= E[(A-EA)X][(A-EA)X)]^T \end{aligned} \tag{2}$$

So,

$$tr(S_x) = X^T[E(A-EA)^T(A-EA)]X \tag{3}$$

Define,

$$G_t = E(A-EA)^T E(A-EA) \tag{4}$$

Matrix G_t with size $n \times n$, is covariance matrix of images, calculated by training images. G_t is a non-negative-definite-matrix according to defination. Suppose that there are N training image samples $A_i(i=1,2\ldots N)$ with size m by n,

$$G_t = \frac{1}{N}\sum_{i=1}^{N}(A_i-\overline{A})^T(A_i-\overline{A}) \tag{5}$$

$$\overline{A} = \frac{1}{N}\sum_{i=1}^{N}A_i \tag{6}$$

Formula(1) can be expressed as:

$$J(X) = X^T G_t X \tag{7}$$

Projection vector X is a column vector making $J(X)$ take the maximum.and corresponding with the maximum eigenvalue. A group of projection shafts X_k need to meet the following conditions:

$$\begin{cases} \{X_1, X_2, \cdots X_k\} = \arg\max J(X) \\ X_i^T X_j = 0, i \neq j, i, j = 1,2,\cdots k \end{cases} \tag{8}$$

Vectors $\{X_1, X_2, \ldots X_k\}$ correspond with top k maximum eigenvalues $\{\lambda_1, \lambda_2, \ldots \lambda_k\}(\lambda_i > \lambda_j, i>j)$ of G_t.k needs to meet the following condition:

$$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \geq \theta \tag{9}$$

θ is a threshold ,usually takes 0.9.Define a projection matrix $U=[X_1, X_2, \ldots X_k]$ with size n by k.

B. Feature extraction

According to $Y_i=AX_i(i = 1, 2, \ldots k)$, Y_i is the eigenvector of image A projected to X_i.Define matrix $B=[Y_1, Y_2, \ldots Y_k]$,

$$\begin{aligned} B &= [Y_1, Y_2, \cdots Y_k] = [AX_1, AX_2, \cdots AX_k] \\ &= A[X_1, X_2, \cdots X_k] = AU \end{aligned} \tag{10}$$

B is the projection feature matrix of image A.

C. W2DPCA

Due to the different contributions of characteristic vectors, an eigenvalue is larger, the greater the contribution of eigenvector.Accordin to literature[6], in order to give prominence to the different contribution of eigenvectors, using eigenvalues as weights of eigenvectors, W2DPCA is produced.

$$Y_i = A\left(\lambda_i^{\omega} X_i\right) \tag{11}$$

λ_i is the i-th maximum eigenvalue of G_t.

$$\begin{aligned} B &= [Y_1, Y_2, \cdots Y_k] \\ &= [A\left(\lambda_1^{\omega} X_1\right), A\left(\lambda_2^{\omega} X_2\right), \cdots A\left(\lambda_k^{\omega} X_k\right)] \\ &= A[\lambda_1^{\omega} X_1, \lambda_2^{\omega} X_2, \cdots \lambda_k^{\omega} X_k] \\ &= AU\Lambda \end{aligned} \tag{12}$$

$$\Lambda = \begin{bmatrix} \lambda_1^{\omega} & & & \\ & \lambda_2^{\omega} & & \\ & & \cdots & \\ & & & \lambda_k^{\omega} \end{bmatrix} \tag{13}$$

D. $(W2DPC)^2A$

According to the formula $Y=AX$, the image A is transformed on row direction,so 2DPCA is called row-direction-2DPCA.Similarly,we can get column-direction-2DPCA.

The covariance matrix

$$G_t = E(A-AX)(A-AX)^T = \frac{1}{N}\sum_{i=1}^{N}(A_i-\overline{A})(A_i-\overline{A})^T \quad (14)$$

The feature matrix

$$B = V^T A \quad (15)$$

In formula(15), $V = [X_1^T, X_2^T, \cdots X_d^T]$

In literature[5], row- direction-2DPCA and row-direction-2DPCA combines into $(2D)^2$PCA.

$$C = V^T AU \quad (16)$$

C is a feature matrix，A is an image matrix，V is a row direction projection matrix，U is a column direction projection matrix.

Compared to 2DPCA,$(2D)^2$PCA can make the dimension of an image characteristic matrix lower，

$$C = \Lambda_V V^T AU \Lambda_U \quad (19)$$

IV. PARTIAL FEATURE EXTRACTION AND INTEGRATION

A. Image pyramid

In order to better get the information of image content, usually, an image is processed in multi-scale. Image pyramid is a kind of effective multi-scale expression structures, using different resolution of an images as the image multi-scale.

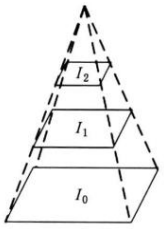


Figure 1. Image pyramid.

Gaussian pyramid[7] construct formula is as follows:

$$G^{(k+1)} = C_{(\downarrow 2)} G^{(k)} \quad (20)$$

$G^{(k)}$ is level k image in Gaussian images.C is a onvolution mask for compressed smooth.“↓”says downsapling.“2”is The sampling rate.We can get level k+1 Gaussian image gaussian smooth and downsapling on level k Gaussian image.

without any decline on recognition effect according to literature[5].

$$C = V^T(AU) = V^T B \quad (17)$$

In fact, V is the projection matrix of B,rather than A.So redefine column direction covariance matrix.

$$G_t' = \frac{1}{N}\sum_{i=1}^{N}(B_i-\overline{B})(B_i-\overline{B})^T \quad (18)$$

V is composed by eigenvectors eigenvector corresponding with top d maximum eigenvalue.We call this method $(2DPC)^2$A.

We can get $(W2DPC)^2$A by setting the weights to two directions according formula(12).

B. Image segmentation and Feature extraction

Because the whole image features can't very well said the detail of the image,and expressions and position changes are more apparent in the local,it could utilize image information more fully that an image is segmented into blocks and recognition feature is composed with each block feature.

In this paper,an image is standardized into 80 x 80,and get different resolution images by Gaussian pyramid.Each of different resolution images is segmented into blocks with size 10×10.

Figure 2. Image blocks

As shown in figure 2,the number of blocks.is 85. Use $(W2DPC)^2$A to calculate the characteristics of each block set which is composed by corresponding blocks.

$$C_{ij} = \Lambda_{V_j} V_j^T A_{ij} U_j \Lambda_{U_j} \quad (21)$$

A_{ij} is block j in image i，C_{ij} is the feature of block j in image i.

C. Local features fusion

In this paper,image similarity is measured by the sum of weighted Euclidean distance between blocks.

$$D_i = \sum_{j=1}^{M} \omega_j d_{ij} \tag{22}$$

$$\sum_{j=1}^{M} \omega_j = 1 \tag{23}$$

D_i is the distance between the image and training image i. d_{ij} is the distance between bolck j in this image and block j image i. ω_j is the weight of block j.M, the number of blocks, equals 85 in this paper.

The weight ω of each block is determinrd by the distance between blocks in testing image and in training image.

$$\sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N}(\frac{d_{ij} - \overline{d_j}}{\overline{d_j}})^2 \tag{24}$$

$$\overline{d_j} = \frac{1}{N}\sum_{i=1}^{N} d_{ij} \tag{25}$$

σ^2_j is the variance of distances between block j in testing image and in training image.

$$\mathrm{P} = \sum_{j=1}^{M} \omega_j \sigma_j^2 \tag{26}$$

The weight ω of each block need to make P minimal and meet formula(23).So we can get:

$$\omega_j = \frac{1}{\sigma_j^2 \sum_{j=1}^{N} \frac{1}{\sigma_j^2}} \tag{27}$$

V. Experiment and Result AnalysiS

A. Experimental Data

In this paper, we have done some experiments on ORL(40 persons,10 image each one,total 400 images) and Yale(15 persons,11 image each one,total 165 images).In the experiments, respectively,we take one image each person as training samples and other images as testing samples.

B. Values of ω_U and ω_V

There is an experiment on ORL to get two weights in $(W2DPC)^2A$. making the recognition result best.

TABLE I. Result on different Ω_U and Ω_V

ω_V \ ω_U	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0	0.61	0.6175	0.6225	0.63	0.64	0.65	0.655	0.655	0.655	0.66	0.6575
0.1	0.61	0.62	0.6275	0.635	0.6525	0.6475	0.6575	0.6575	0.6575	0.6575	0.6625
0.2	0.615	0.6275	0.6425	0.6475	0.66	0.655	0.6575	0.655	0.655	0.655	0.6625
0.3	0.625	0.635	0.6425	0.6575	0.6575	0.666	0.6575	0.6525	0.655	0.6575	0.6625
0.4	0.625	0.635	0.6475	0.66	0.66	0.66	0.655	0.6475	0.6575	0.6625	0.665
0.5	0.6225	0.635	0.6575	0.655	0.655	0.6525	0.65	0.6525	0.66	0.66	0.665
0.6	0.6525	0.6425	0.655	0.66	0.65	0.6525	0.65	0.655	0.6625	0.6625	0.6625
0.7	0.6325	0.645	0.66	0.6625	0.6525	0.655	0.65	0.6575	0.66	0.66	0.66
0.8	0.6425	0.6475	0.6625	0.6625	0.655	0.655	0.6575	0.6675	0.6625	0.6575	0.665
0.9	0.645	0.65	0.6625	0.6625	0.6575	0.6525	0.66	0.665	0.665	0.67	0.6625
1.0	0.6475	0.6625	0.6675	0.66	0.66	0.6575	0.6625	0.6725	0.67	0.6625	0.6625
2.0	0.655	0.67	0.665	0.66	0.6575	0.6575	0.6625	0.665	0.6675	0.67	0.67

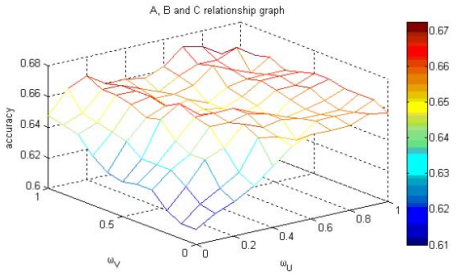


Figure 3. Rresult on different ω_U and ω_V

As showm in TABLE 1, we can get the best result when ω_U =0.7，ω_V =1.0.

C. *Experiments on ORL and Yale*

We have taken 10 groups experiment on ORL and Yale ,randomly selecting one image each person as training samples and using PCA， 2DPCA， $(2D)^2$PCA and $(W2DPC)^2$A respectively.

The results are as follows:

TABLE II. RESULT ON ORL

	Group1	Group2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10	Average
PCA	0.5875	0.575	0.5875	0.5525	0.575	0.5878	0.56	0.5025	0.5525	0.56	0.564
2DPCA	0.6075	0.5875	0.61	0.575	0.575	0.61	0.575	0.5375	0.575	0.5875	0.584
$(2D)^2$PCA	0.61	0.5875	0.6075	0.5875	0.61	0.6175	0.575	0.5225	0.5875	0.575	0.588
$(W2DPC)^2$A	0.6675	0.615	0.6275	0.6075	0.6375	0.65	0.625	0.56	0.635	0.635	0.626

TABLE III. RESULT ON YALE

	Group1	Group2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10	Average
PCA	0.719	0.776	0.724	0.719	0.757	0.583	0.616	0.594	0.687	0.704	0.688
2DPCA	0.745	0.807	0.77	0.719	0.776	0.6	0.637	0.6	0.704	0.719	0.708
$(2D)^2$PCA	0.751	0.812	0.757	0.745	0.782	0.637	0.64	0.583	0.719	0.748	0.714
$(W2DPC)^2$A	0.776	0.848	0.782	0.77	0.812	0.655	0.655	0.6	0.733	0.794	0.742

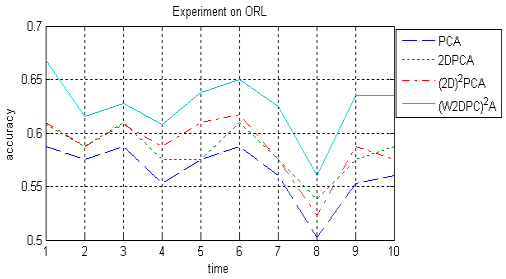


Figure 4. Experiment on ORL

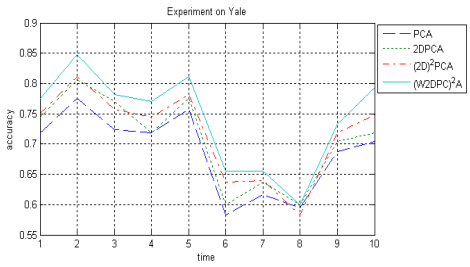


Figure 5. Experiment on Yale

According to TABLE 2 and 3,the result with$(W2DPC)^2$A is better than those with other methods.Among the results,the best one is 0.6675 on ORL and 0.848 on Yale.Illuminations in images in Yale alter obviously ,while expressions and imaging angles in images in Yale alter obviously.So $(W2DPC)^2$A to illuminations can get better robustness than to expressions or imaging angles.

VI. CONCLUSIONS

In this paper,$(W2DPC)^2$A is presentd improving upon $(2D)^2$PCA.According to the results of experiments, $(W2DPC)^2$A can get a higher recognition rate than PCA,2DPCA or $(2D)^2$PCA,when taking proper weights with single trainging sample.It takes much time to calculate because the number of blocks is a little big and the recognition effect is not very ideal because the contribution of each block to recognition is not good reflected. The focus of the future is the strategy of multi-feature fusion.

ACKNOWLEDGMENT

The research work is supported by National Natural Science Foundation of China(60872133), Beijing Municipal Natural Science Foundation(4092015),and Scientific Research Common Program of Beijing Municipal Commission of Education (KM201010772023).

REFERENCES

[1] XiaoJun Liu. Face recognition technology research[D].Beijing: Institute of Electronics,Chinsee Academy of Sciences,2001.

[2] JinYi Wu, DeLong Zhou. Survey of face recognition. Application Research of Computers, 2009， 26(9) : 3205-3209.

[3] Kirby M， Sirovich L.Application of the Karhunen-Loeve procedure for the Characterization of human faces.IEEE Transactions on Pattern Analysis And Machine Intelligence 1990， 12(l):103-108P

[4] Jian Yang,David Zhang.Two-Dimensional PCA:A new approach to appearance based face representation and recognition,IEEE Transactions on Pattern Analysis and Machine Intelligence,(2004),26(1):131-137

[5] Daoqiang Zhang, Zhihua Zhou. Two-direction Two-dimensional PCA for efficient face representation and recognition. Neurocomputing, 2005, 69 : 224-231P

[6] Rajkiran G,Vijayan K.Asari.An improved face recognition technique based on modular PCA approach.Pattern Recognition Letters， 2004， 25:429-436P

[7] YuJin Zhang.Image Engineering(I)Image Processing(Second Edition)[M].Beijing:TsingHua University Press, 2006.3， 371-372.

Optimization-based Multi-scale Texture Synthesis

Fan Xulong
College of Mathematics and Computer Science
FuZhou University
FuZhou 350002, China
dreamoflong@163.com

Chen Zhaojiong
College of Mathematics and Computer Science
FuZhou University
FuZhou 350002, China
chenzj@fzu.edu.cn

***Abstract*—Multi-scale texture synthesis is a rising research in recent years. There are two shortcomings: First, the exemplar design is very difficult; Second, Multi-scale texture synthesis uses pixel matching. Pixel matching can not capture the big structure information of the texture and it requires significant resources. A new multi-scale texture synthesis algorithm based optimization was proposed. The algorithm designs input exemplar textures using graph structure, constructs an exemplar-relation graph, adopts patch matching synthesis, introduces optimization to keep patch matching properly, and the size of the block can be adjusted and a jitter function was proposed. The algorithm absorbed the random characteristic of the pixel matching, and enhanced synthesis efficiency. Experimental results show that the method is effective.**

Keywords: Multi-scale; Texture synthesis; Optimization; Energy function; Exemplar-relation graph

I. INTRODUCTION

In reality, there are a large number of images, which have different structures at different scales, such as satellite or microscope images. Texture has this sort of multi-scale property we call multi-scale texture. Multi-scale texture synthesis aim to synthesize a texture that could preserve the multi-scale property of several input exemplar textures.

Traditionally, sampled-based texture synthesis usually employs one or more exemplar textures with same scale as input, and the synthesis result only keeps the same scale information of the input. Meanwhile, Multi-scale texture synthesis needs several different scale exemplar textures as input and the result could preserve the multi-scale property. The difficulties of multi-scale texture synthesis are that how to properly express the relationship of different scales, and how to exactly preserve the multi-scale property.

In this paper we present a simple algorithm to address the multi-scale texture synthesis problem. Our algorithm adopts exemplar-relation graph to characterize the relationship of different scales, uses up-sample step and coordinate jitter step to blur and randomize the exemplar textures. Finally, we employ an optimization step to make sure the similarity between input exemplar textures and the synthesis result.

A. Previous Work

There are a number of approaches for sampled-based texture synthesis. According to with constraints or without constraints, sampled-based texture synthesis can be classified into synthesis with constraints [1, 2] and synthesis without constraints. The result of synthesis with constraints includes not only the characters of input exemplars but also the shape of constrained texture. According to the synthesis result, sampled-based texture synthesis can be categorized into image texture synthesis and video texture synthesis [3, 4]. Video texture synthesis can be regarded as many image textures combining together with some rule, and showing them in a fixed period of time. According synthesis unit, sampled-based texture synthesis can be classified into pixel matching [5, 6] and patch matching[7-9]. Pixel matching is typically generated one pixel at a time in scan line or spiral order. For each pixel, the partial neighborhood already synthesized is compared with exemplar neighborhoods to identify the most likely pixels. Meanwhile, patch matching generates one patch at a time. Patch matching is faster than pixel matching and it's easier to capture big texel structure of exemplar texture.

Currently, multi-scale texture synthesis is a new research direction. [10, 11] are two typical articles focusing on multi-scale texture synthesis. The algorithm of [10] employs two different scale textures which describe a same object as input, and transforms their two different coordinates system into one. And then, it adopts parallel texture synthesis[12] algorithm to finish the synthesis. The parallel texture synthesis[12] is based on [13]. The algorithm of [12] improves synthesis quality using three new ideas: First, it uses Gaussian image stack to boost synthesis variety; Second, it adopts coordinate upsampling insteading of color upsampling; Third, it splits each neighborhood matching pass into several subpasses to improve correction. [11] is also based on [12]. But it effective uses of exemplar-relation graph to express the scale relationship of input multi-scale exemplar textures. [11] also can create infinite zooms and infinitely detailed textures that are impossible with current example-based methods.

B. Motivation

In article [10], if the number of input exemplar textures exceed two, the algorithm[10] needs complex coordinate transformations. In article [11], the author adopts exemplar-relation graph to express the scale relationship of input exemplar textures. And the synthesis algorithm[11] does not need coordinate transformations. However, the design of exemplar-relation graph is very difficult, a small fault will lead to unacceptable result. Further, both algorithms [10, 11] are based [12] which employs pixel matching. As we know,

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.63

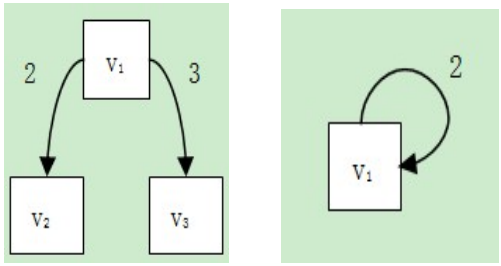
pixel matching algorithm has two shortcomings: First, it can not capture big texel; Second it's very slow. Although we can use GPU to accelerate the speed, the algorithm is complex and needs a lot of resources.

In order to overcome the deficiencies of multi-scale texture synthesis described above, in this paper, we develop a new optimization-based multi-scale texture synthesis algorithm. Our algorithm remedies the shortcoming that the synthesis result is overly dependent on the design of exemplar-relation graph in [11], and can capture big texel. Our algorithm is simple and easy to implement.

II. KEY CONCEPTS

A. Exemplar-relation Graph

In this paper, we introduced the exemplar-relation graph defined in [11], and adapted it to our algorithm. The exemplar-relation graph (V, E), is a reflexive, directed, weighted graph. Vertex set V={v1,v2,v3,…}, each element in V is an exemplar. E, denotes similarity relations between exemplars. The root, v1 serves as the coarsest-level starting point for synthesis. The weight of edge denotes the scale relationship between their vertexes. For example, e = (v1, v2, 3) denotes there is an edge from v1 to v2, the diameter of a pixel in exemplar v1 is three times larger than the diameter of a pixel in exemplar v2. Suppose, V_1 = {v1,v2,v3}, E_1 ={(v1,v2, 2),（v1,v3,3）}, figure 1(a) is its exemplar-relation graph. Figure 1(b) is a one vertex exemplar-relation graph, and the edge (v1, v1, 2) indicates that v1 is similar to a 2x scaling of itself.



(a) (b)

Figure 1: (a) Exemplar-relation graph (V_1, E_1), (b)One vertex exemplar-relation graph.

B. Texture Optimization for Example-based Synthesis

In article [14], the authors define a Markov Random Field (MRF)-based similarity metric for measuring the quality of synthesized texture with respect to a given input sample. In contrast to most example-based techniques that do region-growing, theirs is a joint optimization approach that progressively refines the entire texture. The energy optimization[14] is optimized using an Expectation Maximization[15] (EM)-like algorithm.

A unique aspect of algorithm [14] is that it is intermediate between pixel and patch based methods. The neighborhood size used to define texture energy determines the granularity at which synthesis is performed. They begin synthesis using large scale neighborhood sizes (32×32 pixels), which allows large scale texel of the texture to settle coherently in the output. And then, they refine the texture using smaller neighborhood sizes (16×16,8×8 , etc).

In this paper, we use the ideal of optimization algorithm mentioned above for reference, and properly change it to fit our multi-scale texture synthesis algorithm. We expand the search range of the neighborhood block according to super exemplar (will be addressed III), and the finding block will be more reasonable. We introduce the jitter function to add richness of the output texture. The jitter extension will be properly controlled based on synthesis level to prevent that the synthesis result is unacceptable.

C. Definition of Energy Function

We still adopt energy function that was defined in article [14]. We now describe the definition of energy function. The energy of a single neighborhood to be its distance to the closet neighborhood in the input. The total energy of the texture is then equal to the sum of energies over individual local neighborhoods in the texture.

Let Z denotes the only input sample to be used as reference; X denotes the texture over which we want to compute the texture energy. Firstly, let us introduce the following symbol definitions:

x , denotes the vectorized version of X; it is formed by concatenating the intensity values of all pixels in X.

N_p , denotes the neighborhood centered around pixel p in X.

x_p , denotes the sub-vector of x corresponds to N_p .

z_p , denotes the vectorized pixel neighborhood in Z whose appearance is most similar to x_p under the Euclidean norm.

Then, the texture energy over X to be:

$$E(x;\{z_p\}) = \sum_{p \in X^*} \| x_p - z_p \|^2$$

$X^* \subset$ X, the definition of X^* directly follows that article [14].

III. ALGORITHM DESCRIPTION

A. Algorithm Basic scheme

There are three steps in our algorithm. We now describe them detaily.

1) Upsample

We get a bigger texture image than original input exemplar texture by neighborhood-based upsampling.

2) Jitter

Next, we jitter the coordinates, using the following function to add richness of the output texture.

$Z_i(u) = Z_i(u + J(u)(\mathrm{mod}\, m))$, where $J(u) = \lfloor H(u) r_l + \binom{0.5}{0.5} \rfloor$

H: a hash function, $H : \mathbb{Z}^2 \to [-1,+1]^2$; Z_i denotes the i-th input exemplar texture, and i can not exceed the total numbers of input exemplar textures; u denotes a coordinate of a pixel in Z_i ; $Z_i(u)$ denotes the pixel value correspond to coordinate u; $0 \le r_l \le 1$, l denotes the synthesis level which will be discussed in section super exemplar, r_l denotes jitter intensity in synthesis level l , the higher of the synthesis level, the jitter intensity is smaller. At last, it reaches zero.

3) Energy minimization

The energy step takes the result of step 2 as input and uses energy minimization function to alter the input to recreate neighborhoods similar to those in the exemplar texture.

The algorithm is to iterate on the above three steps until gets the needed result texture. To explain our algorithm detaily, we will construct a super exemplar. Now, we describe how to construct a super exemplar and how to minimize energy over a texture.

B. Super Exemplar

The purpose of constructing a super exemplar is that indicates our algorithm's execution process and the scope of neighborhood searching that uses in every energy minimization step. Take figure 1 (a) for example, we explain how to construct a super exemplar. To construct a super exemplar, there are two key points: First, the size of the output texture, which denotes the number of the synthesis levels while running our algorithm. For example, we need a 128×128 pixels output texture and the input texture is 64×64 pixels, the number of the synthesis levels is two. Second, which exemplar textures include every synthesis level. Figure 2 is the figure 1 (a)'s super exemplar.

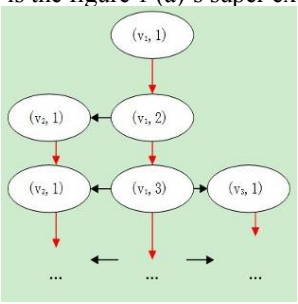


Figure 2 figure 1 (a)'s super exemplar

In this super exemplar, there are two parameters in the brackets. The first parameter is a vertex, corresponding to a vertex in the exemplar-relation graph; the second parameter denotes the synthesis level of the exemplar. In exemplar-relation graph, weight of an edge denotes the scale relation between the two vertexes. While, in the super exemplar, it denotes the energy minimization start point of the exemplar texture represented by the edge's end vertex. For example, edge (v_1, v2, 2) denotes the exemplar texture represented by v2 that begin to take part in energy minimization at synthesis level 2. For root exemplar texture in super exemplar, the start point is synthesis level one. In super exemplar, the red edges represent the step 1 and step 2 of our algorithm, and black edges represent neighborhood searching range of this synthesis level.

C. Energy Minimization

In each synthesis level except the synthesis level one of the root of the super exemplar, energy minimization step must be employed. And the detailed steps as follows:

Step 1: For $\forall p \in X^*$, we find a random neighborhood in the exemplar textures Z_1 , Z_2 , ..., Z_n which are in the same synthesis level. And then we use all these neighborhoods to construct z_p^0 .

Step 2: Computing $x^{n+1} = \arg\min_x E(x : \{z_p^n\})$

Step 3: Computing z_p^{n+1} . For $\forall p \in X^*$, we find the nearest neighborhood of x_p^{n+1} in the exemplar textures Z_1 , Z_2 ,..., Z_n which are in the same synthesis level.

Step 4: If z_p^{n+1} equals z_p^n , then $x = x^{n+1}$, energy minimization finishes. Else the algorithm goto Step 2.

In step 2, for each neighborhood x_p , we find the nearest neighborhood z_p , and then update x , minimize the energy. Since x changes after this update step, the set of closest input neighborhoods $\{ z_p \}$ may also change. Hence we need to repeat the step 2 and step 3 until met the stop conditions of step 4.

In our algorithm, the size of neighborhood is 32×32 pixels at the beginning. We could decrease the neighborhood size step by step.

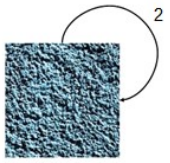
IV. EXPERIMENTAL RESULTS

In our experiments, the size of all input exemplar texture is 256×256 pixels, and the size of the synthesis result is 1024×1024 pixels. Both input textures and the result were zooming out for this paper.

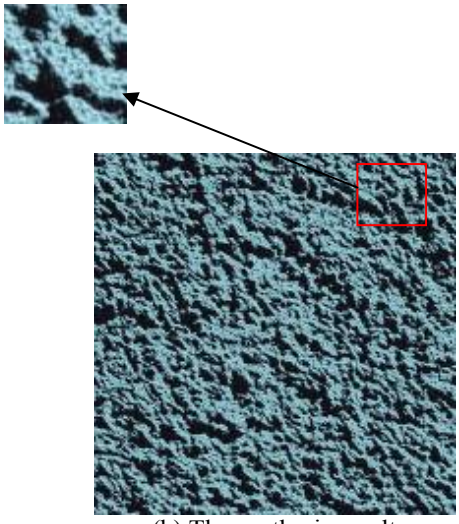
Figure 3, (a) is an exemplar-relation graph which only has one rock exemplar texture. (b) The synthesis result.

Figure 4, (a) is an exemplar-relation graph which has two exemplar textures. The two exemplar textures have two different scales and depict the same area. (b) The synthesis result.

Figure 5, (a) is an exemplar-relation graph which has three exemplar textures. It depicts mountains of southern Taiwan. (b) The synthesis result.

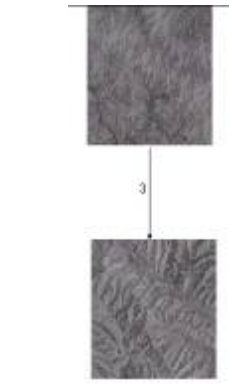


(a) Exemplar-relation graph

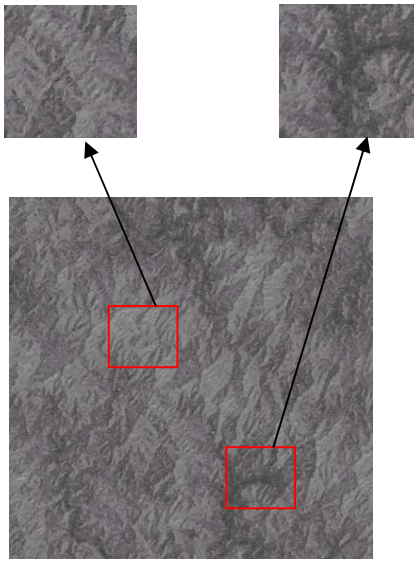


(b) The synthesis result

Figure 3 Rock texture and its synthesis result



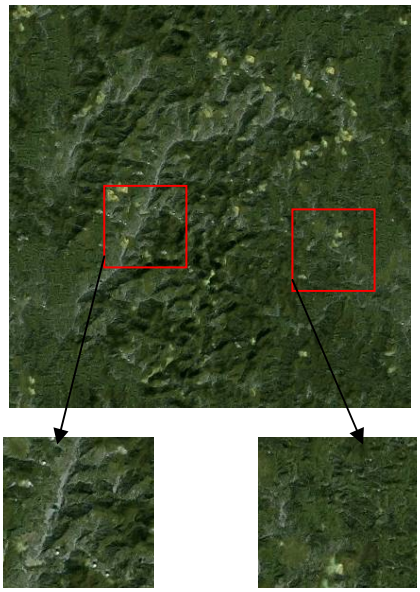
(a) Exemplar-relation graph



(b)The synthesis result

Figure 4 Satellite images and its synthesis result

(a) Exemplar-relation graph



(b)The synthesis result

Figure 5: Mountains of southern Taiwan and its synthesis result

V. CONCLUSION AND FUTURE WORK

The main contribution of this paper is to propose an optimization-based multi-scale texture synthesis algorithm, after the analysis of the current multi-scale texture synthesis algorithm. The algorithm of this paper employs exemplar-relation graph to describe the relationship of input exemplar textures, and upsample, jitter step to add richness of the output texture, energy optimization step to guarantee the rationality of the patch matching. Our algorithm absorbs the advantages of pixel matching and uses patch matching to capture big texel.

Our algorithm provides a good start point for the design of multi-scale texture. In the future, we expect that we could optimize the patch searching. Further, finding some rules for the exemplar-relation graph design is also an interesting aspect.

ACKNOWLEDGMENT

We would like to thank our anonymous reviewers for their insightful comments. This work was supported by NSFC under Grant No. 60805042, Natural Science Foundation of Fujian Province under Grant No. 2010J01329. We also thank Professor Ye Dongyi for his help and encouragement.

REFERENCES

[1] Ashikhmin Michael. Synthesizing natural textures[A]. In: 2001 ACM Symposium on Inter2active 3D Graphics[C], North Carolina: Research Triangle Park, 2001: 217~226.

[2] ZHENG Qiu-Xu, CHEN Zhao−Jiong Texture Design and Synthesis Based on Constrained Multiple Sources 2010,19(1)

[3] Szummer Martin, Picard Rosalind W. Temporal texture modeling[C]. International Conference on Image Processing, 1996, 3: 823~826.

[4] Bar-Joseph Z. Statistical learning of multi-dimensional textures[D]. The Hebrew University of Jerusalem, 1999.

[5] A.A.Efros and T.K.Leung. Texture Synthesis by Non-Parametrie Sampling[C]. IEEE International Conference on Computer Vision, 1999: 1033-1038.

[6] Liyi Wei, Marc Levoy. Fast Texture Synthesis using Tree-structured Vector Quantization. Proceedings of SIGGRAPH 2000, pp. 479-488, August 2000.

[7] Efors A, Freeman WT. Image quilting for texture synthesis and transfer [J]. Proc SIGGRAPH, 2001, (8):341-346.

[8] Liang L, Liu C, Xu Y Q, et al. Real-time texture synthesis by patch-based sampling [J]. ACM Transactions on Graphics, 2001, 20(3): 127-150.

[9] Xu Y, Guo B, and Shum H Y. Chaos mosaic: Fast and memory efficient texture synthesis. Tech.Rep.MSR-TR-2000-32[R], Microsoft Research, 2000.

[10] Sung-Ho Lee, Han-Wook Park, Jung Lee, et al. Multi-scale TextureSynthesis.http://cg-korea.or.kr/Journal/2008/14-2/%5B3%5DMulti-scale%20Texture%20Synthesis.pdf

[11] Charles Han, Eric Risser, Ravi Ramamoorthi, et al. Multiscale Texture Synthesis ACM Transactions on Graphics (Proceedings of SIGGRAPH 2008.27:51:1--51:8).

[12] Sylvain Lefebvre, Hugues Hoppe. Parallel controllable texture synthesis. In SIGGRAPH, 777−786［3］

[13] WEI, L.-Y., AND LEVOY, M. 2003. Order-independent texture synthesis.

[14] Vivek Kwatra, Irfan Essa, Aaron Bobick, et al. Texture optimization for example-based synthesis. In SIGGRAPH 2005, 795–802

[15] MCLACHLAN, G., AND KRISHNAN, T. 1997. The EM algorithm and extensions. Wiley series in probability and statistics. John Wiley & Sons.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Normalization Methods of SIFT Vector for Object Recognition

Hongxia Wang, Kejian Yang
School of Computer Science & Technology
Wuhan University of Technology
Wuhan, Hubei, China
e-mail:whx_green@163.com; ykjcgs@whut.edu.cn

Feng Gao, Jun Li
School of Computer
Wuhan University
Wuhan, Hubei, China
e-mail:gaofeng.sd@gmail.com; tianshia323@gmail.com

***Abstract*—In the clustering of keypoints, the ratio of inter-class distance and intra-class distance is enlarged using Normalized SIFT vectors. In order to significantly make the ratio larger, a new normalization method is proposed, which is termed the 'normalization method based on seedpoints'. In image preprocessing, for the removal of the object background, a SIFT-based entropy method is given, while the removal of the background, effectively retaining the SIFT keypoints of the object. Before clustering the keypoints, SIFT feature vectors are normalized using both Lowe's normalization method and the normalization method based on seedpoints. Experimental result of clustering shows that two kinds of normalized SIFT vectors, compared with the non-normalized SIFT vector, make the ratio of inter-class distance and intra-class distance larger. Especially, the SIFT vector normalized by the normalization method based on seedpoints more significantly enhances the distinctiveness between different classes, and makes the clustering better than Lowe's normalized SIFT vector.**

Keywords- SIFT; normalization; clustering

I. INTRODUCTION

Generic Object recognition is different from the recognition of specific, individual objects such as face recognition and fingerprint identification. Its research goals are to detect objects in images and to determine the object's categories. This requires that the semantic information of the image be obtained through the local feature extraction, clustering and other image processing. In [1], the performance of ten kinds of representative feature descriptors were compared, for example shape context, moment invariants, cross-correlation, SIFT (Scale Invariant Feature Transform) [2-3], etc. The experimental results in [1] showed that SIFT descriptor performed best in the cases of image rotation, scaling, geometric distortion, blurring, illumination change and image compression.

In this paper, because of its robustness, SIFT algorithm is used to extract the keypoints. Firstly, the image background is removed using the SIFT-based entropy method given in this paper, effectively retaining the edge keypoints of the object. Then the extracted SIFT feature vectors are normalized using two methods, the one is Lowe's and the other is proposed in this paper. Finally, in the clustering of the keypoints, to compare the effect on clustering of non-normalized and two kinds of normalized SIFT feature vectors. Experimental result illustrates that the proposed normalization method makes the clustering of keypoint data better than non-normalization and Lowe's normalization method.

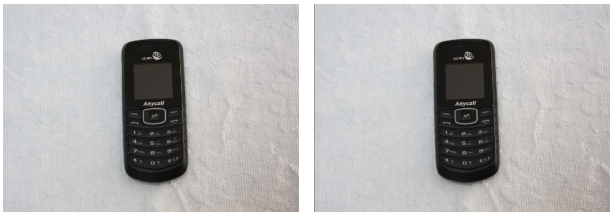
II. IMAGE PREPROCESSING

If keypoints extracted from the image are directly used for experiment, then experimental results will be affected. Therefore, images are must be preprocessed to remove the background in advance. Since some images in the standard image library have been processed, their entropy features have been changed. The images used in the experiment of this paper are taken by us, and the background is a white towel with uneven patterns.

For removal of the background, two methods, filter and entropy, are adopted, but neither is satisfactory. Thus a method termed 'SIFT-based entropy method' is given, that extracting SIFT keypoints firstly, and then comparing entropy mean with entropy threshold to retain SIFT keypoints of the object.

A. Filter Method

This method adopts the most common Butterworth Lowpass filter and median filter to remove the background. The images before and after processed are as shown in Fig. 1. The experimental results show that the effect is not obvious because the pattern background in the filtered image is still clear. Filter is sensitive to the backgrounds like salt-and-pepper noise, etc, but it's not suitable for the experiments in this paper.



(a) Original image (b) Filtered image

Figure 1. Comparison of images before and after filtered

B. Entropy Method

The experiment in A shows that simple filters can not effectively remove the background of the mobile phone, so entropy information of the image is used. The entropy histogram of an image is as shown in Fig. 2, which illustrates

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.62

that the main entropy information of the image is centralized on the right of the histogram, belonging to the mobile phone background, the regions around the peak of the left of the histogram belongs to the mobile phone.

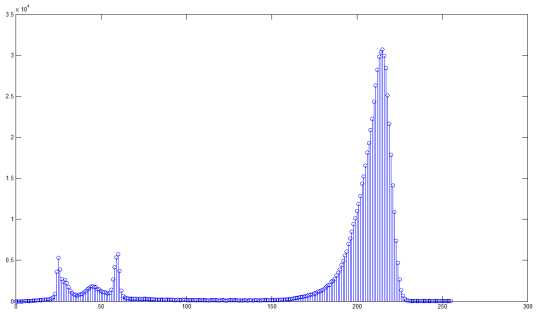


Figure 2. entropy histogram of an image

It is verified through doing experiments that taking $n \times n$ regions in the image, using entropy mean, E_k, of the $n \times n$ regions,

$$E_k = \frac{1}{n \times n} \sum_{i=1}^{n*n} g_i \tag{1}$$

where n is suggested to be an odd number, if E_k is greater than the threshold of the image histogram (median of two peaks), then it is proved that they belong to the background, otherwise, belong to the mobile phone.

The entropy method is good enough to clearly separate out the mobile phone, as shown in Fig. 3. However, this method has the problem of the extraction of keypoints on the edge of the phone, as shown in Fig. 4(a).

As experimental results of the above two methods are not satisfactory, the SIFT-based entropy is proposed in this paper.

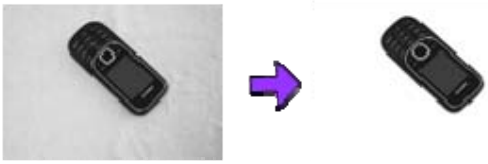


Figure 3. using entropy method to remove background

C. *SIFT-based Entropy Method*

The basic idea of this method: First, SIFT keypoints are extracted from the whole image with the background (SIFT algorithm will be presented in III). Then SIFT keypoint location is decided where the SIFT keypoint is. For a SIFT keypoint, S_k, to calculate entropy mean E_k of the $n \times n$ regions whose center S_k is as. Let q be the entropy threshold, then,

$$\begin{cases} E_k > q, & S_k \text{ is removed} \\ E_k \le q, & S_k \text{ is retained} \end{cases} \tag{2}$$

This method, compared with the entropy method, effectively retains the keypoints on the edge, as shown in Fig. 4(b).

(a) (b)

Figure 4. comparison of extracted SIFT keypoints using two methods of B and C after removing the background

III. SIFT DESCRIPTOR EXTRACTION

In SIFT algorithm, SIFT descriptor are assigned an image location, scale and orientation. Following is the stages of computation used to extract SIFT descriptor:

(1) Constructing the scale space of an image, detecting the scale-space local extrema.

To efficiently detect stable keypoint locations in scale space, Lowe[2] has proposed the scale space defined as (3), the difference-of-Gaussian (DoG) function convolves with the image:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{3}$$

(2) Locating the keypoints accurately, rejecting unstable keypoint candidates.

To reject unstable keypoints is to discard keypoints with low contrast and to eliminate edge keypoints, because keypoints with low contrast are sensitive to noise and the location of edge keypoints are poorly determined.

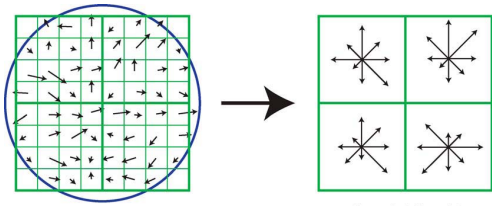
(3) Assigning a consistent orientation to each keypoint.

For achieving invariance to image rotation, each keypoint is assigned the dominant direction by using the gradient orientation histogram formed from the gradient orientations of sample points within a region around the keypoint.

(4) Creating the keypoint descriptors -- SIFT feature vectors.

In order to achieve orientation invariance, the coordinates of the descriptor and the gradient orientations are rotated relative to the keypoint orientation.

Fig.5(a) illustrates the gradient magnitude and orientation of each sample point around the keypoints with small arrows. A 4×4 set of samples are accumulated into gradient histograms to produce a seedpoint of eight directions, as shown in Fig. 5(b), with the length of each arrow corresponding to the sum of the gradient magnitudes with 4×4 sample regions.



(a) image gradients (b) keypoint descriptor

Figure 5. the keypoint descriptor

The figure shows a keypoint descriptor formed from 4 seedpoints, a 2×2 descriptor array. For the robustness in practical applications, the descriptor is formed from 16 seedpoints, a 4×4 descriptor array. It is a 16×8=128 element feature vector for each keypoint.

IV. NORMALIZATION METHODS OF SIFT VECTOR

The previous operations have made the SIFT descriptor, a feature vector of 128 elements, invariant to image scaling, rotation and other geometric distortion factors. Further, the vector is normalized to unit length in order to reduce the effects of illumination change and image grayscale change.

A seedpoint is to a sample position shifted up from 4×4 sample regions, relative to a sample point to a sample region, and a keypoint to 16×16 sample regions. Therefore, according to the size of sample regions for the normalization, the approach proposed by Lowe is termed normalization method based on keypoint and the approach proposed in this paper is termed normalization method based on seedpoints.

A. Normalization Method Based on Keypoint

The steps of normalizing the unit feature vector are as follows:

(1) The SIFT descriptor is extracted using SIFT algorithm:

$$D = \{d_1, d_2, \cdots, d_{128}\} \tag{4}$$

(2) If $d_i > 0.2$, then $d_i = 0.2$, where 0.2 is Lowe's empirical data.

(3) The vector D is normalized, $\overline{D}$,

$$\overline{D} = D / 0.2 \tag{5}$$

B. Normalization Method Based on seedpoints

(1) The SIFT descriptor is extracted using SIFT algorithm:

$$\begin{aligned} D &= \{d_1, d_2, d_3, \cdots, d_{128}\} \\ &= \{D_1, D_2, D_3, \cdots, D_{16}\} \end{aligned} \tag{6}$$

where D_i is the vector for the seedpoint,

$$\begin{aligned} D_i = \{&d_{8(i-1)+1}, d_{8(i-1)+2}, d_{8(i-1)+3}, d_{8(i-1)+4}, \\ &d_{8(i-1)+5}, d_{8(i-1)+6}, d_{8(i-1)+7}, d_{8(i-1)+8}\} \end{aligned} \tag{7}$$

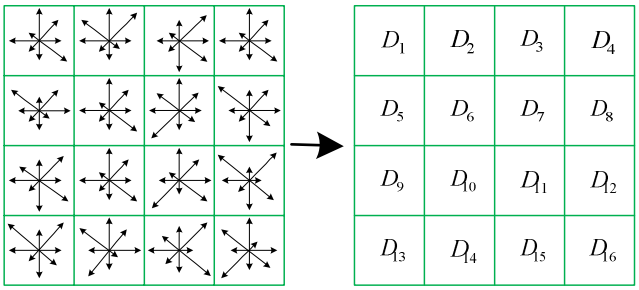


Figure 6. the seedpoint discriptors

(2) The maximum in D_i is found. Then it is assigned to $d_{\max}$.

(3) The vector D_i is normalized, $\overline{D_i}$,

$$\overline{D_i} = D_i / d_{\max} \tag{8}$$

thereby,

$$\overline{D} = \{\overline{D_1}, \overline{D_2}, \overline{D_3}, \cdots, \overline{D_{16}}\} \tag{9}$$

V. CLUSTERING EXPERIMENTS

In recent years, the research of object recognition has gradually shifted from low-level image features to image semantic information in order to eliminate the gap of semantic. At present, the Bag-of-Words method proposed by LiFeiFei[4] becomes one of hotspots of image semantic analysis method. The steps of this method are as follows:

(1) Extracting SIFT keypoints;
(2) Clustering the extracted SIFT keypoints;
(3) Producing the bag of words of a class;
(4) Classifying and recognizing the objects.

Before forming object words, the extracted SIFT keypoints are clustered. The result of clustering directly affects the formation of words, and ultimately decides the effect of object recognition. Therefore, aiming at clustering in the Bag-of-Words method, three kinds of SIFT descriptors before and after normalized are respectively used in the clustering experiments for comparison. Here, the K-Means method is adopted to cluster SIFT keypoints.

In order to compare the different effects on clustering using non-normalized SIFT vectors and SIFT vectors normalized by both normalization method based on keypoint and normalization method based on seedpoints, mobile phone images are taken as experimental object in this experiment, taking all SIFT keypoints of 42 mobile phones, which are 26702 keypoints in total.

Experimental steps are as follows:

(1) Clustering SIFT keypoints of the mobile phone images, they can be clustered into three parts: keypad part, outline part, other part, as shown in Fig. 7.

(a) Outline part (b) keypad part (c) other part

Figure 7. clustering of phone images

(2) After clustering, Euclidean distance formula, $d(x, y)$, is used to calculate inter-class distances and intra-class distances.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{10}$$

The clustering center of each class is taken as its descriptor. Intra-class distance and inter-class distance are defined as follows:

Intra-class distance: The mean of all the Euclidean distances of each point in one class and the clustering center of this class is termed intra-class distance.

Inter-class distance: The mean of all the Euclidean distances of each point in one class and the clustering center of other class is termed inter-class distance. Here, the distance of two clustering centers is not regarded as inter-class distance in order to reduce the deviation.

(3) Computing the ratio of inter-class distance and intra-class distance, as shown in Tab. 1.

Through observing the ratios in Tab. 1, it is found that the ratios of inter-class distance and intra-class distance increase obviously using the normalized SIFT feature vectors, compared with non-normalized SIFT vectors. Especially, SIFT vectors normalized by normalization method based on seedpoints make the ratios larger than normalized by normalization based on keypoint. Data show that SIFT vectors normalized by normalization method based on seedpoints more significantly enhance the distinctiveness between different classes, and are more conducive to the clustering of keypoints.

VI. CONCLUSION

The method termed normalization method based on seedpoints is proposed in this paper. And the normalization method proposed by Lowe is presented. After the background removal, SIFT descriptor extraction, SIFT vector normalization and clustering, experimental results show that the normalized SIFT feature vectors is better for the clustering of keypoints than the non-normalized ones, and the normalization method proposed in this paper makes the clustering better than the normalization method proposed by Lowe. It contributes to the formation of the semantic information, making the object recognition better.

REFERENCES

[1] Krystian Mikolajczyk, Cordelia Schmid, "A Performance Evaluation of Local Descriptors," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 10, pp. 1615-1630, 2005.

[2] David G. Lowe, "Object recognition from local scale-invariant features," Proceedings of International Conference on Computer Vision, vol. 2, no. 9, pp. 1150-1157, 1999.

[3] David G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[4] F.-F. Li, P. Perona, "A Bayesian hierarchical model for learning natural scene categories," In Proc. CVPR, 2005.

[5] Ying Liu, Dengsheng Zhang, Guojun Lu, Weiying Ma, "A survey of content-based image retrieval with high-level semantics," Pattern Recognition 40, pp. 262-282, 2007.

[6] Alex Pinz, "Object Categorization," Foundations and Trends in Computer Graphics and Vision, vol. 1, no. 4 pp. 255-353, 2005.

[7] Jasper R. R. Uijlings, "Real-Time Visual Concept Classification," IEEE Transactions on Multimedia, vol. 12, no. 7, November 2010.

TABLE I. THE RATIO OF INTER-CLASS DISTANCE AND INTRA-CLASS DISTANCE

	Non-normalization	***Normalization based on keypoint***	***Normalization based on seedpoints***
Keypad part - outline part	1.5721	1.6127	2.1156
Keypad part - other part	1.3385	1.7626	2.0476
outline part - other part	1.2789	1.8457	2.0713

A Novel Implementation of JPEG2000 MQ-Coder Based on Prediction

Jiangyi Shi, Jie Pang, Zhixiong Di
School of Microelectronic
Xidian University
Xi'an, China
e-mail: pj112233@126.com

Yunsong Li
National Key Laboratory of Integrated Service Networks
Xidian University
Xi'an, China
e-mail: ysli@mail.xidian.edu.cn

***Abstract*—The main process of the JPEG2000 is Discrete Wavelet Transform (DWT) and Embedded Block Coding with Optimized Truncation (EBCOT). MQ-Coder in the EBCOT Tier-1 is the bottleneck of the system for it is very difficult to be implemented by hardware with a high throughput. Therefore a novel architecture of MQ-Coder is proposed. Although this MQ-Coder processes only one single symbol for each clock cycle, it possesses the advantages of a higher frequency and smaller area which benefit from the prediction of the next index and the optimization of renormalization. The coder can achieve the throughput of 185.43 Msps.**

Keywords-JPEG2000; arithmetic encoder; MQ; throughput

I. INTRODUCTION

JPEG2000 is the image compression standard developed by ISO/IEC which has recently become one of the most popular image coding standards for its new functions such as progressive image transmission by quality or resolution, lossy and lossless compressions, region-of-interest encoding, and good error resilience[1]. So it can be widely used in many applications such as digital photography, printing, mobile applications, medical imagery, and Internet transmissions.

The key algorithms involved are Discrete Wavelet Transform (DWT) as its transform coding algorithm and Embedded Block Coding with Optimized Truncation (EBCOT) as its entropy coding algorithm. In the EBCOT algorithm, Tier-1 is a complete context-based entropy coder consisting of two processing stages: the bit-plane coding and the arithmetic coding (MQ-Coder). The bit-plane coding module generates the context of symbols while the arithmetic coder (MQ-Coder) performs entropy coding. The critical path of EBCOT exits in MQ-Coder, so if we shorten the critical path, both the frequency and the throughput will be improved. In this paper, a novel architecture is proposed to solve the above problems.

II. MQ ARITHMETIC CODING

A. Principle

The MQ coder adopted in JPEG2000 is Binary Arithmetic Coder (BAC). It processes the context (CX) together with the decision value (D) to obtain entropy encoded symbols which will be output to Tier-2 to form separated bit streams for each code block[4]. As shown in Fig. 1, the decision (D) and context (CX) are the inputs of the MQ encoder to produce compressed image data (CD) output.

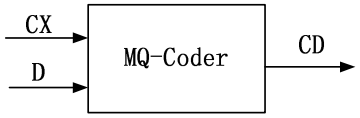


Figure 1. The model of MQ-coder

The coder comprises an IT (Index Table) which stores totally 19 contexts and 19 corresponding indexes. IT is looked up to obtain the index when one of the contexts comes into the coder. Then the index from IT will be used to look up another table called PET (Probability Estimation Table). There are 47 entries in the PET ROM[2], and each entry contains NLPS, NMPS, and Qe in JPEG2000 standard. NLPS and NMPS are used to update the PET after one CXD pair is coded. The flowchart[4] of MQ-Coder is shown in Fig.2.

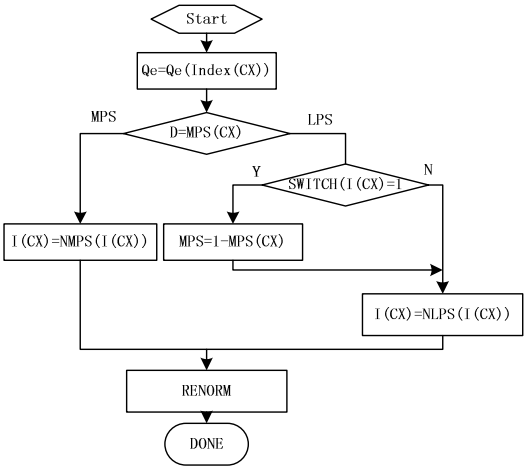


Figure 2. MQ-Coder procedure in JPEG2000

The MQ encoder selects the probability estimation Qe from PET according to the current index. The Qe is used to determine whether the PET will be updated and how the process of renormalization will be performed. With each binary decision (D), the current probability interval A is subdivided into two sub-intervals, and if the interval A is less than 0x8000[7] after coding the CXD pairs, it will be prevented from overflowing by left shifting the register A. It

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.8

should be noted that the roles of MPS and LPS can be switched when the new interval is less than the probability [3]: Qe > A – Qe (or A < 2Qe). The C register has the precision of 16 bits and is left shifted by the same amount as the A register during renormalization. Finally, the bits that are left shifted out of the C register form the Compressed Data (CD).

B. Challenge

The main challenge is that the above algorithm must be performed strictly step by step, in other words, the coding of current CXD pairs depends on the previous one, and the algorithm can't be implemented easily with the feature of hardware. As a result, the frequency of MQ-Coder is limited by the sequential coding steps in JPEG2000 Standard.

In recent years, many methods have been proposed to improve the coding speed, however, there are still some limitations in these designs. The MQ-Coder, for example, in paper [2] [3] [9] processes two symbols per clock, but the sequential relation between the two symbols is not broken, and the combinational logic has to calculate the two symbols one after another in one clock, so the long timing path produced by preceding process determines the design can't be implemented with a high frequency. Moreover, the hardware resource is wasted for there are two symbols to be dealt with. The MQ-Coder in [5] working at the frequency of 112 MHz improves the throughput, but it is far from enough.

For above reasons, we come back to the idea that the coder only processes one single symbol per clock with the advantages of smaller area and higher frequency. To achieve this target, some novel mechanisms which can obviously improve the frequency are adopted to make up the loss of throughput caused by less symbols per clock.

III. PROPOSED MQ-CODER ARCHITECTURE

A. Overview of the Architecture

As is shown in Fig.3, the proposed architecture of MQ-coder is designed in a three-stage pipeline scheme: Probability Estimation Stage, Renormalization Stage and Byte out Stage. In the first stage, a mechanism called the prediction of the index is adopted to help obtain the Qe faster. Meanwhile the equation_sel signal has to be calculated to determine which equation will be used in the Stage of renormalization. In the second Stage, register A and register C are updated according to the value of Qe and equation_sel signal. The adjustments in this stage are the simplified condition of renormalization and a partly parallel architecture. In the last stage, the high order bits of the register C are moved out to temporal buffer B as register C is left shifted the same amount as the register A. When the CT[6] reaches 0, the previous value of the register B is released to the output compressed data.

Probability Estimation Stage and Renormalization Stage are very significant for the coder, because the most of coding operations are performed in them. The efficiency of these designed stages influences the throughput of the coder directly.

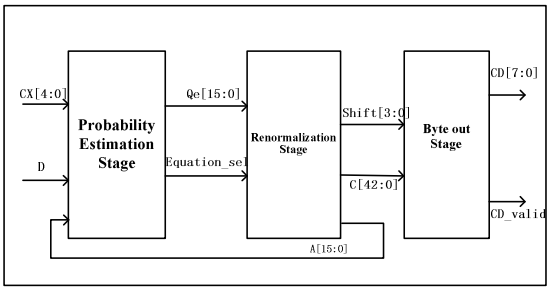


Figure 3. The architecture of the MQ-coder

B. Prediction of the Index

The index of PET is obtained from a table called IT (Index Table) according to the current CX. Usually, the IT should be updated before it is used, but if the result of the update can be predicted, the IT could be looked up without any waiting time before it is updated.

In order to meet the timing requirement of the circuit, the CX that comes into the coder must be firstly saved in register CX_R. The CX_R is set for two purposes: firstly, it can be used for the update of the IT which is one clock later than the current CX. If the current CX is not equal to the previous one, the later update can be ignored for it has no effect on the coding process. Secondly, it can be used to predict the next index of PET if the current CX is the same as the previous one. A precise process of the mechanism is shown in Fig. 4. Due to the existence of CX_R, we could insert a register called Qe_R which could cut off the long timing path between CX_R and A. Besides, during the process of update, NLPS and NMPS which can be reused in the process of prediction should be looked up from PET. The area is saved a lot for the reuse of CX_R, NLPS and NMPS.

```
If   A>Qe+0x8000
   Predict=I(CX1)
Else
   If D1==MPS
     Predict=nmps
   Else
     Predict=nlps
```

```
While CX1≠CX2
 Index=I(CX2)

While CX1==CX2

Index=P(D,MPS,Predict)
```

Figure 4. The mechanism of prediction.

C. Simplified Condition of Renormalization

As shown in Fig.5, conditions of the renormalization of register A are complicated. If D is equal to MPS(CX), MPS coding is utilized, if not, LPS coding is utilized, then we have to check the results of A>=Qe+0x8000 and A<2Qe [8] to decide which equation will be used. The whole process is sequential and will create a long timing path which could slow down the speed of the coder. It is found that the operation A=Qe works only under the specific condition of

D≠MPS(CX) and A>2Qe, so the conditions are concluded to one expression represented in Fig.6 where the complex condition is replaced by (D=MPS(CX))□(A>=2Qe). Because the calculation of equation_sel will be added to the path shown in Fig.7, the less time the coder consumes to gain the signal equation_sel, the faster the coder could work.

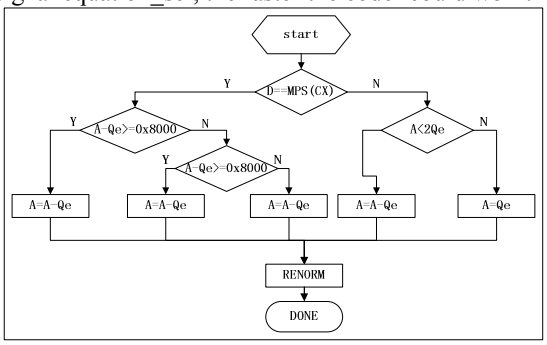


Figure 5. The renormalization of register A

```
If   (D=MPS(CX))⊙(A>=2Qe)
       Equation_sel=1
       A=A-Qe
       C=C+Qe
Else
       Equation_sel=0
       A=Qe
       C=C
```

Figure 6. The optimized conditions of renormalization

D. *Partly Parallel Architecture*

As is shown in Fig.6, because the logic of generating the signal Equation_sel is much more complicated than a 16 bits adder, the register A has to wait the arrival of the signal before updating itself, consequently a new critical path appears. To solve this problem, a partly parallel architecture is put forward which is expounded in Fig.7. This architecture enable the coder could calculate the A-Qe(or Qe) before getting the information of which equation will be used, in other words, the logic shown in Fig.7 is parallel with the logic of calculating (D==MPS(CX))□(A>=2Qe).

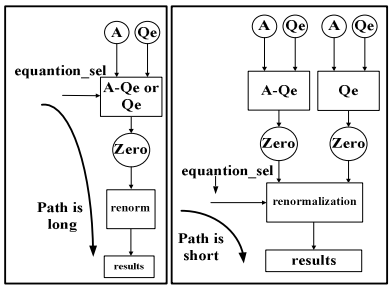


Figure 7. The partly parallel architecture

E. *Adjustment of the Encoder Registers*

The Encoder Register Structures are illustrated in Table 1. The register C is expanded from 28 bits[8] to 43 bits. The purpose of this change is that a specific situation can be handled when the register A reaches the value of 0x1. If it happens, the coder has to byte-out two times for the register C is left shifted 15 bits according to the rules of JPEG2000. The pipeline has to pause at the very moment C is left shifted 8 bits, therefore the efficiency of the coder is decreased. After the adjustment of the register C, it is not necessary for pipeline to stop to wait the process of Byte Out and the register C can be shifted 15 bits at one clock without any pause. Then the result of shift is released to Byte Out Stage, and it will be used to byte out two compressed data.

TABLE I. ENCODER REGISTER STRUCTURES

Register	
C	cbbb bbbb bsss xxxx xxxx xxxx xxxx
C expanded	000 0000 0000 0000 cbbb bbbb bsss xxxx xxxx xxxx xxxx
A	aaaa aaaa aaaa aaaa

IV. EXPERIMENTAL RESULTS

The proposed MQ-Coder is implemented with the virtex4 XC4VFX140 FPGA device using ISE 10.1i tool. The resource consumption and coding speed of the designed hardware are illustrated in Table 2 and the maximum frequency is 185.43 MHz. The sequence diagram of the coder is shown in Fig.8. From the table reported by the FPGA system, it can be seen that our encoder's throughput is higher than other architectures with less resource consumption, as is shown in Table 3.

TABLE II. THE RESOURCE CONSUMPTION OF THE CODER

Technology Parameter	One symbol
Number of Slices	495
Number of Flip Flops	382
Number of LUTs	893
Number of BRAMs	2
Maximum Frequency	185.43MHz

TABLE III. THE COMPARISON OF DIFFERENT ARCHITECTURES

Architecture	Frequency	Throughput
Architecture1[3]	53.92MHz	66.38 Msps
Architecture2[7]	48.3MHz	96.6 Msps
Architecture3[9]	50.1MHz	100.2 Msps
Our coder	185.43MHz	185.43 Msps

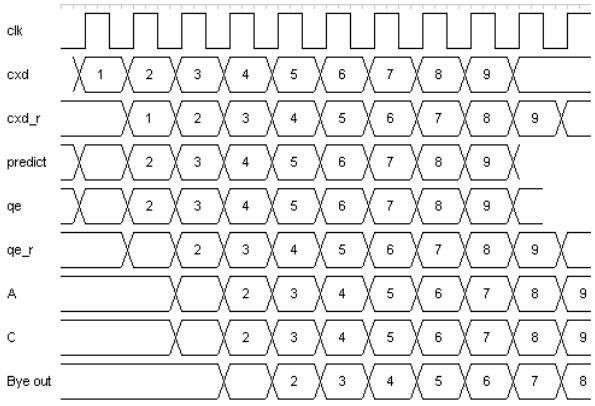


Figure 8. The sequence diagram of the coder

V. CONCLUSION

In this paper, optimized high speed, area efficient sequential MQ coder architecture is proposed. It adopts the mechanism called the prediction of the index, and simplifies the condition of renormalization. Moreover, a partly parallel architecture also contributes much to the high throughput of the MQ-Coder. The MQ coder is described with Verilog HDL at RTL, and the throughput and clock frequency have been markedly improved compared with other architectures. Throughput can achieve 185.43 Msps and clock frequency can reach 185.43MHz.

VI. ACKNOWLEDGE

Support for this work was provided by Natural Science Basic Research Plan in Shaanxi Province of China (2010JM8015).

REFERENCES

[1] Jen-Shiun Chiang, Chun-Hau Chang, Chang-Yo Hsieh, and Chih-Hsien Hsia, "High Efficiency EBCOT with Parallel Coding Architecture for JPEG2000", EURASIP Journal on Applied Signal Processing, 2006, pp. 1–14.

[2] PENG Zhou and ZHAO Baojun "High-throughout Hardware Architecture of MQ Arithmetic Coder", ICSP2010, 2010,pp. 430-433.

[3] Michael Dyer, David Taubman and Saeid Nooshabadi, "Concurrency Techniques for Arithmetic Coding in JPEG2000[J]". IEEE Transactions on Circuits and Systems for Video Technology, 2006, vol.53, pp. 1203–1213.

[4] Nopphol Noikaew and Orachat Chitsobhuk, "Dual Symbol Processing for MQ arithmetic coder in JPEG2000", 2008 Congress on Image and Signal Processing, 2008, pp. 521-524.

[5] Taoufik Saidani, Mohamed Atri and Rached Tourki, "Implementation of JPEG 2000 MQ-Coder", 2008 International Conference on Design & Technology of Integrated Systems, 2008, pp. 1-4.

[6] Kishor Sarawadekar and Swapna Banerjee, "LOW-COST, HIGH-PERFORMANCE VLSI DESIGN OF AN MQ CODER FOR JPEG 2000" ICSP2010, 2010, pp.397-400.

[7] Kai Liu, Yu Zhou b and Yun Song Li, "A high performance MQ encoder architecture in JPEG2000", INTEGRATION, the VLSI journal, 2010, pp. 305–317 .

[8] "JPEG2000 part 1 020719 (final publication draft)," ISO/IEC JTC1/

[9] A. K. Gupta, S. Nooshabadi and D. Taubman, "Realizing low-cost high-throughput general-purpose block encoder for JPEG2000[J]". IEEE Transactions on Circuits and Systems for Video Technology, 2006, vol.16, pp. 843–858.

A Novel Unsupervised Optimal Discriminant Plane

Su-Qun Cao
Faculty of Mechanical Engineering
Huaiyin Institute of Technology
Huai'an 223003, China
E-mail: caosuqun@126.com

Shi-Tong Wang
School of Digital Media
Jiangnan University
Wuxi 214122, China
E-mail: wxwangst@yahoo.com.cn

Quan-Yin Zhu
Faculty of Computer Engineering
Huaiyin Institute of Technology
Huai'ai 223003, China
E-mail: zqy@hyit.edu.cn

Abstract—**Optimal discriminant plane is an important feature extraction method in machine learning, pattern recognition and image processing, etc. Based on this, Zhao et al. presented a hybrid optimal discriminant plane method by integrating uncorrelated and orthogonal discriminant vectors together. As the same to optimal discriminant plane, it is a supervised feature extraction method. This paper extends Zhao's optimal discriminant plane to the unsupervised pattern. With the orthogonal constraint and the conjugated orthogonal constraint of the fuzzy total-class scatter matrix, two vectors which maximize the fuzzy Fisher criterion are obtained. These two vectors make up a novel unsupervised optimal discriminant plane. Experimental results on UCI datasets show its efficiency.**

Keywords- Optimal discriminant plane; Feature extraction; Unsupervised pattern

I. INTRODUCTION

Fisher's linear discriminant (FLD) analysis is a very useful technique for supervised pattern analysis [1]. It attempts to find out an optimal feature space using the Fisher criterion function and it can extract the discriminant features from the data sets which best separate two or more classes of data samples. Sammon [2] extended FLD by finding out an optimal discriminant plane (ODP) by computing orthogonal discriminant vectors from data. Jin et al. [3] proved that two discriminant vectors in ODP are essentially correlated and then proposed the uncorrelated optimal discriminant plane (UODP) method which is more powerful than ODP in such applications as face recognition. Furthermore, Zhao et al. presented a hybrid optimal discriminant plane method by integrating uncorrelated and orthogonal discriminant vectors together [4].

Zhao's optimal discriminant plane (ZODP) is based on Fisher discriminant criterion. It can only be used in a supervised way. Such a limitation makes it very inappropriate for unlabeled data. This study here attempts to extend it to the unsupervised case.

The rest of this paper is organized as follows. Section 2 introduces Fisher criterion and ZODP. Section 3 presents fuzzy Fisher criterion and proposed optimal discriminant plane. Some tests are performed in later section.

II. FISHER CRITERION AND ZHAO'S OPTIMAL DISCRIMINANT PLANE

Assume that X is a d-dimensional sample set with N elements belonging to c classes, S_b is the between-class scatter matrix, S_w is the within-class scatter matrix and S_t is the total-class scatter matrix. The Fisher criterion can be defined as follows:

$$J_{FC} = \frac{\omega^T S_b \omega}{\omega^T S_w \omega} \tag{1}$$

where ω is an arbitrary vector in d-dimensional space. The Fisher optimal discriminant vector ω_1 corresponding to maximum of $J_{FC}(\omega)$, can be derived as the eigenvalue of the following eigen-system equation:

$$S_b \omega_1 = \lambda S_w \omega_1 \tag{2}$$

Based on the Fisher' vector ω_1, Zhao et al. [4] obtained the second direction ω_2 which maximizes the Fisher criterion $J_{FC}(\omega)$ with the orthogonal constraint in the following Eq.(3) and the conjugated orthogonal constraint in the following Eq.(4). Thus, ω_1 and ω_2 make up

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.10

Zhao's optimal discriminant plane (ZODP).

$$\omega_2^T \omega_1 = 0 \tag{3}$$

$$\omega_2^T S_t \omega_1 = 0 \tag{4}$$

III. Fuzzy Fisher Criterion and Proposed Optimal Discriminant Plane

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}:

$$S_{fw} = \sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij}^m (x_j - m_i)(x_j - m_i)^T \tag{5}$$

the following fuzzy between-class scatter matrix S_{fb}:

$$S_{fb} = \sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij}^m (m_i - \bar{x})(m_i - \bar{x})^T \tag{6}$$

and the following fuzzy total scatter matrix S_{ft}:

$$S_{ft} = \sum_{i=1}^{c} \sum_{j=1}^{N} u_{ij}^m (x_j - \bar{x})(x_j - \bar{x})^T \tag{7}$$

Thus, we can define a novel fuzzy Fisher criterion as follows:

$$J_{FFC} = \frac{\omega^T S_{fb} \omega}{\omega^T S_{fw} \omega} \tag{8}$$

By Maximizing J_{FFC}, several formulas as follows are obtained [5].

$$S_{fb}\omega = \lambda S_{fw}\omega \tag{9}$$

where λ is taken as the largest eigenvalue.

$$m_i = \frac{\sum_{j=1}^{N} u_{ij}^m (x_j - \frac{1}{\lambda}\bar{x})}{\sum_{j=1}^{N} u_{ij}^m (1 - \frac{1}{\lambda})} \tag{10}$$

$$u_{ij} = F_1 / F_2 \tag{11}$$

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

$$\begin{aligned} &u_{ij} = 1 \text{ and } u_{i'j} = 0 \text{ for all } i' \neq i, \text{ if} \\ &\omega^T (x_j - m_i)(x_j - m_i)^T \omega \\ &\leq \frac{1}{\lambda} \omega^T (m_i - \bar{x})(m_i - \bar{x})^T \omega \end{aligned} \tag{12}$$

According to Zhao et al.'s ideas [4], we can obtain the second ω_2 with both the orthogonal constraint and conjugated orthogonal constraint by Theorem 1. Thus, ω_1 and ω_2 make up a novel unsupervised optimal discriminant plane (NUODP).

Theorem 1. The second discriminant vector ω_2 which maximizes the fuzzy Fisher criterion $J_{FFC}(\omega)$ with both the following orthogonal constraint and conjugated orthogonal constraint:

$$\omega_2^T \omega_1 = 0 \text{ and } \omega_2^T S_{ft} \omega_1 = 0 \tag{13}$$

is the eigenvector corresponding to maximal eigenvalue of the following eigen-system:

$$PS_{fb}\omega_2 = \lambda S_{fw}\omega_2 \tag{14}$$

where

$$P = I - S_{ft}\omega_1 M^{-1} N - \omega_1 A^{-1} B$$

$$M = (\omega_1^T S_{fw}^{-1}\omega_1)(\omega_1^T S_{ft} S_{fw}^{-1} S_{ft}\omega_1) - (\omega_1^T S_{ft} S_{fw}^{-1}\omega_1)(\omega_1^T S_{fw}^{-1} S_{ft}\omega_1)$$

$$N = (\omega_1^T S_{fw}^{-1}\omega_1)(\omega_1^T S_{ft} S_{fw}^{-1}) - (\omega_1^T S_{ft} S_{fw}^{-1}\omega_1)(\omega_1^T S_{fw}^{-1})$$

$$A = (\omega_1^T S_{fw}^{-1} S_{ft}\omega_1)(\omega_1^T S_{ft} S_{fw}^{-1}\omega_1) - (\omega_1^T S_{ft} S_{fw}^{-1} S_{ft}\omega_1)(\omega_1^T S_{fw}^{-1}\omega_1)$$

$$B = (\omega_1^T S_{fw}^{-1} S_{ft}\omega_1)(\omega_1^T S_{ft} S_{fw}^{-1}) - (\omega_1^T S_{ft} S_{fw}^{-1} S_{ft}\omega_1)(\omega_1^T S_{fw}^{-1})$$

Here, I is the identity matrix.

Based on the fuzzy Fisher vectors ω_1, ω_2, the following linear transformation from d-dimension to 2-dimension can be defined:

$$Y = \begin{bmatrix} \omega_1^T \\ \omega_2^T \end{bmatrix} X \tag{15}$$

In summary, a novel feature extraction algorithm based on NUODP can be described as follows:

Algorithm NUODP

Step 1. Set the given threshold ε and the number of classes c. Use K-means to initialize $U = [\mu_{ij}]_{c\times N}$, $m = (m_1, m_2, ..., m_c)$;

Step 2. Compute S_{fw}, S_{fb}, S_{ft} using Eq.(5), Eq.(6), Eq.(7) respectively;

Step 3. Compute the largest eigenvalue λ and the corresponding ω using Eq.(9);

Step 4. Update m_i and μ_{ij} using Eq.(10), Eq.(11) and Eq.(12) respectively;

Step 5. Compute J_{FFC} using Eq.(8). If the difference between J_{FFC} and its last value larger than ε, back to Step 2.

Step 6. Compute the second fuzzy Fisher vector ω_2 using Eq.(14). The obtained ω_1 and ω_2 make up NUODP.

Step 7. Perform the linear transformation from the original data space to the 2-dimensional feature space using Eq. (15).

IV. EXPERIMENTAL RESULTS

In this experiment, two benchmarking UCI datasets Iris, and Wdbc [6] were chosen to test the performance of feature extraction methods. Table 1 illustrates the basic information of the datasets.

TABLE I. THE BASIC INFORMATION OF DATASETS

Datasets	Classes	Samples	Features
Iris	3	150	4
Wdbc	2	569	30

Supervised feature extraction method ZODP and unsupervised feature extraction methods including principal component analysis (PCA) and NUODP have been applied to reduce the dimensions of these datasets. Classification experiments have been done on the reduced-dimension datasets. 1-Nearest Neighbor (1-NN) and LibSVM [7] methods on the reduced-dimension datasets were performed by leave-one-out cross validation (LOOCV) in this experiment. LOOCV involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. Table 2 lists the classification accuracy of 1-NN and LibSVM methods on the reduced-dimension datasets extracted by ZODP, PCA and NUODP using LOOCV.

TABLE II. CLASSIFICATION ACCURACY OF 1-NN, LIBSVM FOR REDUCED-DIMENSION DATASETS EXTRACTED BY ZODP, PCA AND NUODP USING LOOCV

Datasets	ZODP		PCA		NUODP	
	1-NN	*LibSVM*	*1-NN*	*LibSVM*	*1-NN*	*LibSVM*
Iris	0.96	0.973	0.953	0.96	0.953	0.98
Wdbc	0.963	0.953	0.914	0.627	0.946	0.926

From Table 2, we can see:

1 Supervised feature extraction method ZODP has the best average performance in the experiments. This means

that the class information of samples is important to feature extraction.

2 Whether classified by 1-NN or LibSVM, the classification accuracy of the reduced-dimension Iris and Wdbc datasets extracted by NUODP is superior to those extracted by PCA. This shows the proposed method's efficiency.

ACKNOWLEDGEMENTS

This work is supported by Natural Science Major Basic Research Program of Jiangsu Colleges and Universities, China (09KJA460001, 10KJA460004), Huaian Science and Technique Program (HAG2010039, HAG2010062) and Qing Lan Project of Jiangsu Province.

REFERENCES

[1] R.A. Fisher, "The use of multiple measurements in taxonomic problems". Ann. Eugenics, vol. 7, 1936, pp.178-188.

[2] J.W. Sammon, "An optimal discriminant plane", IEEE Trans on Computers, vol.19, no.9, 1970, pp.826-829.

[3] Jin Zhong, Lou Zhen, Yang Jingyu, "An optimal discriminant plane with uncorrelated features," PR & AI, vol.12, no.3, 1999, pp.334-338.

[4] Zhao Haitao, Jin Zhong, "An improved optimal discriminant plane," Journal of Nanjing University of Science and Technology, vol. 24, no. 1, 2000, pp. 88-92.

[5] Su-Qun Cao, Shi-Tong Wang, Xiao-Feng Chen et al., "Fuzzy fisher criterion based semi-fuzzy clustering algorithm," Journal of Electronics & Information Technology, vol.30, no.9, 2008, pp.2162-2165.

[6] C.L. Blake and C.J. Merz, UCI repository of machine learning databases,http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.

[7] Chih-Chung Chang, Chih-Jen Lin, LIBSVM : a library for support vector machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

Generation of Panoramic View from 360 ° Fisheye Images Based on Angular Fisheye Projection

Xiao Shouzhang
College of Information Science and Engineering of Northeastern University
Shenyang City, Liaoning Province, China
xiaoshou-zhang@163.com

Wang Fengwen
Northeastern University at Qinhuangdao
Qinhuangdao City, Hebei Province, China
wfw0335@126.com

***ABSTRACT*—this paper proposes a algorithm to generate the panoramic view from 360 ° fisheye images based on angular fisheye projection model. Firstly, the image plane will be mapped to the view plane, the sphere plane, to obtain the coordinates of spherical points .then use the angular fisheye projection model to get fisheye images' radius. Secondly, calculate the coordinates of fisheye images' points with the radius of fisheye images and the coordinates of spherical points. Finally, map the points of fisheye images to the target plane by the way of backward mapping. Experiments show that this algorithm is simple and fast, and can obtain the ideal images.**

Keywords: Angular Fisheye Projection, Fisheye Images, Panorama Expand, Backward Mapping

I. INTRODUCTION

Fisheye lens can capture the scene for large field of view. When fisheye lens toward front, the angle of view can reach about 180° . As was hemispherical fisheye lens, the horizon can reach 360° or so, when the shooting is down or upward. Therefore, the characteristics of its super view has been widely applied to many areas of computer vision, such as Virtual reality technology, Robot Navigation, Geometry image-based rendering, Visual Surveillance, Shooting venue and so on. Especially in the field of virtual reality, to get a very vivid visual world, panoramic image is essential. Panoramic image of the past is to be the direct use of the general camera or video camera to get real image sequences, and then obtained using the way of image mosaic. This method requires original image as more, so each image has more complex information, panoramic image not only the speed of generation fails to meet requirements, and the result is not satisfactory. Because of the large fish-eye perspective of the image, generally using two or several, or even one image can meet the requirements of panoramic images. However, due to serious distortion of fish-eye images, it gives generation of panoramic image inconvenience.

This study is the algorithm to obtain panoramic image by expanding one 360° fisheye image. Traditionally, the main methods of generation of the panoramic image by a fisheye image are Half-cube-based texture mapping expand and panoramic expand based on the Cylinder Model. Both methods are based on hemispheric fisheye projection model. Result of Half-cube-based texture mapping expand not only does not match with actual density, and distribution in five plane is not conducive to observation. Panoramic expand based on the Cylinder Model is too slowly, and result is not satisfactory. Both algorithms have their own blind spots to start and cannot be expand points of the central region of the image. As these deficiencies, this paper presents an algorithm, based on the angular fisheye projection model using 360° circular fisheye image to generate panoramic images.

II. THE ANGULAR FISHEYE PROJECTION MODEL

The angular fisheye projection model is defined that distance from point in fisheye image to the center of fisheye image and direction angle from the connection line of corresponding point on the sphere and the spherical center point to the optical axis of the camera are proportional. By definition, in angular fisheye projection model, the whole fisheye image pixel distribution is uniform. The angle of view of angular fisheye projection model can up to 360° , but at the outermost edge of the 360° angular projection image, just one point is projected onto the image plane, it is not easy to understand. As shown in Figure 1. , the angle of angular fisheye projection model is 180° . It can be observed that the distance from projection point on the graph to the graphics center and the angle from the point of the projection sphere to main optical axis are directly mapped. 180° angular projection of fisheye lens is only half of the scene projected onto the image plane, if the angle is 360° , all of the scenes can be obtained.

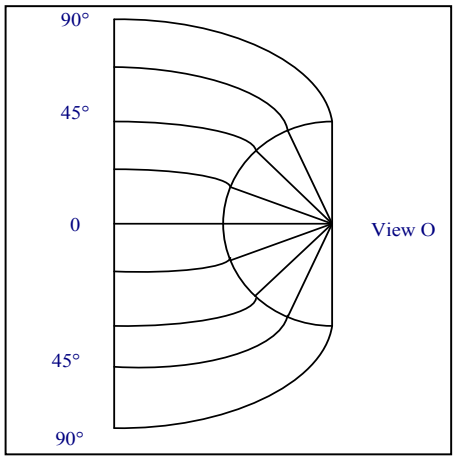


Figure 1. The angular fisheye projection model

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.49

III. ALGORITHM THEORY

Place the pixel on each radius of 360° fisheye image parallel to the target plane in turn, then the ideal panorama can be expanded. Here, Figure 2(a) shows the fisheye image model, and Figure 2(b) shows the mapping plane, also call it target plane. Horizontal axis of the target plane is the longitude of spherical coordinates, vertical axis is latitude. A radius (OQ) of fisheye image corresponds to the line (O'Q') on the target plane. Pixel value is also one to one. Get the mapping formulas of coordinates between fisheye image and the target plane, then, Figure 2(a) can be expanded into the desired panorama.

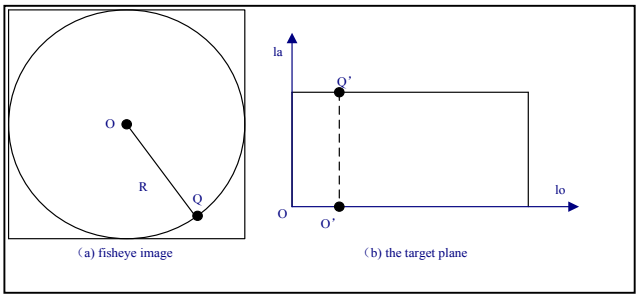


Figure 2. pixel mapping diagram

IV. STEPS OF ALGORITHM

The ultimate goal is to derive the mapping formulas of coordinates between fisheye image and the target plane.

Two methods of deriving the mapping formulas are forward mapping method and backward mapping method. Forward mapping method is that calculate the coordinates of points on target plane with the coordinates of points on fisheye image by a series of transformations, then, assign pixel value of this point to the destination point. This is the most general idea of the mapping. Backward mapping method is that starting from the target plane, get the coordinates of corresponding point on fisheye image to assign the pixel value.

The size of the target image getting by forward mapping method may not match the size of fish eye image, and result in a lot of pixel gap. So, in this paper, use backward mapping method.

A. Algorithm flow chart

Algorithm flow chart is showed in Figure 3.

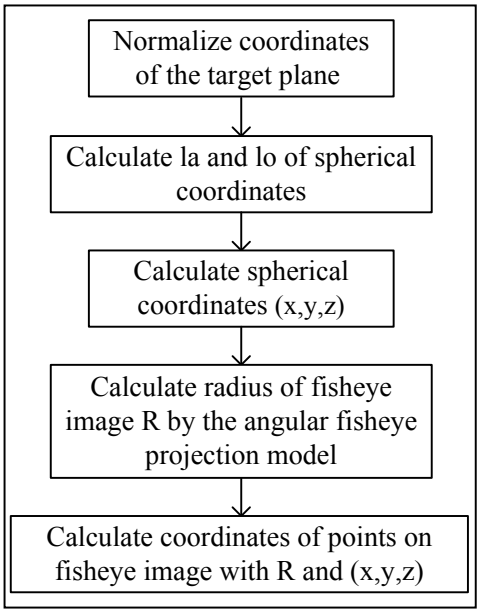


Figure 3. algorithm flow chart

B. Normalize the target plane coordinates

Let coordinates of the target plane be expressed as (i, j), coordinates after normalizing is expressed as (u, v). After normalizing the target plane, the mapping function formulas of two planes are defined as:

$$u = 2i / width - 1 \quad (1)$$

$$v = 2j / height - 1 \quad (2)$$

C. Calculate spherical coordinates (x,y,z)

Calculate longitude lo and latitude la with (u, v),

$$lo = (1-u) \times \frac{\pi}{2} \quad (3)$$

$$la = (1-v) \times \frac{\pi}{2} \quad (4)$$

Due to the value range of u and v is [-1, 1], the value range of lo and la is [0, Π]. As shown in Figure 4, la is the angle between OP and the direction of X-axis, lo is the angle between OP and the direction of Y-axis.

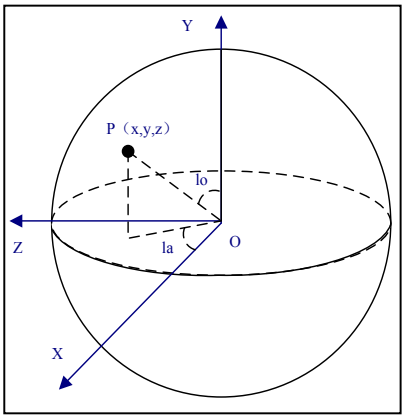


Figure 4. latitude and longitude diagram

Assume radius of the sphere is 1, with lo and la, formulas of coordinates (x, y, z) of spherical point P can be defined as：

$$x = \sin(la) \times \cos(lo) \tag{5}$$

$$y = \cos(la) \tag{6}$$

$$z = \sin(la) \times \sin(lo) \tag{7}$$

D. Calculate radius of fisheye image R and coordinates of points on fisheye image

The angular fisheye projection model is defined that distance from point in fisheye image to the center of fisheye image and direction angle from the connection line of corresponding point on the sphere and the spherical center point to the optical axis of the camera are proportional. In this algorithm, Fisheye lens corresponds to hemispherical, so the angle of view of fisheye lens is 180° , and the view size of the length and width of projection plane corresponding is π. It can be obtained the distance of line on the projection plane corresponding to the view of fisheye image. Let distance express as Dist, and width of fisheye image express as Wid. So,

$$Dist = Wid / \pi \tag{8}$$

With spherical coordinates (x, y, z), the length of the hypotenuse of corresponding to the point (x, y) is defined as r.

$$r = \sqrt{x\text{^}2 + y\text{^}2} \tag{9}$$

As shown in Figure 5, direction angle from the connection line of corresponding point (x, y, z) on the sphere and the spherical center point to the main optical axis of the camera is phi.

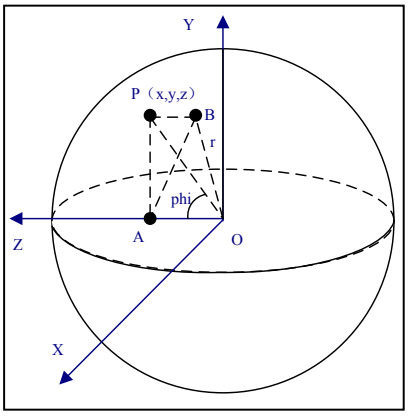


Figure 5. r and radian of r corresponding diagram

It can be obtained from Figure 5, phi is defined as：

$$phi = a\cos(z) \tag{10}$$

According to the definition of angular fisheye projection, the formula of R is

$$R = dis \tan ce \times phi \tag{11}$$

From the above, coordinates of the point (x_src, y_src) on fisheye image can be calculated by

$$x\_src = R \times \frac{x}{r} \tag{12}$$

$$y\_src = R \times \frac{y}{r} \tag{13}$$

Then we have derived the mapping function from point (i, j) of the target plane to the point (x_src, y_src) on fisheye image. At last, it only need to assign the pixel value for point (i, j).

E. Experimental results and analysis

As shown in Figure 6, the result is ideal. Firstly, it can expand whole scene on fisheye image, including the scene in center of image, and no blind area. But experimental result in Figure 7 shows that algorithm based on Cylinder Model cannot do this. Secondly, the speed of calculation is faster. Take Figure 6(a) for example, experimental environment is that experimental software is MATLAB 7.6.0(R2008), the computer is Intel(R) Pentium (R) 4, and CPU is 3.00GHZ 2.99GHZ 1.00GB, the size of picture is 400×400 (pixels). The time of panoramic expand based on the Cylinder Model is 24.96s, but the angular fisheye projection model just cost 0.73s. Thirdly, program structure is simple, logic is clear and concise. Finally, the pixel distribution of panorama expanded is uniform, and the definition of image meets the requirements of experiments.

REFERENCES

[1] Miyamoto K.. Fish eye lens .Journal of Optical Society of America, 1964：54(8)：1060-1064

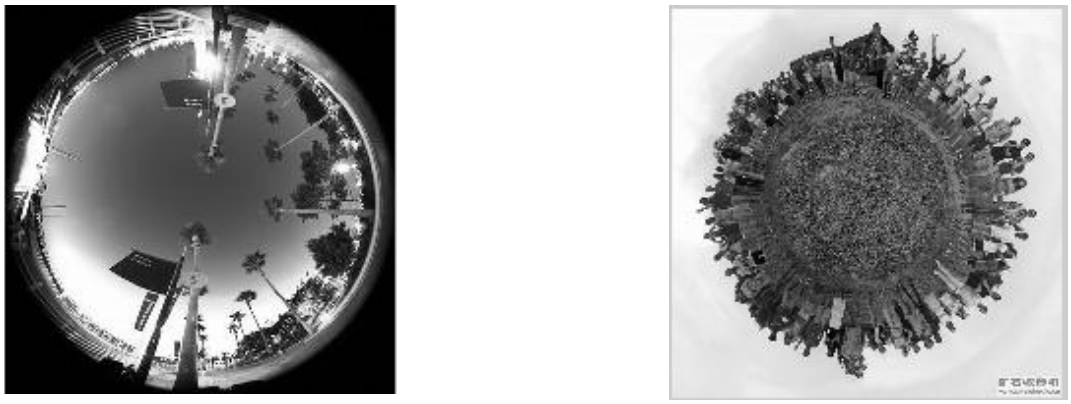
[2] Wang Junjie，Xu Xiaogang，Hu Yunfa. Fisheye projection in the Application of Virtual Reality [J]. Small micro-computer system，2004,25(2)：287-290.

[3] Xue Juntao，He Huaiqing. A texture mapping technology using fish-eye lens method for rapid correction [J]. Electronic technology Applications，2006，8：38-41.

[4] Zhou Hui，Luo Fei，Li Huijuan，Feng Bingshu.. Research of fish-eye image correction method based on Cylinder model [J]. Computer Applications，2008.10，28（10）：2664-2666.

[5] Lin Yufeng，Zhu Qidan，Wu Zixin. Panoramic cylinder theory of visual images expand algorithm and its improvement [J]. Applied Science and Technology，2006，33（9）：4-6.

[6] Paul Bourke .Computer Generated Angular Fisheye Propjections,http://astronomy.swin.edu.au/,2001



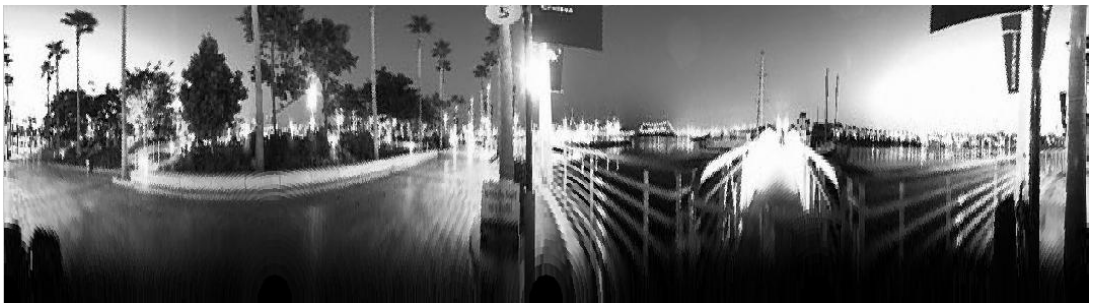
（a） original picture

（b） experimental result 1

（c） experimental result 2

Figure 6. experimental result based on Angular Fisheye Projection Model

（a） original picture



（b） experimental result

Figure 7. experimental result based on Cylinder Model

Fabric Defect Detection Based on Projected Transform for Feature Extraction

Xingye Zhang, Wenbo Xu
School of Internet of Things
Jiangnan University
Wuxi, China
e-mail: zxy1995@163.com

Ruru Pan, Jihong Liu, Weidong Gao
Key Laboratory of Eco-Textile, Ministry of Education
Jiangnan University
Wuxi, China
e-mail: prrsw@163.com

***Abstract*—In order to resolve the key technology problem of automated fabric defect detection, projected transform is proposed to extract features of the fabric image making use of fabric characteristic in this paper. Automated fabric defect detection scheme is divided into two phases, which are the study phase and the detection phase. During the study phase, features of normal fabric image are extracted to get the feature data set of normal fabric. During the detection phase, the method of anomaly detection is developed using features of fabric image to detect defect. Testing on general fabric by this method, experimental results demonstrate that each of feature values is in the normal range for normal fabric image, and one feature value at least is abnormal for defective fabric image. Defects can be located according to the location of abnormal value.**

Keywords- fabric; defect detection; projected transform; feature extraction; anomaly detection

I. INTRODUCTION

Image analysis techniques are being increasingly used to automate the detection of fabric defects in recent years. The approaches of fabric feature extraction has been mainly classified into two categories, one kind of approach is realized in the spatial domain, such as Cooccurrence matrix [1], Local Entropy [2] etc. Another approach is realized in the frequency domain, such as FFT [3], Gabor Filter[4] and wavelet transform[5] etc. The approach of the frequency domain is computational expensive. An automatic fabric defect detection scheme based on projected transform in the spatial domain and anomaly detection is presented in this study. Projected transform is employed to extract features making full use of the fabric characteristic in this paper, and the method of anomaly detection is developed using features of fabric image to detect defect.

II. FEATURE EXTRACTION BASED ON PROJECTED TRANSFORM

A. Principle

Feature extraction according to the fabric image characteristic is a key step in the fabric defect detection system. According to some previous research, light gradient exists obviously in the middle of yarn, the other part of the yarn and the interstice between yarns in the fabric image. The lightness from high to low sequence is the middle of yarn, the other part of the yarn and the interstice between yarns. The light change of the fabric image reflects the change of the yarn location. Projection in the horizontal direction and in the vertical direction can be obtained by using the above principle. The texture of the normal fabric image has a certain period principle [6] [7]. The texture of fabric defect area is different from normal regular fabric texture because most defects are caused mainly by weaving irregularly.

Projected transform is proposed to extract features of the fabric image making use of the above mentioned fabric character. The method of computing projected value is as follow:

Given the size of fabric image is M×N, coordinate system is set up in the fabric image first, x axis (horizontal direction) parallels weft yarn, and y axis(vertical direction) parallels warp yarn. The coordinate of each pixel is (x, y),and the gray value is f(x,y), then the projected value of the pixels in the warp direction is as follows[8]:

$$L(x)=\frac{1}{M}\sum_{y=0}^{M-1}f(x,y) \quad (1)$$

The projected value of the pixels in the weft direction is as follows:

$$L(y)=\frac{1}{N}\sum_{x=0}^{N-1}f(x,y) \quad (2)$$

In order to analyze easily, the projected discrete dots are joined to get the curve. Then the gray projected transform image can be obtained. Y axis is the gray value, and x axis is the row or column of the projection. It can be seen from the gray projected transform image in Figure 1, the curve varied approximately sinusoidally, which reflects interlacing law of the fabric.

B. The selection of resolution

For the different projected image with different resolutions which are 600dpi、1200dpi、2400dpi repectively, it's found that higher resolutions lead to more smooth projected curve. But high resolution leads to expensive computation. In real application, the selection of resolution has relation with fabric density and cover factor. The warp density and weft density of the woven fabric shown in Figure 2 are 70.2 ends/inch and 70.2 picks/inch, and its cover factor is 80%. The resolution is at least 657 dpi if the interstice between two warp yarns is computed as one pixel. In order to improve precision, resolution can be set

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.42

higher properly. The resolution of fabric image in this paper is set to 1200 dpi in order to give consideration to accuracy and computational amount.

III. ANOMONY DETECTION AND LOCATION

A. Feature value

For normal fabric, the peak value and the valley value fluctuate in a certain range and each extreme value distributes equably in the gray projected image. From the view of extracting features by projected transform, there are three kinds of defects: gray value anomaly, space anomaly, both gray and space anomaly. For mean gray value defects such as oil spot, its mean gray value or peak value and valley value may change drastically and exceed the normal range of the gray feature values. For structural defects such as double warp, they destroy the regularity of yarn arrange, so the peak space or valley space may change obviously.

During the course of weaving fabric, weft is weaved one by one, so the distribution of weft is uniform relatively and the space between two wefts is clear. One or more warp are inserted through the dent, and the warp density is high, so the regularity of projection in the weft direction is more obvious than the projection in the warp direction. The fluctuation of gray value in the warp direction is larger than the fluctuation in the weft direction, but the peak space and the valley space have a good regularity, as shown in Figure 2. By the above analysis, nine features are extracted: mean gray value of the fabric image, maximum peak value, minimum valley value, peak space and valley space which are obtained by projection in the warp and weft direction respectively.

B. Anomaly detection and location

Automatic fabric defect detection scheme is divided into two phases, which are the study phase and the detection phase respectively. In the study phase, feature values are extracted by a set of normal fabric samples first. The range of each feature value is acquired by statistical method. Maximum value and minimum value of feature values of normal fabric image have been applied to certain confidence interval of feature values. In the detection phase, the feature values are computed for a set of windows covering the image. Most of the windows contain the basic, defect-free pattern, while some windows may contain defects. Since the feature values in the windows which contain defects are different from the feature values of normal fabric obviously, the method of anomaly detection is deployed to detect defects. If only one feature value is not in the normal range, defects can be determined to be detected. Defects can be located according to the location of abnormal value.

C. The size of the window

It must be noted that the size of the window is very important. Owning to the characteristics of the fabric during the weaving course and the flexibility of fabric in itself, the projected values of the normal fabric fluctuate in a certain range. Smaller size of window is not enough to include the fluctuation of feature value sufficiently. Larger size of window results in decreased sensitivity of the gray value and only large defects are detected. Therefore, the size of the window will have to be freely adjusted according to the required sensitivity and the resolution. For the fabric used in this paper, the size of the window is set to 128×128, the real size of the window is $\frac{128}{1200}\times 25.4=2.7$mm. The minimum size of the defect which can be detected is 0.5mm, then $\frac{0.5}{2.7}\times 100\%$ is near to 20% and the variation of the gray value of the defects is enough to influence the mean gray value in the window.

IV. EXPERIMENT AND ANALYSIS

The fabric samples used in this paper are raw-white cotton. The feature values are extracted by 50 normal samples of fabric image and each of their size is 128×128. Then the defects are detected by the method as mentioned above. Some typical defective fabric images and their projected transform image are shown in Figure 3~ Figure 6. The results of the experiments are shown in Table 1. The range of each feature values of normal fabric and the abnormal values of defective fabric are listed in Table 1. For the fabric which is required to detect, it can be taken as normal fabric if its each feature value is in the normal range or it will be taken as defects if only one feature value is not in the normal range.

For defective fabric A as shown in Figure 3, mean gray value is 174, minimum valley value in the weft direction is 131 and minimum valley value in the warp direction is 131. These three feature values are all lower than the normal range. The peak space and valley space are both in the normal range, so the defective fabric A is a gray value defect (oil spot)). For defective fabric B as shown in Figure 4, all of feature values are in the normal range except that the peak space between 47 row and 70 row is 23 and the valley space between 38 row and 63 row is 25 in its projected transform image in the weft direction. These two feature values both exceed the normal range, so it can be concluded that there is a defect in the weft direction of the defective fabric B. For defective fabric C as shown in Figure 5, the valley space between 16 column and 23 column is 7, the valley space between 23 column and 41 column is 18, the peak space between 30 column and 45 column is 15, the valley space between 49 column and 67 column is 18, the peak space between 57 column and 73 column is 16 in its projected transform image in the warp direction. All of these feature values exceed the normal range, so it can be concluded that there is a defect in the warp direction of the defective fabric C. The same analysis can be applied in defective fabric D as shown in Figure 6. In its projected transform image in the warp direction, the maximum peak

value is 239 and is higher than the normal range. The valley space between 66 column and 88 column is 22, the peak value between 62 column and 76 column is 14, and the peak value between 76 column and 92 column is 16. All of these feature values exceed the normal range, so it can be concluded that there is a defect in the warp direction of the defective fabric D. From the above analysis, it can be seen that defects can not only be detected by inspecting abnormal values, but can be located by the location of abnormal values.

V. CONCLUSIONS

The automatic fabric defect detection scheme based on projected transform and anomaly detection is presented in this study. The feature data set of normal fabric can be obtained by extracting features of normal fabric images in the study phase. The method of anomaly detection is developed using nine features of fabric image to detect defect. The intent of the method is to make the calculations simple and fast for the system to be suitable for real-time applications.

The performance of the proposed woven fabric defect detection scheme has been evaluated by using a set of plain woven fabric images. The experimental results obtained have indicated that the scheme performs very well in detecting woven fabric defects. The future work is to apply the woven fabric defect detection scheme into real-time industrial environment.

REFERENCES

[1] Y. Shimizu. "Expert system to inspect fabric defects by pattern recognition," IEEE Transactions on Pattern Recognition , vol. 46(3), 1990 , pp. 460-469.

[2] X. Y. Qing, H. Duan, J. M. Wei . "Defect Inspection and Reco.gnition Based on Local Entropy Method," Journal of Textile Research, vol. 25 (5), 2004 , pp. 57 - 58

[3] I. S. Tsai , M. C. Hu. "Automatic inspection of fabric defects using an artificial neural network technique," Textile Research Journal , vol. 66(7), 1996,pp. 472-482.

[4] D. L. Yuan, Y. M. Song, "Fabric Defect Detection Based on Optima Gabor Filters," Journal of Image and Graphics, vol. 11(7) 2006, pp. 954-958.

[5] L. Q. Li, X. B. Huang, "Woven Fabric Defect Detection with Features Based on Adaptive Wavelets," Journal of DongHua University. vol. 27 (4) 2001 ,pp. 82 – 87

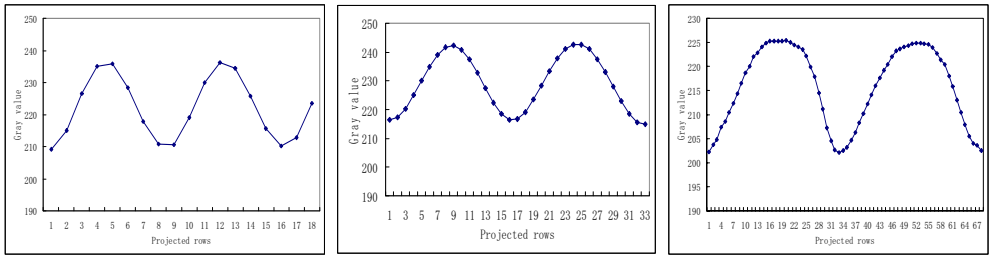
[6] D. Chetverikova, A. Hanbury. "Finding defects in texture using regularity and local orientation," Pattern Recognition vol. 35, 2002 pp. 2165-2180.

[7] A. Abouelela, H. M. Abbas, H. Eldeeb,etc. "Automated vision system for localizing structural defects in textile fabrics," Pattern Recognition Letters vol. 26 ,2005, pp. 1435–1443.

[8] W. D. Gao, J. H. Liu, B. J. Xu,etc. "Automatic Identification of Weft Arrangement Parameters in Fabric," Cotton Textile Technology. vol. 30(1) 2002, pp. 28-31.

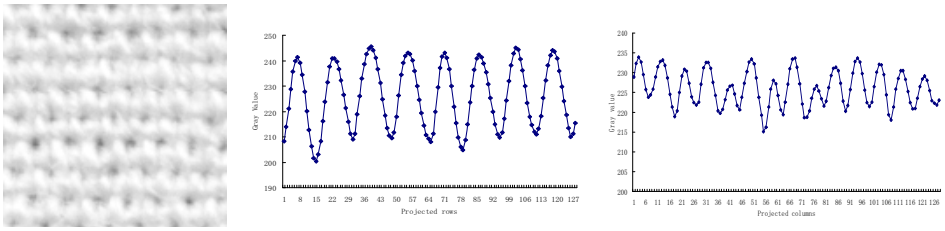
TABLE I. ANOMALY ANALYSIS OF FEATURE VALUES OF DEFECTIVE FABRIC IMAGE

Feature value		Mean gray value	Weft direction				Warp direction			
			Max peak value	Min valley value	Peak space	Valley space	Max peak value	Min valley value	Peak space	Valley space
Normal range		222-229	249	190	14-19	14-19	235	210	8-12	8-12
Defective fabric	A	174	—	131	—	—	—	—	—	—
	B	—	—	—	23	25	—	—	—	—
	C	—	—	—	—	—	—	—	15、16	7、18
	D	—	—	—	—	—	239	—	14、16	22



(a) (b) (c)

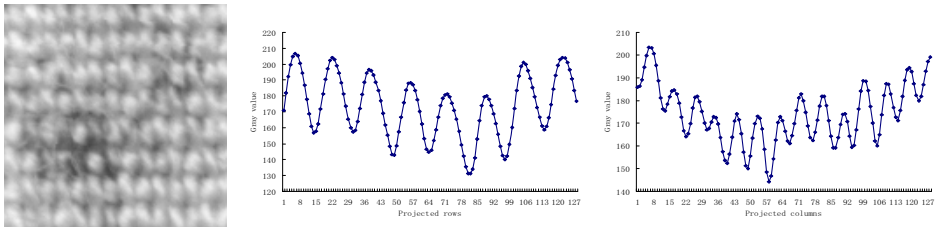
Figure1. Gray projected images of different resolution
(a)600 resolution (b) 1200 resolution (c) 2400 resolution



(a) (b) (c)

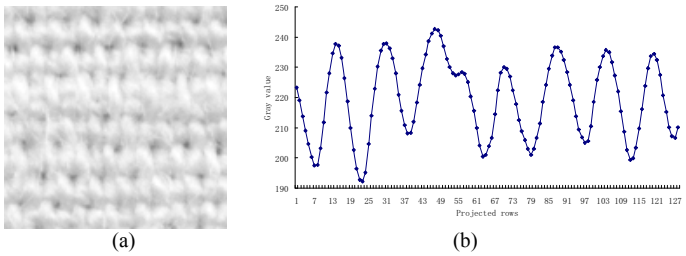
Figure.2 Plain fabric image and gray projected image

(a) fabric image (b) Gray projection in the weft direction (c) Gray projection in the warp direction



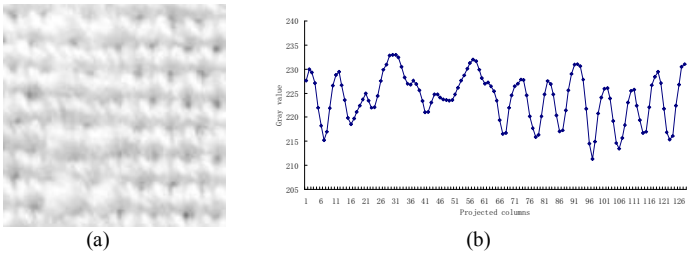
(a) (b) (c)

Figure.3 Defective fabric A
(a) Fabric image (b) Gray projection in the weft direction (c) Gray projection in the warp direction



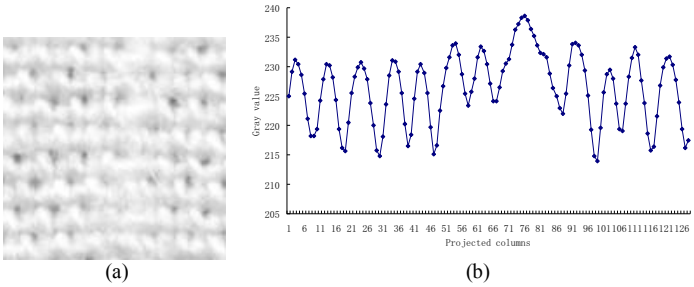
(a) (b)

Figure.4 Defective fabric B (a) Fabric image (b) Gray projection in the weft direction



(a) (b)

Figure.5 Defective fabric C (a) Fabric image (b) Gray projection in the warp direction



(a) (b)

Figure.6 Defective fabric D (a) Fabric image (b) Gray projection in the warp direction

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

An Improved GAC Model Combining with GNGVF

Yanqing Guo, Meiqing Wang*
College of Mathematics and Computer Science,
Fuzhou University,
Fuzhou, China
e-mail: gyqing_123@163.com, mqwang@fzu.edu.cn

Choi-Hong Lai
School of Computing and Mathematical Sciences,
University of Greenwich,
London, UK
e-mail: c.h.lai@gre.ac.uk

***Abstract*—Geodesic active contour (GAC) model is a common used method for image segmentation. But one drawback of this model is that it's difficult to control the number of iterations and sometimes may produce over-segmentation results. In this paper, the generalized GVF in the normal direction (GNGVF) is proposed, and an improved GAC model which combines GAC with GNGVF is proposed. When the distance between the evolution curves produced by successive iterations is smaller than some given threshold, the on-off function will change and different force will affect. The new force can extend the capture range and stop the curve at the boundary stably avoiding over-segmentation. The experimental results show the curve can converge to boundary well.**

Keywords- image segmentation; GAC model; GNGVF field; L2-distance

I. INTRODUCTION

Active contour models [1-3,5,6] are used widely for detecting boundaries in computer vision, object recognition, biomedical image processing and so on. One of the famous geometric active contour models is geodesic active contour (GAC) model proposed by Caselles et al. (1997) [2]. It doesn't depend on parameters, and can split or merge the curve freely because of using the level set. However, it also has some drawbacks. In this paper, what we interest in is the problem of choosing the number of final iterations.

Due to the existence of the shrinking force, GAC model will lead to over-segmentation if the iterations surpass some value. To deal with this problem, we once proposed a convergence-to-boundary method [7]. There are two steps in this method, and the second step combining GAC with GVF is used when the distance between the evolution curves produced by successive iterations is smaller than some given threshold. It can extend the capture range and stop at the boundary stably avoiding over-segmentation. In this paper, the two steps are united into one model. Gradient vector flow (GVF) is replaced by generalized GVF in the normal direction (GNGVF) which is combined NGVF field and GGVF field. The experimental results show that the curve under the improved model can converge to boundary better than the old method.

II. GEODESIC ACTIVE CONTOUR MODEL

Geodesic active contour model is proposed by Caselles et al. [2] in 1997. Its energy functional is defined as

$$L_R(C)=\int_0^{L(C)} g\left(\left|\nabla I[C(s)]\right|\right)ds \tag{1}$$

where s is the arc length parameter, and $L(C)$ indicates the total length of the curve *C*. The "edge detector" function *g* should be monotonically decreasing, and $g:[0,+\infty]\to R^+$, $g(0)=1$, $g(x)\to 0$ as $x\to\infty$. Usually the following function is used:

$$g(r)=\frac{1}{1+(r/K)^p},\ p=1,2 \tag{2}$$

Where r is a point on the image, and K is a constant for controlling the decreasing velocity of *g*.

Minimizing (1) by using calculus of variations, we can get the corresponding gradient descent flow as

$$\frac{\partial C}{\partial t}=g(c)\kappa N-(\nabla g\cdot N)N \tag{3}$$

where κ is the curvature and N is the normal vector of curve C. The first item is an internal power provided by the curvature motion. When the curve approaches to boundary of objects, g becomes smaller, so the internal power becomes weaker and the curve stops moving. The second item is an external power provided by the gradient ∇g derived from the image. Since ∇g always directs away from the boundary whether the curve *C* is inside or outside the boundary, the curve would approach to and stop at the boundary stably.

For dealing with the convergence to boundary concavities efficiently, Caselles et al. proposed a generalized GAC model by adding a "shrinking power" gcN and then . Equation (3) is modified as following:

$$\frac{\partial C}{\partial t}=g(c+\kappa)N-(\nabla g\cdot N)N \tag{4}$$

This shrinking power not only can fast the convergence speed at smoothing areas, but also make the curve shrink when the curvature is negative and segments the concave boundary properly.

Using the level set method, the corresponding PDE of (4) with the embedding function *u* is

*: corresponding author

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.23

$$\frac{\partial u}{\partial t}=\left[g(c+\kappa)-\nabla g\cdot\frac{\nabla u}{|\nabla u|}\right]|\nabla u| \qquad (5)$$
$$=cg|\nabla u|+div\left(g\frac{\nabla u}{|\nabla u|}\right)|\nabla u|$$

where the function u usually is a sign distance function defined as following,

$$u(x,y)=\begin{cases} d[(x,y),C], & (x,y)\text{is outside of C} \\ -d[(x,y),C], & (x,y)\text{is inside of C}\end{cases}$$

III. Generalized Gradient Vector Flow in The Normal Direction

Xu et al. presented a new external force field-GVF field, which is composed by gradient vector flow [4]. The vectors of GVF field vary smoothly on homogeneous image regions and point to the object boundary when they are near to the boundary.

The GVF field is defined to be the vector field $V(x,y)=(u(x,y),v(x,y))$ that minimizes the energy functional

$$\varepsilon=\iint\mu\left(u_x^2+u_y^2+v_x^2+v_y^2\right)+|\nabla f|^2|V-\nabla f|^2\,dxdy \qquad (6)$$

where $f(x,y)$ is an edge map derived from the image $I(x,y)$ having the property that it is larger near the image edges. For a gray-level image, the following edge maps are often used: $f(x,y)=-|\nabla(G_\sigma(x,y)*I(x,y))|^2$ or $f(x,y)=-|\nabla I(x,y)|^2$.

In (6), when $|\nabla f|$ is small, the energy is dominated by the first term, yielding a smooth field. On the other hand, when $|\nabla f|$ is large, the second term dominates the integrand, and is minimized by setting $V=\nabla f$. μ is a regularization parameter governing the tradeoff between the first term and the second term. This parameter should be set according to the noise of the image. More noise in image, larger μ should be set.

Using the calculus of variations, the following gradient descent flow can be obtained:

$$\begin{cases} u_t=\mu\nabla^2u-(u-f_x)\left(f_x^2+f_y^2\right) \\ v_t=\mu\nabla^2v-(v-f_y)\left(f_x^2+f_y^2\right)\end{cases} \qquad (7)$$

where ∇^2 is the Laplacian operator. The main advantages of the GVF-snake model are that it can capture a snake from a long range—from either side of the object boundary—and can force it into concave regions (Fig.1).

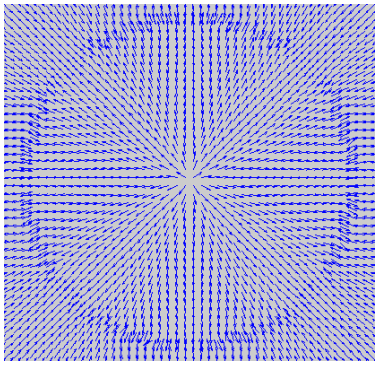


Figure 1. The GVF field of a circle.

Xu et al. improved GVF to GGVF field [9] which is as following,

$$\begin{cases} u_t=g(|\nabla f|)\nabla^2u-h(|\nabla f|)(u-f_x) \\ v_t=g(|\nabla f|)\nabla^2v-h(|\nabla f|)(v-f_y)\end{cases} \qquad (8)$$

where $g(|\nabla f|)=e^{-(|\nabla f|/K)}, h(|\nabla f|)=1-g(|\nabla f|)$. This pair of weighting functions will conform to the edge map gradient at strong edges, but will vary smoothly away from the boundaries.

Ning Jifeng et al. proposed an improved external force field based on GVF which is called as NGVF [8]. It decomposed Laplacian terms along the tangent and normal directions:

$$\nabla^2w=w_{TT}+w_{NN} \qquad (9)$$

$$w_{TT}=\frac{1}{|\nabla w|^2}\left(w_x^2w_{yy}+w_y^2w_{xx}-2w_xw_yw_{xy}\right) \qquad (10)$$

$$w_{NN}=\frac{1}{|\nabla w|^2}\left(w_x^2w_{xx}+w_y^2w_{yy}+2w_xw_yw_{xy}\right) \qquad (11)$$

where w_{TT} and w_{NN} denote the second derivatives of w in the tangent direction and normal direction, respectively.

According to the analyses in [8], w_{NN} has the best interpolating effects among ∇w, w_{TT} and w_{NN}. Then the NGVF force field is defined as following,

$$\begin{cases} u_t=\mu u_{NN}-(u-f_x)\left(f_x^2+f_y^2\right) \\ v_t=\mu v_{NN}-(v-f_y)\left(f_x^2+f_y^2\right)\end{cases} \qquad (12)$$

NGVF has a bigger time step than GVF, and can help the curve enter into long, thin concavities.

For getting a better force field, we combine NGVF with GGVF and get the following vector partial differential equation:

$$\begin{cases} u_t = g\left(|\nabla f|\right)u_{NN} - h\left(|\nabla f|\right)\left(u - f_x\right) \\ v_t = g\left(|\nabla f|\right)v_{NN} - h\left(|\nabla f|\right)\left(v - f_y\right) \end{cases} \tag{13}$$

where $g\left(|\nabla f|\right) = e^{-\left(|\nabla f|/K\right)}, h\left(|\nabla f|\right) = 1 - g\left(|\nabla f|\right)$. We call this new force field as GNGVF (generalized GVF in the normal direction). Therefore, the vector of the force field will conform to ∇f at strong edges, but will vary smoothly away from the edges. The comparison between NGVF and GNGVF is depicted as Fig.2.

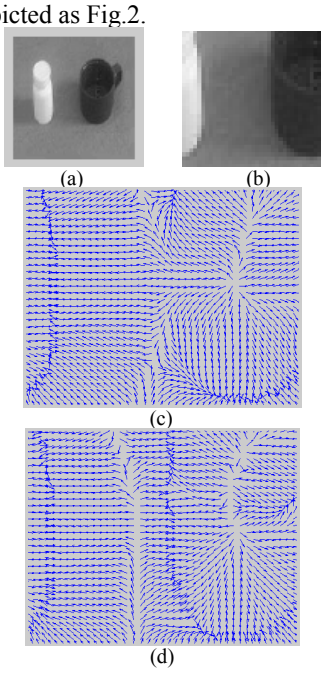


Figure 2. (a) Image "gac1.bmp". (b) The central part of image. (c) The corresponding NGVF field of (b), and the edge of black cup has been smoothed. (d) The corresponding GNGVF field of (b), and the edge of black cup is preserved perfectly.

IV. An Improved GAC Model Combining with GNGVF

A. Analysis of the GAC Model

As running the process of GAC model, we can find that the curve will move fast away from the boundaries and move slowly at the boundaries. In theory, if $u_t = 0$ (u is a level set function), the energy will arrive to the minimum which means the curve is on the edges. Because its slow velocity and the limit that the initial outline should be around the curve, people usually use the generalized GAC model which adds a shrinking force. The following GAC model in this paper is short for the generalized GAC model. But another problem that u_t can't arrive to zero produces.

The reasons producing this problem are the shrinking force and the small capture range of ∇g. The shrinking force gcN always makes the curve shrink because c is a positive constant. Although it decreases due to the influence of $g\left(|\nabla I|\right)$, it doesn't decrease to zero in fact. Therefore, if the number of iterations is too big, the position of the curve would be out of the capture range of ∇g. Then the shrinking force will increase and lead to over-segmentation.

So when running the GAC model, it's difficult to control the number of iterations. The number of iterations is pre-assigned and if it is too small, the evolution curve doesn't reach the right segmentation; but if too big, it will be over the right segmentation (Fig.3). Therefore, if the shrinking force gcN can be instead by another force at edges, the problem will be solved. The characteristic when the curve reaches edges is that u_t is very small.

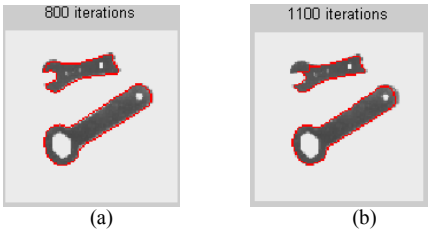


Figure 3. Use GAC model. (a)The number of iterations is 800.(b) When the number of iterations is 1100, the left part of object is over-segmentation.

B. The Distance Between Two Curves

The distance between two curves C_1 and C_2 is defined as the distance between the two corresponding signed distance functions u_{C_1} and u_{C_2} as following [10]: $d\left(C_1, C_2\right) = d_D\left(u_{C_1}, u_{C_2}\right)$, where D denotes the space of signed distance functions, and $d_D\left(\cdot,\cdot\right)$ means the metric in space D. The metric we usually use is L_2-distance:

$$d_{L_2}\left(u_1, u_2\right) = \left(\int_\Omega \left|u_1(x) - u_2(x)\right|^2 dx\right)^{\frac{1}{2}} \tag{14}$$

C. An Improved Model Combining GNGVF with GAC

In this paper, we propose an improved model which converges to the boundary of objects. In this method, the model combining GVF with GAC is used when the distance between the evolution curves produced by successive iterations is smaller than some given threshold. It can extend the capture range and stop at the boundary stably avoiding over-segmentation.

An on-off function is defined as

$$h\left(d\left(C_i, C_{i+1}\right)\right) = \begin{cases} 1 & , \quad d\left(C_i, C_{i+1}\right) > \varepsilon \\ 0 & , \quad otherwise \end{cases}$$

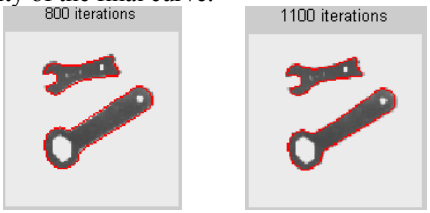
ε is a pre-assigned threshold .

The improved model is defined as following,

$$C_t = h(d(C_i, C_{i+1}))(g(c+\kappa)N + (\nabla g \cdot N)N) + (1 - h(d(C_i, C_{i+1})))sign(V \cdot N)N \quad (15)$$

where V is the vectors in GNGVF field, and N is the unit normal vector that points to the inside of the curve.

When the distance between two curves on successive iterations $d(C_1, C_2)$ is greater than ε, only the first item plays a role and the model becomes GAC model. When $d(C_1, C_2)$ is smaller than ε, the curve is determined by the second item which just like a balloon force. $sign(V \cdot N)$ denotes the sign of $V \cdot N$. If $V \cdot N = 0$, then we define $sign(V \cdot N) = 1$ .We can see that V always points to the boundary. If the curve is outside of the boundary, then V and N have the same direction and the curve continue to deflate. Otherwise, if the curve is inside of the boundary, then V and N have the same direction and the curve continue to inflate. Moreover, because GVF field extends the capture range of the contour and there is no influence of the shrink force, the curve will attract to boundary stably just like Fig.4. We can stop the circle of the process through the invariability of the final curve.



(a) (b)

Figure 4. Using the improved GAC model. (a)The number of iterations is 800.(b)The number of iterations is 1100.The boundary is segmented rightly.

V. EXPERIMENT RESULTS

In this paper, the experiments are implemented in Matlab. In the following experiments, the parameters in GNGVF is set the same as $\mu = 0.25$, $K = 0.05$, $\Delta t = 0.05$. We do three experiments and shrink image by 0.5 times during the process for convenience. During each group of experimental images, the first image is the original image and the red curve is the initial contour. The second image is the result of segmentation. The third image is the normalized GNGVF field of the first image.

The first experiment is to segment the image "eee.bmp". In the improved model, we need two time steps Δt_1 and Δt_2 for the first and second item respectively. The parameters are set as $\Delta t_1 = 0.07$, $\Delta t_2 = 0.5$, $c = 3$, $\varepsilon = 7$ .The segmentation result is gained when the iterations are 477. The result is shown in Fig.5. The second experiment is to segment the image "3.bmp". The parameters are set as $\Delta t_1 = 0.05$, $\Delta t_2 = 0.5$, $c = 3$, $\varepsilon = 5$. The result is shown in Fig.6. The segmentation result is gained when the iterations are 601.

For the inhomogeneous areas and noise images, this model also can deal with them. The third experiment is to segment the image "twocolor.bmp" which has inhomogeneous areas. The parameters are set as $\Delta t_1 = 0.07$, $\Delta t_2 = 0.5$, $c = 3$, $\varepsilon = 7$. The segmentation result is gained when the iterations are 317. The result is shown in Fig.7. The last experiment is to segment the image "3noise.bmp" with salt and pepper noise. The parameters are set as $\Delta t_1 = 0.05$, $\Delta t_2 = 0.5$, $c = 3$, $\varepsilon = 5$. The segmentation result is gained when the iterations are 1345. The result is shown in Fig.8.

Compared with the old method combining GAC with GVF which is performed in two steps, the improved GAC model which is added GNGVF has better results (Fig.9) .

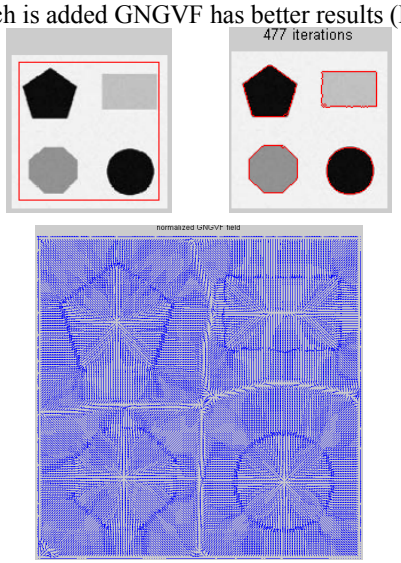


Figure 5. The result of "eee.bmp" by the improved GAC model and the corresponding GNGVF field.

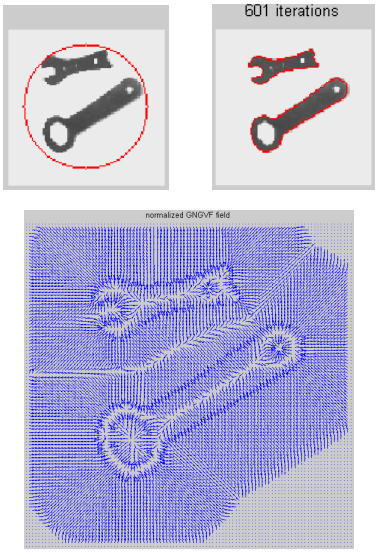


Figure 6. The result of "3.bmp" by the improved GAC model and the corresponding GNGVF field.

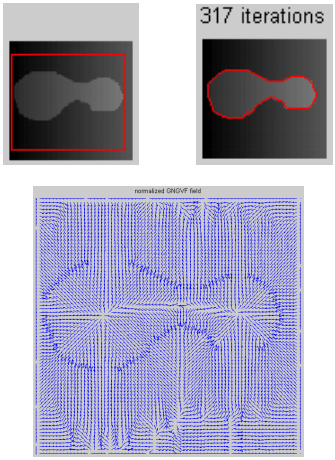


Figure 7. The result of "twocolor.bmp" by the improved GAC model and the corresponding GNGVF field.

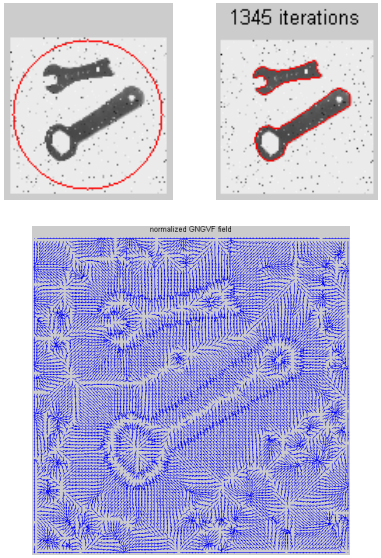


Figure 8. The result of "3noise.bmp" by the improved GAC model and the corresponding GNGVF field.

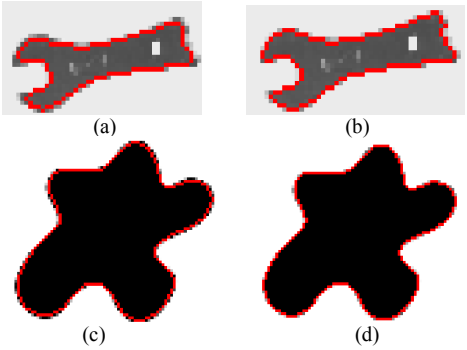


Figure 9. (a) Part of "3.bmp" by the old method. (b) Part of "3.bmp" by the improved GAC model. (c) The result of "draw1.bmp" by the old method. (d) The result of "draw1.bmp" by the improved GAC model.

VI. Conclusion and Future Work

In this paper, we propose an improved GAC model. This model combines GAC with GNGVF, and it can reduce the trouble of controlling the number of iterations by people. But the improved model doesn't overcome the drawbacks of both GAC and GNGVF such as bad effects for blurred images, the limit of the initial curve's position. All of these will be our future work.

Acknowledgment

This work was partially supported by NSFC under Grant No.11071270, Natural Science Foundation of Fujian Province under Grant No. 2010J01328 and the Technology Innovation Platform Project of Fujian Province under Grant No.2009J1007.

References

[1] Michael Kass, Andrew Witkin, and Demetri Terzopoulos, "Snakes: Active Contour Models," International Journal of Computer Vision, 1988, pp.321-331.

[2] Vicent Caselles, Ron Kimmel and Guillermo Sapiro, "Geodesic Active Contours," International Journal of Computer Vision, vol. 22(1), pp. 61-79, 1997.

[3] Tony F. Chan and Luminita A. Vese, "Active Contours Without Edges," IEEE Transactions on Image Processing, 266-277, 2001.

[4] Chenyang Xu and Jerry L. Prince, "Gradient Vector Flow: A New External Force for Snakes," IEEE Proc. Conf. on Comp. Vis. Patt. Recog, 1997, pp 66-71.

[5] Nikos Paragios, Olivier Mellina-Gottardo and Visvanathan Ramesh, "Gradient Vector Flow Fast Geodesic Active Contours," IEEE Trans Pattern Analysis and Machine Intelligence, vol. 26(3), pp. 402-407, 2004.

[6] Kaihua Zhang, Lei Zhang, Huihui Song and Wengang Zhou, "Active contours with selective local or global segmentation: A new formulation and level set method," Image and Vision Computing, vol. 28, pp. 668-676, 2010.

[7] Yanqing Guo, Meiqing Wang, Choi-Hong Lai. "A Convergence-to-boundary Segmentation Method Combining GVF with GAC," submitted to CISP 2011.

[8] Ning Jifeng, Wu Chengke, Liu Shigang, Yang Shuqin, "NGVF: An improved external force field for active contour model," Pattern Recognition Letters, vol. 28, pp. 58-63, 2007.

[9] Chenyang Xu and Jerry L. Prince, "Generalized Gradient Vector Flow External Forces for Active Contours," Signal Processing, vol. 71, pp. 131-139, 1998.

[10] Junmo Kim, Mujdat Cetin and Alan S. Willsky, "Nonparametric Shape Priors for Active Contour-based Image Segmentation," Signal Processing, vol.87, pp. 3021-3044, 2007.

Image registration with a modified quantum-behaved particle swarm optimization

Yu Bao
School of Digital Medium, Jiangnan University, China
No. 1800, Lihu Avenue
Wuxi, Jiangu 241222, China
e-mail:494795639@qq.com

Jun Sun
Key Laboratory of Advanced Process Control for Light Industry (Jiangnan University), Ministry of Education, P.R. China
e-mail: sunjun_wx@hotmail.com

Abstract—**A revised quantum-behaved particle swarm optimization (RQPSO) with novel iterative equation is proposed, which uses hybrid of double exponential and normal probability distributions to sample the particle's position. The RQPSO is applied to medical image registration and the performance comparison testifies that it outperforms the PSO and QPSO algorithms on the tested image registration problems.**

Keywords- particle swarm optimization, quantum-behaved particle swarm optimization, image segmentation, probability distribution

I. INTRODUCTION

The particle swarm optimization (PSO) method was originally proposed by J. Kennedy and R.C. Eberhart as a simulation of social behavior of bird flock and was initially introduced in [1, 2]. Since its origin, the PSO algorithm has already been used in a large variety of application areas due to its easy implementation and low requirement in memory and CPU speed. However, as proved by F. Van Den Bergh [3], PSO is not a global convergence-guaranteed algorithm because the particles are restricted to a finite search space at each iteration, which weakens the global search ability of the algorithm and makes the PSO algorithm easily trap into the local optima. In order to overcome this shortcoming, Sun et al. proposed a novel variant of PSO, named quantum-behaved particle swarm optimization (QPSO), in which each individual particle is assumed to have quantum behavior [4]-[6]. In this paper, we introduce a QPSO with a hybrid probability distribution (QPSO-HPD) and apply it to medical image registration, which has already played an increasingly important role in surgery planning, functional studies, computer-guided therapies and other biomedical researches.

The rest part of the paper is organized as follows. The principle of QPSO is reviewed in Section 2. The QPSO-HPD is presented in Section 3. Section 4 is the results of medical image registration using QPSO-HPD. Concluding remarks are provided in the last section.

II. QUANTUM-BEHAVED PARTICLE SWARM OPTIMIZATION

In the original PSO with M individuals, each individual is treated as an infinitesimal particle in the D-dimensional space, with the position vector and velocity vector of particle i $X_i(t)=(X_{i1}(t),X_{i2}(t),\cdots,X_{iD}(t))$ and $V_i(t)=(V_{i1}(t),V_{i2}(t),\cdots,V_{iD}(t))$. The particle moves according to the following equations:

$$\begin{aligned} V_{id}(t+1) &= V_{id}(t)+c_1r_1(P_{id}(t)-X_{ij}(t)) \\ &\quad +c_2r_2(P_{gd}(t)-X_{id}(t)) \end{aligned} \tag{1}$$

$$X_{id}(t+1)=X_{id}(t)+V_{id}(t+1) \tag{2}$$

where i=1,2,…, M, d=1,2,…, D, c_1 and c_2 are called the acceleration coefficients which are usually set within [0,2]. Parameters r_1 and r_2 are two different sequence of random numbers distributed uniformly on (0,1). The personal best (*pbest*) position $P_i=(P_{i1},P_{i2},\cdots,P_{iD})$ is the best previous position which gives the best fitness so far of the particle i, and $P_g=(P_{g1},P_{g2},\cdots,P_{gD})$, known as global best (*gbest*) position, is the position of the best particle among all the particles. Generally, the value of V_{ij} should be restricted in the interval $[-V_{\max},V_{\max}]$.

In [7], it has been proved that if c_1 and c_2 are properly selected, each particle will converge to $p_i=(p_{i1},p_{i2},\cdots p_{iD})$, and whose coordinates are

$$p_{id}(t)=\varphi\cdot P_{id}(t)+(1-\varphi)\cdot P_{gd}(t) \tag{3}$$

where $\varphi=c_1r_1/(c_1r_1+c_2r_2)$. It can be seen that the local attractor is a stochastic attractor of particle i that lies in a hyper-rectangle with P_i and P_g being two ends of its diagonal. If c1 equals c2, φ will be a random number uniformly distributed on (0,1).

Assume that each individual particle move in the search space with a δ potential well on each dimension, of which the center is the point p_{id} . The update equation of particle position is

$$X_{id}(t+1)=p_{id}(t)\pm\frac{L_{id}(t)}{2}\cdot\ln(1/u) \tag{4}$$

where $L_{id}(t)$ is the characteristic length of the δ potential well. In [5], a global point called mean best position of the population is introduced into PSO. The global point, denoted as C, is defined by the mean of the personal best positions among all particles. That is

$$C(t)=(C_1(t),C_2(t),\cdots,C_D(t))=\left(\frac{1}{M}\sum_{i=1}^{M}P_{i1}(t),\ \frac{1}{M}\sum_{i=1}^{M}P_{i2}(t),\ \cdots,\ \frac{1}{M}\sum_{i=1}^{M}P_{iD}(t)\right) \tag{5}$$

where M is the population size and P_i is the personal best position of particle i. Then the value of $L_{id}(t)$ is evaluated by $L_{id}(t)=2\alpha\cdot\left|C_d(t)-X_{id}(t)\right|$ and the position are updated by

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.51

$$X_{id}(t+1) = p_{id}(t) \pm \alpha \cdot \left| C_d(t) - X_{id}(t) \right| \cdot \ln(1/u) \tag{6}$$

where parameter α is called contraction-expansion (CE) coefficient, which can be tuned to control the convergence speed of the algorithms. The PSO with equation (6) is known as quantum-behaved particle swarm optimization (QPSO), where parameter α must be set as $\alpha < 1.781$ to guarantee convergence of the particle [7]. In most cases, α can be controlled to decrease linearly from $\alpha_0$to α_1 ($\alpha_0 < \alpha_1$).

III. A REVISED QPSO

Beside the quantum Delta potential well model, other quantum potential models such as harmonic oscillator potential well also can be established for PSO. The particles are subject to different distributions in different potential models, which influences the converge rate. In order to improve the performance of QPSO, other position distribution (or potential well) could be employed. However, it has been shown by preliminary experiments that the exponential distribution maybe the best unitary distribution for QPSO. It could be expected that a hybrid distribution may improve QPSO.

Now the iterative equation (6), considering the one-dimensional case, can be rewritten as:

$$x(t+1) = p(t) + A(t) \tag{7}$$

where $A(t)$ is a random sequence that should converge to zero to guarantee that $x(t)$ reach to point $p(t)$. In QPSO, as written above, $A(t)$ is subject to double exponential distribution. It could be assumed $A(t)$ as the sum of two random sequences, that is

$$A(t) = a(t) + b(t) \tag{8}$$

where *a(t)* and *b(t)* are two random sequences subject to different probability distributions. In our proposed method, we adopt that a(t) and $b(t)$ are subject to normal distribution and double exponential distribution respectively. Concretely, these two random sequences can be given by

$$a(t) = \pm\alpha \mid p(t) - x(t) \mid \ln(1/u) \tag{9}$$

$$b(t) = \beta \left| C(t) - x(t) \right| \cdot Rn \tag{10}$$

where u is a uniformly distributed random number on the interval (0,1) and Rn is random number with standard normal distribution. Both α and β are two algorithmic parameters and also called contraction-expansion coefficients. Therefore the iterative equation can be rewritten as

$$\begin{aligned} x(t+1) = p(t) &\pm \alpha \left| p(t) - x(t) \right| \ln(1/u) \\ &+ \beta \left| C(t) - x(t) \right| Rn \end{aligned} \tag{11}$$

In the D-dimensional search space, the iterative equation of each particle is

$$\begin{aligned} X_{id}(t+1) = p_{id}(t) &\pm \alpha \mid p_{i,d}(t) - X_{id}(t) \mid \ln(1/u) \\ &+ \beta \mid C_d(t) - X_{id}(t) \mid \cdot Rn \end{aligned} \tag{12}$$

Thus the proposed RQPSO can be described as follows

RevisedQPSO Algorithm

Initialize the population;
while the stop criterion is not met **do**
Compute the mean best position C by equation (5);
  for i = 1 to swarm size *M*
    If $f(X_i) < f(P_i)$ then $P_i = X_i$; **Endif**
    Update P_g;
    for d=1 to D
      φ=rand(0,1); u=rand(0,1);
      $p = \varphi * P_{id} + (1-\varphi) * P_{gd}$;
      if (rand(0,1)>0.5)
        $X_{id} = p + \alpha * \text{abs}(p - X_{id}) * \ln(1/u) + \beta * \text{abs}(C_d - X_{id}) * Rn$;
      Else
        $X_{id} = p - \alpha * \text{abs}(p - X_{id}) * \ln(1/u) + \beta * \text{abs}(C_d - X_{id}) * Rn$;
      Endif
    Endfor
  Endfor
Endwhile

The values of parameter α and β can be adjusted to balance the exploration and exploitation of the algorithm.

IV. MEDICAL IMAGE REGISTRATION USING RQPSO

This section presents the application QPSO-HPD to medical image registration. The imaging modalities employed can be divided into many categories, include X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (US) and portal images. There are many approaches to medical image registration. The "gold standard" utilizes markers placed on the subject [8], [9]. Increasing work has recently used intensity values to measure similarity between the images, which is known as intensity-based approaches, for they don't generally require extensive preprocessing, such as segmentation or feature extraction. There are generally three important considerations [9], [10] in intensity-based registration: search space, similarity metric and search strategy. The whole process can be formulated as $T : P_A \to P_B \Leftrightarrow T(P_A) = P_B$, where T denotes the spatial transformation that maps features or coordinates (spatial locations) from P_A to P_B. For the 3-dimensional rigid body registration, the mapping of coordinates $p = [x \quad y \quad z]^T$ into $p' = [x' \quad y' \quad z']^T$ can be formulated as follows:

$$\begin{bmatrix} p_B \\ 1 \end{bmatrix} = T \begin{bmatrix} p_A \\ 1 \end{bmatrix} \Leftrightarrow$$

$$\begin{bmatrix} x_B \\ y_B \\ z_B \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\beta\cos\gamma & \cos\alpha\sin\gamma + \sin\alpha\sin\beta\cos\gamma & \sin\alpha\sin\gamma - \cos\alpha\sin\beta\cos\gamma & t_x \\ -\cos\beta\sin\gamma & \cos\alpha\cos\gamma - \sin\alpha\sin\beta\sin\gamma & \sin\alpha\cos\gamma + \cos\alpha\sin\beta\sin\gamma & t_y \\ \sin\beta & -\sin\alpha\cos\beta & \cos\alpha\cos\beta & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_A \\ y_A \\ z_A \\ 1 \end{bmatrix} \tag{13}$$

That is, the goal of the optimization is to determine the parameters t_x, t_y, t_z, α, β and γ. Similarity metrics for image registration must be robust, and here mutual

information is chosen to calculate the similarity of floating and reference images.

$$I(A,B) = H(A) + H(B) - H(A,B) = \sum p_{AB}(a,b) \log \frac{p_{AB}(a,b)}{p_A(a)\, p_B(b)} \quad (14)$$

where *H* denotes Shannon entropy:

$$H(A) = -\sum_a p_A(a) \log p_A(a) \quad (15)$$

$$H(A,B) = -\sum_{a,b} p_{AB}(a,b) \log p_{AB}(a,b) \quad (16)$$

As the optimization task in image registration is to maximize similarity, similarity metric values are denoted as object function *f*(*X*). Nowadays normalized mutual information was selected frequently as the similarity measure in the current study [8], [11], [12]. This metric is robust, in the sense that it usually attains its maximum at correct alignment, but is till generally nonsmooth and has many local optima, as is the case with other metrics. For this reason, global optimization approaches are needed. Local methods, such as Powell's direction set method, conjugate gradient, Levenberg-Marquardt, or the Nelder-Mead simplex algorithm [13], [14], are generally used in image registration. However, there methods are prone to trap in local optima, as the global optimum may not be present in lower resolutions [8], [15]. Therefore, global optimization methods including simulated annealing [16], genetic algorithms [17], [18], evolutionary strategies [19], and the tabu search [20], are efficient tools to tackle this kind of problems.

V. Experimental Results

As mentioned above, the most three important considerations in registration are the transformation in search space, similarity metrics and search strategy. In this study, our experiments focus on registering 3-D volumes to 3-D volumes. All the data used in this paper were provided by *The Retrospective Image Registration Evaluation (RIRE) Project* database from Vanderbilt University, Nashville, TN, USA. For this patient dataset whose ID is 008, all CT images and all PET images are registered to the MR images which used as reference images. The volume modalities, size, and properties are shown in Table I.

When it comes to the real clinical use, various noises are often found in real medical images, which may rise the possibility of trapping in the local optimum, as shown in Fig. I. That is to say, registration method used in real clinical environment should be robust. In order to simulate the real medical images, Gaussian noise with mean=0, variance=0.001 were added into the original data of RIRE project. Registration transformation is limited to six-parameter rigid-body type (3 translations and 3 rotations) transformations. Experiments for performance comparison were made between QPSO, PSO, DRQPSO and Powell, a classical local search strategy which has been proved to be effectively in medical image registration without noises. All the parameters in QPSO, RQPSO and PSO were set as follows. For all the algorithms, the population size was 20, and the maximum number of iterations was 1000. For QPSO the CE coefficient was set to decrease linearly with the iteration number from 1.0 to 0.5. For PSO, the inertia weighted PSO [21] was used with the inertia weight decreasing linearly from 0.9 to 0.4 and both of the acceleration coefficients were set to 2. For RQPSO, the parameter α was fixed at 0.6 and the value of β decreased linearly from 0.9 to 0.4. Six CT-MR pairs and six PET-MR pair images were involved in the experiments which were conducted on a PC, having a 1.73Ghz Intel Pentium processor and a 1G DDR memory.

TABLE I. Image Characteristics

	Rows	Columns	Slices	Pixel Size	Slice thickness
CT	512	512	29	0.653595	4
PET	128	128	15	2.590723	8
MR-PD	256	256	26	1.250000	4
MR-T2	256	256	26	1.250000	4
MR-PD rectified	256	256	26	1.263903	4.102400
MR-T2 rectified	256	256	26	1.271000	4.072800

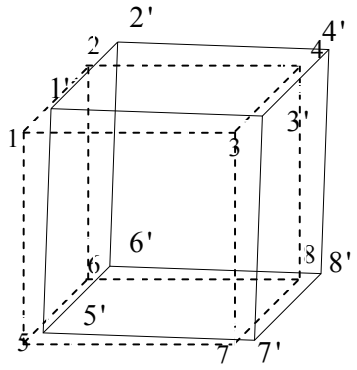


Fig. I. error calculation by standard transformation

Results of all algorithms can be compared with the standard transformation parameters which were provided by RIRE project, as shown in Figure I. All distances are measured in millimeters. For all 6 pairs CT-MR registration, $\sqrt{1.25^2 + 1.25^2 + 4.0^2} \approx 4.373(mm)$ is used as the standard of measuring whether reaching subvoxel accuracy or not, while $\sqrt{2.590723^2 + 2.590723^2 + 8.0^2} \approx 8.799(mm)$ for 6 pairs PET-MR registration. The mean values and standard deviations of distances for 50 runs of each registration solution are recorded in Table II, where those out sub-pixel standard results are in bold.

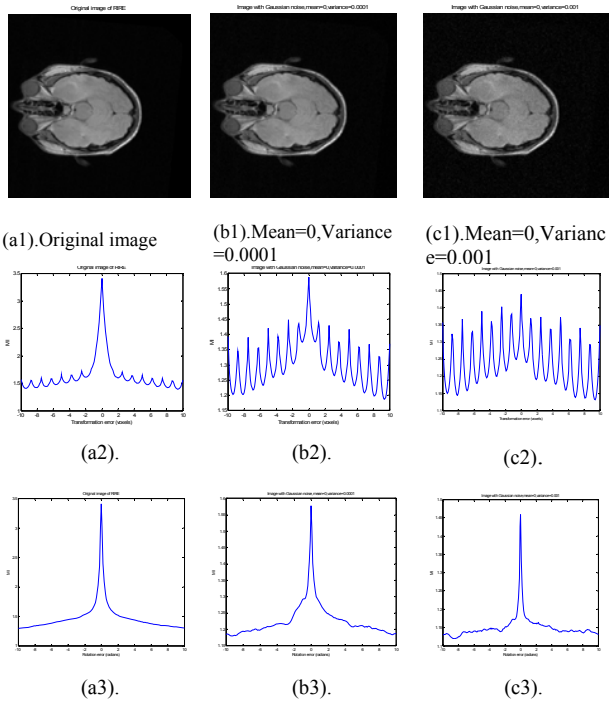


Fig. II. (a1) slice from PD MRI head volume. (b1) and (c1) corresponding image adding different Gaussian noises. (a2), (b2) and (c2) mutual information as a function of translation errors. (a3), (b3) and (c3) mutual information as a function of rotation errors.

Table II. Mean number and standard deviation of each algorithm.
Beyond sub-pixel results are in boldface

		Powell	PSO	QPSO	RQPSO
CT-PD	Mean	3.8070	6.6526	4.3393	3.7787
	St.Dev.	0	2.1038	1.5296	0.6275
CT_PD_r	Mean	227.0322	5.6223	5.6095	4.0592
	St.Dev.	0	0.0157	0.0121	1.2411
CT-T1	Mean	2.69636	3.7189	3.6861	2.8370
	St.Dev.	0	1.9408	2.0359	1.2856
CT-T1_r	Mean	147.29511	30.7214	13.5595	8.2910
	St.Dev.	0	0.0135	2.2463	1.3274
CT-T2	Mean	3.72182	4.3630	4.0219	3.8990
	St.Dev.	0	1.2853	1.3800	0.0108
CT_T2_r	Mean	6.51213	8.3525	4.7299	3.8165
	St.Dev.	0	1.9230	2.0571	0.2170
PET-PD	Mean	7.32626	8.1082	8.0929	7.8733
	St.Dev.	0	2.0311	1.1988	1.5072
PET-PD_r	Mean	12.84138	15.3906	8.0327	7.3222
	St.Dev.	0	1.8210	1.9853	1.2944
PET-T1	Mean	36.83191	11.852	10.3944	8.0169
	St.Dev.	0	2.1301	2.2507	1.7393
PET-T1_r	Mean	47.79181	17.7823	8.2901	7.8337
	St.Dev.	0	1.1095	2.3838	1.2432
PET-T2	Mean	7.10214	9.7373	8.4396	7.6574
	St.Dev.	0	0.8264	1.2884	1.2069
PET-T2_r	Mean	4.41907	8.9753	6.4027	4.2871
	St.Dev.	0	1.2933	1.5352	0.3921

As the results showed, the Powell is extremely unstable when dealing with images with noises, resulting good results in only 50% of the cases, while the other half were exactly unsatisfactory. As shown in Fig II, Powell made 4 inaccurate results which exceed the line of sub-pixel. The PSO on the other hand, was relatively stable, and however, the rate of generating results which didn't reach the sub-pixel standard was much higher than Powell. In terms of the accuracy, the QPSO largely performed better than POWELL and PSO, but the standard deviation is relatively large. Finally, for DRQPSO, although it was not capable of finding the best solution every time among all these four algorithms, its stability was very strong. Among the 12 groups of experiments, only one generated inaccurate results.

VI. Conclusion

A novel version of QPSO, RQPSO was proposed in this paper. As the hybrid distribution was used, the proposed RQPSO algorithm enhanced the global search ability of QPSO effectively. Performance comparison was made among RQPSO, QPSO, PSO and Powell by testing these algorithms on image registration problems. It was shown that the proposed RQPSO algorithm performed effectively in medical image registration, as has been shown by 12 sets of experiments on 3-D medical image registration with noises. It has also been shown that RQPSO worked most stably among all tested algorithms on image registration.

References

[1] J. Kennedy, R.C. Eberhart, "Particle Swarm Optimization", Proc. IEEE Int'1 Conference on Neural Networks, IV. Pisataway, NJ: IEEE Service Center, 1995, pp. 1942-1948.

[2] J. Kennedy, "Small worlds and Mega-minds: Effects of Neighborhood Topology on Particle Swarm Performance", Proc. 1999 Congress on Evolutionary Computation. Piscataway, NJ(1999) 1931-1938

[3] F. Van den Bergh, "An Analysis of Particle Swarm Optimizers", PhD Thesis. University of Prtoria, Nov 2001

[4] J. Sun et al, "Particle Swarm Optimizer with Particles Having Quantum Behavior". Proc. 2004 Congress on Evolutionary Computation, Piscataway, NJ (2004) 325-331.

[5] J. Sun et al, "A Global Search Strategy of Quantum-behaved Particle Swarm Optimization," Proc. 2004 IEEE Conference on Cybernetics and Intelligent Systems.

[6] J. Sun et al, "Adaptive Parameter Control for Quantum-behave Particle Swarm Optimization on Individual Level", Proceedings of 2005 IEEE International Conference on Systems, Man and Cybernetics, pp. 3049-3054.

[7] M. Clerc and J. Kennedy, "The particle swarm-explosion, stability and convergence in a multidimensional complex space," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 58–73, Feb. 2002.

[8] J. Sun et al, "Quantum-Behaved Particle Swarm Optimization: Analysis of the Individual Particle's Behavior and Parameter Selection," Evolutionary Computation, In Press.

[9] D.L.G. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," Phys. Med. Biol., vol. 46, pp. R1-R45, 2001.

[10] J.B.A. Maintz and M.A.Viergever, "A survey of medical image registration," Med. Image Anal., vol.2, pp. 1-37, 1998.

[11] L. Brown, "A survey of image registration techniques," ACM Comput. Surv.,vol. 24, pp.325-376,1992.

[12] D.L.G.Hill and P.Batchelor, "Registration methodology: concepts and algorithms," in Medical Image Registration, J.V.Hajnal, D.L.G..Hill,

and D.J.Hawkes, Eds. Boca Raton, FL: CRC, 2001.

[13] B.Likar and F.Pernus, "A hierarchical approach to elastic registration based on mutual information," Image Vis. Comput., vol. 19, pp. 33-44, 2001.

[14] J.L.Bernon, V.Boudousq, J.F.Rohmer, M.Fourcade, M.Zanca, M.Rossi, and D.Mariano-Goulart, " A comparative study of Powell's and downhill simplex algorithms for a fast multimodal surface matching in brain imaging," Comput. Med. Imaging Graph., vol. 25, pp.287-297, 2001.

[15] F.Maes, D.Vandermeulen, and P.Suetens, "Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information," Med. Image Anal., vol.3, pp. 373-386, 1999.

[16] M.Jenkinson and S.Smith, "A global optimization method for robust affine registration of brain images," Med. Image Anga., vol. 5, pp. 143-156, 2001

[17] G.K.Matsopoulos, N.A.Mouravliansky, K.K.Delibasis, and K.S.Nikita, "Automatic retinal image registration scheme using global optimization techniques," IEEE Trans. Inform. Technol. Biomed., vol.3, pp. 47-60, Mar. 1999.

[18] J.M.Rouet, J.J.Jacq, and C.Roux, "Genetic algorithms for a robust 3-D MR-CT registration," IEEE Trans. Inform. Technol. Biomed., vol.4, pp.126-136, June 2000.

[19] R.He and P.A.Narayana, "Global optimization of mutual information: application to three-dimensional retrospective registration of magnetic resonance images," Comput. Med. Imaging Graph., vol.26,no.4,pp.277-292, 2002.

[20] M.P. Wachowiak and A.S.Elmaghraby, "The continuous tabu search as an optimizer for 2D-to-3D biomedical image registration," in Lecture Notes in Computer Science, W.Niessen and M.Viergever, Eds. New York: Springer-Verlag, 2001, Proc. MICCAI 2001-2208, pp. 1273-1274.

[21] Y.Shi, R. C. Eberhart, "A Modified Particle Swarm", Proc. 1998 IEEE International Conferenceon Evolutionary Computation, pp. 1945-1950.

A new digital watermark algorithm based on the DWT and SVD[1]

Yuhui Li, Wei Gou and Bo Li
Kunming University of Science and Technology，Kunming 650500，China
zbpjlc@163.com

Abstract—**In this paper, we consider the application of the discrete wavelet transform (DWT) and singular value decomposition (SVD) in the digital watermark systerm.Based on this algorithm, it is efficient to protect the rights of the owners' digital content.This paper examines a technique for digital watermarking which utilizes properties of the DWT and SVD.The exeperimental results show that it is roubust against rotating, zoom, compression, adding noise, filtering and cropping.**

Keywords: digital watermark, DWT, SVD, roubust.

I. INTRODUCTION

With the development of internet and communication multimedia techniques, copyright protection of multimedia has become increasingly prominent[7]. Watermarking technology,the most effective method on the issue of copyright protection, has attracted more and more attention at present[6]. As a potential and effective way to solve this problem, digital watermarking becomes a very active research area of signal and information processing [8]. A digital watermark is a signal embedded in a digital image or video sequence which allows one to establish ownership, identify a buyer, or provide some additional information about the digital content[1].This paper proposes a steady image watermarking algorithm based on DWT domain.

In this paper, the next section will cover SVD, section III will cover wavelets, and then Section IV will focus on the Arnold chaotic method. The Section V and Section VI discuss the method used.Section VII presents results, and section VIII concludes the paper.

II. SVD

Singular value decomposition (SVD) is a kind of matrix diagonalization numerical method. And one of the most effective tool in the numerical linear algebra. It has been widely used in statistical analysis, signal and image processing , system theory and control and so on[3].

Let A be an $m \times n$ matrix. The singular value decomposition of A is the factorization $A = U \sum V^*$ where

U is $m \times m$ unitary (the left singular vectors of A)

V is $n \times n$ unitary (the right singular vectors of A)

Σ is $m \times n$ diagonal (the singular values of A)

The singular values reflect the intrinsic properties of the image, with strong stability; therefore the image won't have a big change when given a small disturbance. The feature, robustness of digital image watermarking, has vital significance [10].

III. DWT

Wavelet transform as a new subject branch of mathematical develops rapidly in recent years. Wavelet transform quickly replaces Fourier transformation of position that once regarded as "the most perfect analysis means" in signal processing field.

Various fields including mathmatics, quantus physics, electrical engineering, and seismic geology developed the concept of wavelets independently. Many analysis methods have applied in digital signal analysis, such as image compression and edge detection [2]. Compared with Fourier transform, the wavelet transform is space (time) and frequency of local transform, so it can effectively extract information from signal [3]. Through arithmetic functions such as translations anti dilations of function or signal multi-scale refined analysis[5].it can solve many difficult problems which the Fourier transform can not do so, therefore it also known as "math microscope".

The DWT separates an image into a lower resolution approximation image (LL) as well as horizontal (HL),vertical (LH) and diagonal (HH) detail components.For example,HL means that we used a high-pass filter,and a low-pass filter.Figure 4 illustrates this concept. If necessary, we can iterate the above method with LL again and again.The details at each succeeding octave are one-fourth the size of the previous octave.See Figure 1 and Figure 2.

[1] This work is partially supported by Yunnan science and technology plan projects (2008CA012-4) and Scientific and technological personnel service enterprise action items (2009GJF30050).

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.7

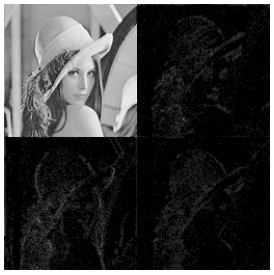


Figure 1. Example of DWT

LL	LH
HL	HH

Figure 2. Structure Diagram of DWT

As shown, a picture of most of information contained in the low-frequency coefficients, so a wartermark will be added in the low frequency content in this paper.

IV. Watermark scrambling

In this part, we main focus on the Arnold chaotic method which incread the system security. Mathematically the two-dimensional Arnold chaotic map is defined as the following:

Let $X_1 = \begin{bmatrix} X \\ Y \end{bmatrix}$ be the n × n matrix, then the Arnold transformation is,

$$\Gamma: \begin{bmatrix} X' \\ Y' \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} (\mathrm{mod}\, n)$$

Note: the mod is reminder the $(X+Y, X+2Y)$ and n.

We scramble the original watermark firstly, and choose the count of the scrambling as a key. See Figure 3 and 4.

Figure 3 the original watermark

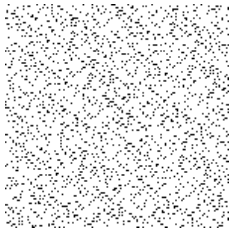


Figure 4 the scrambled watermark

V. Watermark embedding

1. First of all, according to the stated above Arnold scramble algorithin scrambles the watermark, and get the key K.Below we use SVD decompose the scrambled watermark.

$$\grave{W} = U_w S_w V_w^T \qquad (5\text{-}1)$$

The $S_w = diag(\delta_1, \delta_2, \cdots, \delta_p)$ and $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_n \geq 0$, σ_i (i=1,2,...,n) is the singular values of S_w. We need to properly preserve the U_w, V_w^T and K ， which will be used in the behind of the watermarking extraction algorithm.

2. The DWT separates the host image into a lower resolution approximation image (LL) as well as horizontal (HL), vertical (LH) and diagonal (HH) detail components. Then taking the low-frequency coefficient and they are arranged in descending order. It is convenient to express An (An is the n top maximum).

3. Then Adding the an into the lower resolution approximation image

$$A_n = A_n + \alpha \times S_w \qquad (5\text{-}2)$$

Where α is the scaling factor which decides the embedding strength.

4. Use $A_n^{'}$ instead of A_n and produce a new low-frequency coefficient matrix $A^{'}$.

We use the inverse discrete wavelet transform (IDWT) to reconstruct the image $A^{'}$.

5. Embedding process diagram as shown in figure 5.

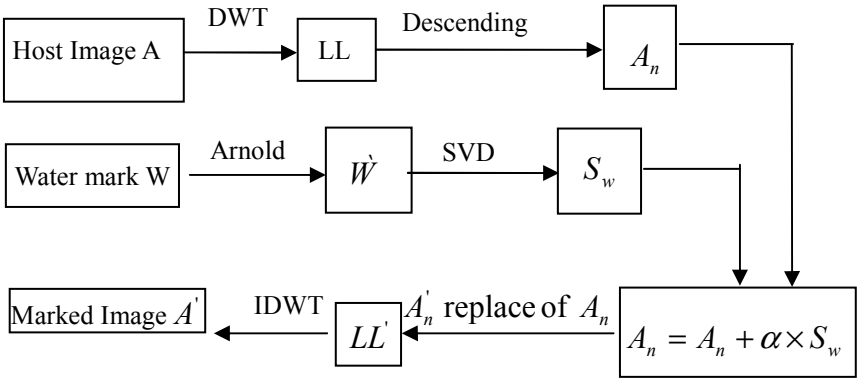


Figure 5. Watermark embedding

VI. WATERMARK EXTRACTING

1.We use the DWT separates the Marked image $\grave{A}$ into LL,HL,LH and HH,and sort LL vector into descending order (A_n^* ,n=1,2,…,n).

2. Then we can deduce the S_w^* and W^* by using the formulas below

$$S_w^* = 1/\alpha(A_n^* - A_n) \quad (6\text{-}1)$$ where α , A_n and A_n^* from the part V.

$$W^* = U_w S_w^* V_w^T \quad (6\text{-}2)$$

where W^* is the Scrambling watermark

3. Finally, we can get the wartermark by Arnold chaotic method (The secret key has reserved).

4. Extracting process diagram as shown in figure 6.

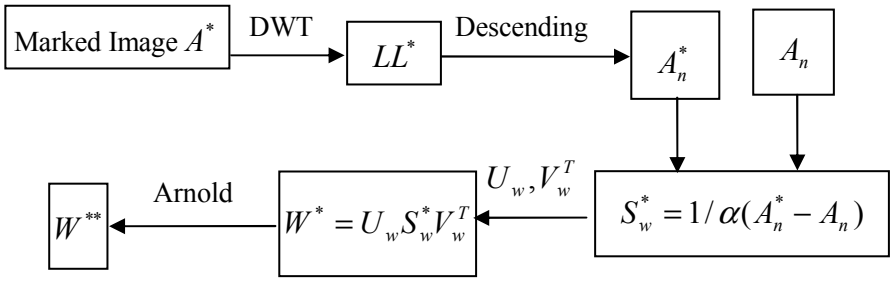


Figure 6. Extracting process

VII. EXPERIMENTAL RESULTS

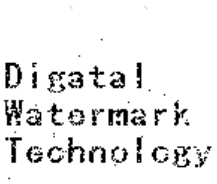
In this part, we will give the achieving results by using the above method.the host image seen in figure 1 and watermark seen in figure 2.firstly, Adding the watermark in the host image. secondly, we perform a variety of attack to marked image by StirMark4.0.finally， extracting the watermark in the attacked images. In the following， we will give the results of experiment.

Affine

snr: 16.702416; psnr: 16.097180

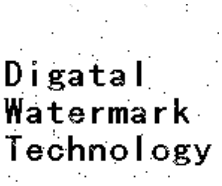
Crop

snr: 17.187082; psnr: 17.566863

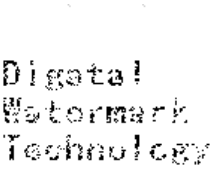
JPEG

snr: 14.691918; psnr: 14.990930

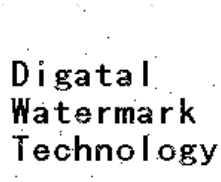
MEDIAN

snr: 17.306474; psnr: 17.689837

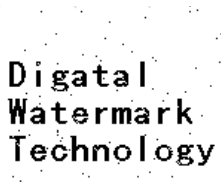
NOISE

snr: 12.552523; psnr: 12.748146

ROT

snr: 17.488619; psnr: 17.863347

PSNR

snr: 16.931616 psnr: 17.320465

From the above results, we can clearly see that the proposed algorithm has good resistance to attack.

VIII. CONCLUSION

The DWT and SVD enable the embedding of a watermark at a higher level, for both can give consideration to the host image and watermark recovered, visibility and objectivity much better. We measured the effect extracted by SNR and PSNR. In this paper, binary image was used as watermark, it also indicates that we can use color image as watermark in the further work. However, it remains true that the DWT and SVD have some advantages for the digital watermark, which can be good to resist geometric attacks and frequency domain attacks.

REFERENCES

[1] Xiangwei Kong, Yu Liu, Huajian Liu, and Deli Yang, Object Watermarks for Digital Images and Video, Image and Vision Computing, Elsevier, Vol.22, No.8,Aug 2004, pp.583-595.

[2] Evelyn Brannock,Michael Weeks,Robert Harrison,The effect of wavelet families on watermarking,journal of computers,vol.4,no.6,june 2009.

[3] I. Daubechies, The wavelet transform, time-frequency localization and signal analysis, IEEE Trans. Inform. Theory 36 (5) (1990), pp. 961-1004.

[4] E. Koch, J. Zhao, Towards robust and hidden image copyright labeling, Proc. IEEE Workshop on Non-Linear Processing, June 1995, pp. 452-455.

[5] P.-C. Su, C.-C. J. Kuo, and H.-J. M.Wang, Blind Digital Watermarking for Cartoon and Map Images, Proceedings of SPIE Security and Watermarking of Multimedia Contents, Vol.3657, 1999,pp.296-306.

[6] C.C. Chang, K.F. Hwang, M.S. Hwang, Robust authentication scheme for protecting copyrights of images and graphics, IEE Proceedings -Vision, Image and Signal Processing 149 (1) (2002) 43-50.

[7] J.T. Brassil, S.Low, N.F. Maxemchuk, L.o'Goman, Electronic marking and identification techniques to discourage document copying,IEEE Journal on Selected Areas in Communications,1995, pp. 1949-1504.

[8] G. Caronni, Assuring ownership rights for digital images, Proc. Reliable IT Systems, VIS 95, June 1995.

[9] Ikpyo Hong, Intaek Kim, Seung-soo Han. A blind watermarking technique using wavelet transforms, Proceedings of IEEE International Symposium on Industrial Electronics, Vol. 3, and 2001.1946-1950.

[10] Andrews H, Patterson C.Singular value decomposition (SVD) image coding, IEEE Transaction on Communications, 24(4) (1976), pp.425-432.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Fast Image Retrieval Method Based on Visual Word Tree Word

Zhu Liang
Computer Department. Chongqing Industry Polytechnic College,
Computer College. Chongqing University
Chongqing ,China
e-mail: Kcsj_2005@126.com

***ABSTRACT*—Fast content-based image retrieval is one of the core problems in multi-media technology and computer vision. Bag-of-words belongs to a currently very popular class of algorithms that work with an efficient visual vocabulary. However, the vocabulary is planar structure to limit the size of vocabulary, representativeness of words and lead to high computational cost. A hierarchical vocabulary scheme called visual word tree is presented. Firstly, features are extracted from training images and hierarchical k-means performed recursively on the descriptor vectors to build a visual word tree, which has k-branch factor and L-layers. The similarity scoring of a database image to the query image is accomplished using inverted files. Due to hierarchical structure, the visual word tree allows a larger and more discriminatory vocabulary to be used efficiently, which has a lower computational cost. We show experimentally a dramatic improvement in retrieval speed on Caltech-101object categories.**

Keywords-image retrieval; bag of words; visual word; hierarchical tree

I. INTRODUCTION

Fast image retrieval technology [1,2] is an important technology of Multimedia Search Engine and an important content for graphical analysis. Text-based Image Retrieval ((TBIR) and Content-based Image Retrieval (CBIR) are two basic image retrieval technologies. Initial retrieval systems usually use text messages (keywords and annotations) for retrieval with two disadvantages: there are subjective differences between understandings of people to images and the textual descriptions; rigmarole man-made tagging can not follow the updating speed of digital media, so multimedia retrieval can hardly be carried out on the internet. The purpose of Content-based Image Retrieval (CBIR) is, with given image for retrieve, to search and find out pictures conforming to the retrieve conditions according to the information of the retrieve and retrieve standards. The indexed approach of this method is to project the images to feature space through extracting the low-level features, then, the similarity of the two images is examined

Although advanced image search engines created conditions for fast searching for image information on the internet, but Content-based Image Retrieval (CBIR) provides a more accurate and direct approach for image data retrieve based on visual conditions, there are still some core problems to be researched. The major core problems are: how to extract low-level image characteristics to make holographic representations of the content information of the images being easy for follow calculation and more suitable for visual sensation of people; how to build up image representation mode to across the Semantic Gap [5] between low-level image characteristics and high-level image semantics as far as possible; how to design the organizing approach of image content characteristics and the image match codes to meet retrieve requirements with higher speed, high efficient and magnanimity.

Based on the idea of BoW, Bag of Words in this article [6], a new organizing mechanism for visual characteristics is built up: Visual Word Tree. Visual Word Tree is different with the general visual vocabulary in terms of the structures, in the visual word tree calculation, firstly, image features, k-branch factor and L-layers of the visual word tree are extracted, a hierarchical vocabulary scheme called visual word tree is produced by using Hierarchical Algorithms for clustering at each level. For example, using Hierarchical K-means Clustering to carry out clustering of all sample characteristics to obtain K numbers of classes (all sample characteristics are classified into K numbers of branches), recursively performing k-means clustering to the each branch to gain K sub-branches at next level until the appointed max L-layers is reached. Word Tree uses Inverse Document structure memory image in text retrieval and calculates Inverse Document Frequency to each tree node as representative value, then, simple Score Scheme is defined for comparing the similarity of the images. The new Visual Word Tree, due to its hierarchical structure, has optimal characteristics comparing with the traditional visual vocabulary with planar structure: 1. The computational cost for image retrieval and leaf nodes are of logarithmic growth, but in the planar structures, computational cost is of linear growth with the word number increases which influences efficiency of the algorithm. 2. The visual word tree has no limitation for the numbers of the visual words and the feature representational ability to the image is improved with the increase of the word numbers to obtain better retrieval. 3. The hierarchical tree structure has greater depth search capability comparing with the planar structures and is able to use various mature searching algorithm being more suitable for fast retrieval for images with mass storage. 4. The tree structure established a simple relationship between the visual words (father-son, brother-brother) which is facilitates image recognition and retrieval, but in the traditional Bag of Words mode, the words are assumed independent each other, much useful original image information is lost.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.44

II. BAG-OF-WORDS OF VISUAL IMAGE

In recent years, Bag-of-words in the text processing has been successfully introduced into image area, using Bag-of-Visual Words for representation of the image content has achieved good application effects in the areas of robot navigation, Web image search, image semantic modeling, scene classification and so on. The basic idea of the BoW is the assumption of an image is combined based on different distributions of low-level image characteristics of the independent so-called visual primitives, first, unsupervised clustering is carried out to low-level characteristics of the image to integrate the class centers into visual primitives (visual words), so the continuous high dimensional image feature space is quantified into manageable discrete Visual Vocabulary. The image analysis and understanding are carried out according to different distributions of the image on the Visual Vocabulary. This method provides a low dimensional, structuring and unified description framework for the image demonstrating an excellent performance in resolving problems of diversity of the classes, similarity between the classes, complicated backgrounds, partial occlusion, location differences, illumination changes and so on. Fig. 1 indicates the major calculation steps of image Bag of Words and the application process [7] in the image classification and recognition.

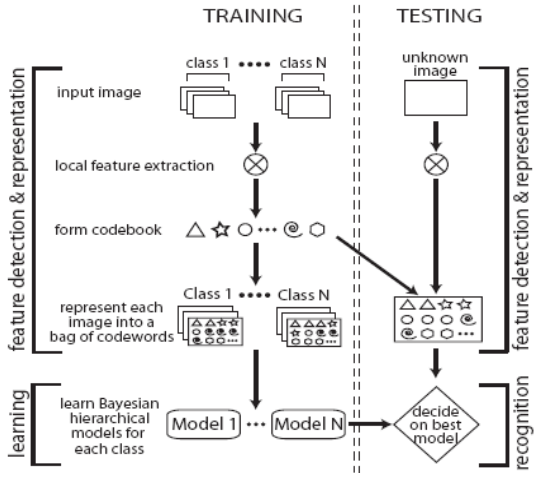


Figure 1. The flow of bag-of-words and image classification

III. CONSTRUCTION ALGORITHM FOR HIERARCHICAL VISUAL WORD TREE

A. Visual Word Tree Construction Based on Hierarchical K-means Clustering

The visual word tree based on hierarchical k-means clustering defined a hierarchical quantification method, unsupervised training is carried out by exacting characteristic description vector of the sample image sets to produce the tree. Different with traditional clustering, K, numbers of the classes, is defined as k-branch factor (numbers of branches or sons of each tree node). First, carrying out K-means clustering to feature sets of the training images, classifying the sample characteristics into K numbers of branches, recursively performing k-means clustering to the each branch to gain K sub-branches at next level until the appointed max L-layers is reached. In the application stages, each characteristic vector of the new image only needs to be compared with K numbers of candidate class centers layer by layer from top to bottom to select the closest class. Only K times of Dot Product are required to be carried out at each layer to make the whole process with the complex rate of KL times of dot product, the algorithm produces very good execution efficiency when K values are not high. Using hierarchical tree structure for defining the visual word tree is different with the planer structure of traditional visual vocabulary, it can be designed as a algorithm of nearest neighbor method for searching matched words more effectively, it also can define the follow-up, more flexible scoring for the computation of similarity of the images. In the planer structural vocabulary, the computational cost of algorithm of the whole image retrieval is increased sharply as the dimensionalities of the vocabulary are increased, but in the hierarchical word tree, the computational cost of algorithm and the leaf numbers are of logarithmic growth, the numbers of represented words at L-layer K-branch are: $\sum_{i=1}^{L} k^i = \frac{k^{L+1}-k}{k-1} \approx k^L$, if the ward is the characteristic vector of D-dimension, one word tree ocuppys Dk^L bytes memory.

B. Measurement of Image Similarity Based on Score Scheme

Once the visual word tree is constructed, the similarity codes need to be defined between the retrieval image and the database image, by using layered characteristics of the tree structure and inheritance relations between the nodes, we just need to compare the similarities of the paths from top to bottom in the word tree of the different images for determine the similarities between different images. Shown as Fig.2, the characteristic vector of each picture is projected to the visual word tree to obtain the responding path information, then, quantitative comparisons of the path similarities are performed by using the score scheme.

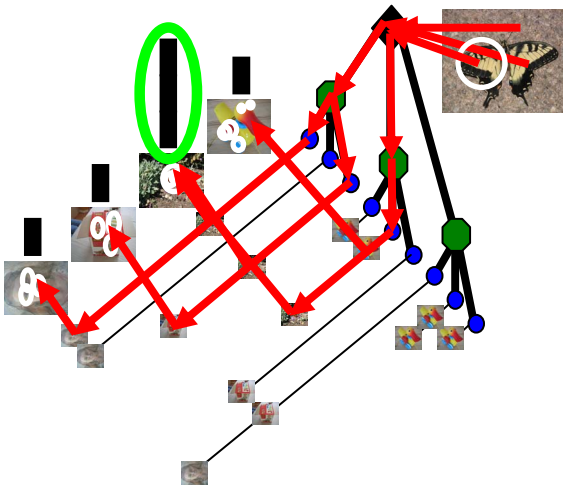


Figure 2. An illustration of the process of building the visual word tree and image matching

First, distributing one weight to each tree node of the visual word tree, then, defining the retrieval characteristic vector q_i and database characteristic vector d_i respectively as follows:

$$q_i = n_i w_i$$
$$d_i = m_i w_i \quad (1)$$

Where n_i and m_i respectively represent the numbers of the characteristics of the retrieval image and the database images passed node i. An empirical value can be designed for node weight w_i according to the experiment; in this article, node weight w_i is determined by using typical Inverse Document Frequency (IDF) weighting method of text retrieval.

$$w_i = \log \frac{N}{N_i} \quad (2)$$

N is the total numbers of the images in database, N_i is the image numbers with at least one characteristic path passed node i.

Similarity scoring S of each database image can be defined based on the difference between query vector and database vector:

$$s(q,d) = \left\| \frac{q}{\|q\|} - \frac{d}{\|d\|} \right\| \quad (3)$$

IV. Fast Image Retrieval Method Based On Visual Word Tree

For improving the scoring computational efficiency of large database, inverted file structure is used to indicate Bag-of-words of the image, each node of the visual word tree associated to an image list including the characteristic of the word and the emerged word frequency m_i of the characteristic in the corresponding images. In the algorithm execution process, just inverse documents of the leaf nodes need to be figured out, the inverted file structures of the intermediate nodes can be obtained by combining the child node inverted files of the corresponding branches. Reverse document length stored in the node is used for calculating weight w_i of the node, if the length is longer than the expected length; it means that the corresponding word of the node frequently emerged in the image with low representativeness, just like text processing, the words with less information can be removed as Stop Words. Supposing the representative value of each node has been obtained from known database training, the scoring vector of database image can be precomputed based on search path of the word tree and normalized as $\mathrm{d} = \frac{d}{\|d\|}$. Similarly, calculating the normalized scoring vector of query image in the visual word tree: $\mathrm{q} = \frac{q}{\|q\|}$, calculating the similarities between the query image and database images under L_p norm:

$$\begin{aligned} s(\mathrm{q},\mathrm{d}) &= \|\mathrm{q}-\mathrm{d}\|_p^p \\ &= \sum_i |q_i - d_i|^p \\ &= \sum_{i|d_i=0} |q_i|^p + \sum_{i|q_i=0} |d_i|^p + \sum_{i|q_i\neq 0, d_i\neq 0} |q_i - d_i|^p \\ &= \|\mathrm{q}\|_p^p + \|\mathrm{d}\|_p^p + \sum_{i|q_i\neq 0, d_i\neq 0} \left(|q_i - d_i|^p - |q_i|^p - |d_i|^p\right) \\ &= 2 + \sum_{i|q_i\neq 0, d_i\neq 0} \left(|q_i - d_i|^p - |q_i|^p - |d_i|^p\right) \end{aligned} \quad (4)$$

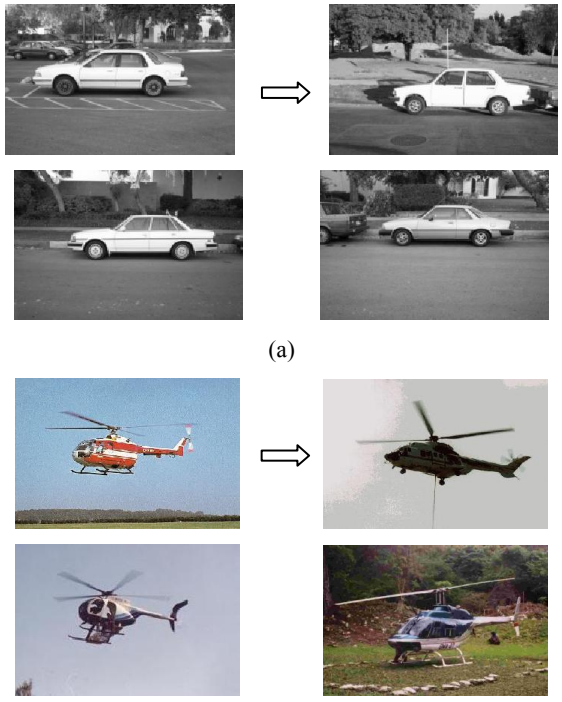
Under common 2-norn definition, similarity computing formula is simplified as:

$$s(\mathrm{q},\mathrm{d}) = \|\mathrm{q}-\mathrm{d}\|_2^2 = 2 - 2 \sum_{i|q_i\neq 0, d_i\neq 0} q_i d_i \quad (5)$$

The significance of the similarity computing formula is to significantly reduce the computational complexity of vector distance. Because of using reverse document index structure, the originally complicated distance calculation is simplified as just requiring through If Statement to perform accumulative summation to the elements on nonzero value corresponding dimensions in query vector and database vector. Due to the numbers of visual words are usually great, the query image library is also large, so the retrieval efficiency is significantly improved by this simplified calculation.

V. EXPERIMENTAL ANALYSIS

For the verification of image retrieval results of the visual word tree, in the experiment of this article, the most frequently used Caltech-101 image library [8] is selected as a labeled training sample set for carrying out the experiment. This standard image library has been collected and accomplished by Fei-Fei Li, Marco Andreetto, Marc Aurelio Ranzato and so on (on September, 2003) including 101 target class composition, the each class includes about 40~800 images of 300×200. Firstly, the image features are extracted by using SIFT feature detection algorithm [9] , using the algorithm of this article to establish visual word tree, k=10，L=6. Fig.3 is a query image with the given car side and helicopter and retrieved pictures with most similarities.

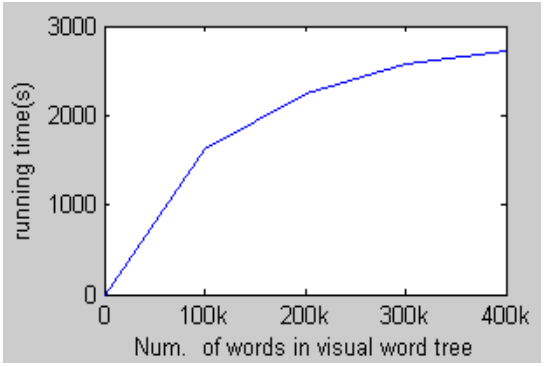


(a)

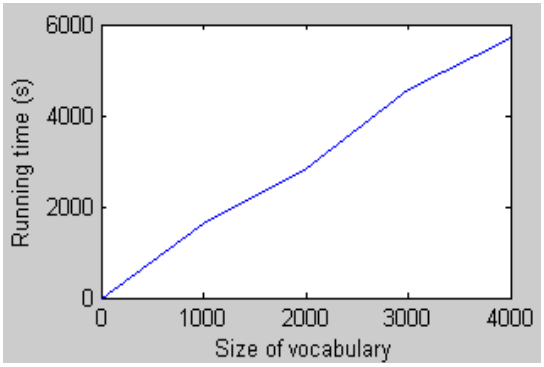
(b)

Figure 3. The 3 images have highest scores among retrieval results

Fig.4 is a comparison for algorithm retrieval efficiency. To verify the relationship between algorithm retrieval efficiency and the size of the vocabulary, we selected 10, 20, 30 and 40 pictures for each class of image to perform 4 tests, so the each test has 1010, 2020, 3030 and 4040 pictures. Fig. 4 (a) is the curve diagram of retrieval running time when numbers of the words in this article are respectively: 100K, 200K, 300K and 400K, it is shown that the running efficiency is with exponential growth as numbers of the words in the visual vocabulary increase. But Fig.4 (b) is the curve diagram of retrieval efficiency of visual vocabulary based on planar structure, it is shown that the calculation load is with linear growth when the sizes of the vocabulary are 1000, 2000, 3000 and 4000 respectively, much lower than the word numbers of the visual word tree.



(a)



(b)

Figure 4. The comparison of running time

VI. CONCLUSION

With the rapid development of Multimedia Technology, traditional text-based retrieval technology is gradually becoming unsatisfied, so content-based image retrieval has become a central issue for the retrieval research. Bag-of-words in text processing has been successfully applied in the various areas for image modeling and image analysis, the core of Bag-of-words is to build up a high efficient visual vocabulary, but the traditional visual vocabulary has a planar structure which has extremely limitations on the sizes of the vocabulary, characterizations of the words, computational cost and so on. Based on this idea, a new organizing mechanism of visual features is established in this article: visual word tree. The visual word tree has different construction comparing with the general visual vocabulary, for the algorithm of the visual word tree, firstly, the image features are extracted, branches of k and layers of L are defined, using Hierarchical Algorithms for clustering at each level to produce the vocabulary with the tree structure. Based on the tree structure, reverse document index in textual retrieval is used to represent images and calculating Inverse Document Frequency to each tree node as representative value to define the score scheme for matching the similarities. Because the layer structure is established, the new visual word tree, comparing with the traditional visual vocabulary with planar structure, has the advantages of low calculation load, unlimited word

numbers, fast search ability and so on. The rapidity and high efficiency of the algorithm in this article is verified in the retrieval test of Caltech-101 image library.

REFERENCES

[1] Xiangyoujun and Xieshengli, Overview of Image Retrieval Technology, (Natural Science) 2006, Vol.8(03)

[2] R. Datta, J. Dhiraj, J. Li, J. Z. Wang. Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys, 2008(40): 1-60.

[3] S. Deb, Y. C. Zhang. An Overview of Content-based Image Retrieval Techniques. AINA 2004

[4] Xuqing, Yangweiwei, Chengshengtan. Content-based Image Retrieval Technology, Computer Technology and Development, 2008(01)

[5] R. Zhao, W. I. Grosky. Bridging the Semantic Gap in Image Retrieval. Distributed multimedia databases: Techniques and applications, 2001: 14-36.

[6] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. ICCV, 2005: 604-610.

[7] F. F. Li, P. Perona. A Bayesian Hierarchical Model for Learning Natural Scene Categories. CVPR, 2005: 524--531.

[8] F. F. Li, R. Fergus, P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In IEEE CVPR, Workshop on Generative Model Based Vision, 2004. http://www.vision.caltech.edu/Image Datasets/Caltech101.

[9] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 2004, Vol.60(4): 91-110.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Quantitative Steganalysis Based on Wavelet Domain HMT and PLSR

Ziwen Sun
School of Internet of Things Engineering
Jiangnan University
Wuxi, China
sunziwen@jiangnan.edu.cn

Hui Li
School of Internet of Things Engineering
Jiangnan University
Wuxi, China
jndxlh0208@yahoo.com.cn

***Abstract*—Aiming at the problem of estimation of secret message length in steganalysis, this paper presents a quantitative steganalysis method based on HMT (Hidden Markov Tree) and PLSR (Partial Least Squares Regression) to solve the problem. In this paper, three 2-State HMT models are modeled respectively for wavelet coefficients in the horizontal, vertical and diagonal directions. In order to calculate the parameters of HMT, EM (Estimation and Maximization) algorithm is adopted to train the HMT models. The parameters are used as the 66-D feature of image. Then, the quantitative steganalyzer which is used to estimate the message length is established by combining HMT with PLSR. The proposed scheme is evaluated by constructing quantitative steganalyzers for F5, outguess and MB, simulation results demonstrate that these quantitative steaganalyzers can estimate the message embedding rates accurately and fast.**

Keywords-hidden markov tree model; partial least squares regression; embedding rate; quantitative steganalysis.

I. INTRODUCTION

Steganography is a technology to realize covert communication in the public channel. However, it may cause an immeasurable loss once they are used for illegal acts. Under such a background, steganalysis, the counter measure to steganography, attracts more and more attention in information hiding field and becomes a hotspot. According to different aims, steganalysis can be divided into three stages: 1) determine the presence of hidden information; 2) estimate the hidden message length and position; 3) extract the secret message combined with the cryptology. The current study is focused on the first two stages of steganalysis.

Most steganalysis algorithms only finish the first stage, which determine whether an image contains secret messages or not [1-6]. In [1], the transition probability matrix which describes correlations of the quantized DCT coefficients in the multi-directions was constructed, and two different calibrations were utilized to form 96 dimensional features. In [2], a steganalysis method based on the differential image's histogram in frequency domain was proposed. Three differential images were calculated in horizontal, vertical and diagonal directions respectively, and the features were extracted from the DFT of the histogram of differential images. In [3], statistical model based on high-order wavelet decomposition was built to distinguish stego images from cover images, and ANOVA (Analysis of Variance) was applied to test which wavelet statistics were more sensitive to hidden message. In [4], 17 wavelet subbands were obtained based on three-level wavelet decomposition including the further decomposition of the first scale diagonal subband. The first three order statistical moments of CF (characteristic function) of each band were selected to form 51 dimensional features. In [5], an image was decomposed at first, and then the co-occurrence matrix of the adjacent wavelet coefficients was calculated and Laplace transform was applied to the co-occurrence matrix, at last, the Laplace transformed variances and CF moments of co-occurrence matrix were combined as image features. In [6], six kinds wavelet filters namely Daubachies, Coiflets, Symlets, Discrete Meyer, Biorthogonal and Reverse Biorthogonal families were utilized in steganalysis and evaluated by detecting F5 and MB steganography.

In recent years, few literatures proposed steganalysis algorithms that can estimate message length [7-10]. In [7], SPA (Sample Pair Analysis) was proposed to detect LSB (least significant bit) steganography and estimate the message length. In [8], a steganalysis technique was proposed to estimate the LSB embedding message rates by constructing equations with the statistics of difference image histogram. The results showed that their method was more accurate and reliable than other difference image histogram methods. Steganographic algorithm based on QIM (quantization index modulation) has the following phenomena: values at jumping points vary as embedding rate changes and the distance between two jumping points varies as the quantization step changes. In [9], an estimation equation been used to estimate the secret message length of QIM-based data hiding was constructed by analyzing above phenomena quantitatively. But all above quantitative steganalysis algorithms rely on full knowledge of the embedding algorithms. In [10], a novel universal quantitative steganalysis method was proposed by constructing the mapping model between features and embedding change rates. The embedding rates of testing image were predicted according to the testing image features and the mapping model.

Wavelet transform has nice time-frequency local and multi-resolution characteristic, and is widely used in image processing field [3-6, 11]. This paper constructs image features based on wavelet domain HMT and adopts PLSR as regression analyzer to train the mapping model, and the obtained mapping model is called quantitative steganalyzer.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.72

II. PROPOSED APPROACH

A. Feature Extraction Based on Wavelet Domain HMT

The input image is decomposed into high frequency subbands (vertical, horizontal, diagonal, denoted by LH, HL, HH respectively) and low frequency subbands by wavelet filter banks. Repeat this process to obtain Multi-scale wavelet coefficients w^j, where j is the decomposition scale. "Haar" wavelet is orthogonal and tight supported, and moreover "Haar" wavelet filter performed well in steganalysis [6], so "Haar" wavelet is chosen in this paper.

Wavelet coefficients have two properties, namely clustering and persistence. Clustering means: if a given wavelet coefficient is large/small, and its adjacent coefficients are very likely to be large/small. Persistence means: large/small value of wavelet coefficients tend to propagate across scales [11]. HMM (Hidden Markov Model) is a doubly stochastic model: the state is not directly visible, but the observation, which depends on and reflects the state, is visible. HMM is widely used in the signal estimation, classification and prediction field. Wavelet domain HMT model is a kind of HMM that describes the dependencies of multi-scale wavelet coefficients through the state probabilities of the wavelet coefficients. HMT model not only fully considers the clustering within scale and persistence across scale of wavelet transforms, but also shows the statistical dependencies and non-Gaussian statistics of wavelet coefficients [11].

Wavelet coefficient w is viewed as observation value and s denotes the corresponding state variable. This paper models each wavelet coefficient as being in one of two states: S_1 denoting the corresponding coefficient containing "high" signal energy, or S_2 denoting the corresponding coefficient containing "low" signal energy. $p(s)$ is the PMF (Probability Mass Function) of state variables s, and satisfies (1):

$$p(s = S_1) + p(s = S_2) = 1 \tag{1}$$

Under the condition of the given hidden state S_m (m=1,2), "high" state is considered as high-variance (σ_{S_1}), zero-mean ($\mu_{S_1} = 0$) Gaussian model while "low" state is considered as low-variance (σ_{S_2}), zero-mean ($\mu_{S_2} = 0$) Gaussian model. As a result, the wavelet coefficients fit two-state mixture Gaussian distribution [11]. The PDF (Probability Density Function) of wavelet coefficients is given by (2):

$$f(w) = \sum_{m=1,2} p(s = S_m) f(w \mid s = S_m) \tag{2}$$

where $p(s = S_m)$ denotes the probability of the state s is S_m, $f(w \mid s = S_m)$ denotes the conditional probability of the hidden state of w is S_m.

The dependencies of wavelet coefficients across scale are described using a probabilistic tree between the hidden state variable of parent node and the state variable of each of its child nodes in HMT. For n-dimensional signal, each parent hidden state is connected to its 2^n child states, e.g image is a kind of 2-dimensional signal, and the HMT model for the image has a natural quadtree structure as showed in Fig. 1. A node at scale J has its child notes at scale J-1, node (w_1, s_1) at scale J is the parent of the nodes $\{(w_2, s_2), (w_3, s_3), (w_4, s_4), (w_5, s_5)\}$ at scale J-1.

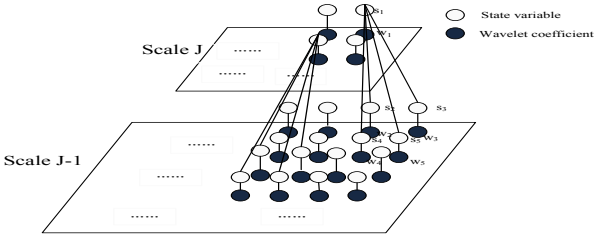


Figure 1. wavelet domain HMT for image Quadtree. Each black node represents a wavelet coefficient and each white node represents a hidden state variable.

The state transition probabilities from parent node to its child nodes are defined as (3):

$$A^{S_m \to S_{m'}} = p_j(S_{m'} \mid S_m) \qquad j = 2,3,\ldots,J \tag{3}$$

where $p_j(S_{m'} \mid S_m)$ denotes the transition probabilities from the parent node at scale j to its child nodes. S_m and $S_{m'}$ ($m, m' \in \{1,2\}$) denotes the hidden state of parent node and the hidden state of child node respectively, J is the maximum decomposition level. For a $N \times N$ image, J is calculated as (4):

$$J = \log_2(N) \tag{4}$$

The PMF of hidden states of child nodes can be calculated by the PMF of hidden states of parent nodes and state transition probabilities as (5):

$$\begin{aligned} p(s_j = S_m) &= \sum_{m,m'=1,2} p(s_{j+1} = S_{m'}) p(s_j = S_m \mid s_{j+1} = S_{m'}) \\ &= \sum_{m,m'=1,2} p(s_J = S_{m'}) p(s_{J-1} = S_m \mid s_J = S_{m'}) \cdots p(s_{j+1} = S_m \mid s_j = S_{m'}) \end{aligned} \tag{5}$$

where $p(s_j = S_m)$ denotes the PMF when the hidden states of wavelet coefficients at scale j is S_m, $p(s_j = S_m \mid s_{j+1} = S_{m'})$ denotes the state transition probabilities from node w^j to its child notes w^{j-1}.

The wavelet domain HMT parameter set θ, which best characterize the wavelet coefficients, includes initial states probabilities, transition probability matrix and variances. For each high-frequency direction, the parameter set θ of HMT can be described as (6):

$$\begin{aligned} \theta = \{ & p(s_J = m), \sigma_j(m), A_2^{S_{m \to m'}}, A_3^{S_{m \to m'}}, \cdots A_J^{S_{m \to m'}} \mid \\ & m, m' \in \{1,2\}, j = 1,2,\cdots,J \} \end{aligned} \tag{6}$$

According to (6), if the image is decomposed into n-level of Haar wavelet and the hidden states of wavelet coefficient are divided into two states: "high" and "low", then the numbers of initial states probabilities, transition probability matrix and variances are $m, m \times n, m \times m \times (n-1)$ respectively. So there are $3 \times [m + m \times n + m \times m \times (n-1)]$ features can be obtained in three directions.

EM algorithm is adopted to calculate the parameter set θ of dependence tree model [11, 12]. According to the

clustering property, wavelet coefficients at the same scale and the same direction are modeled as a mixture Gaussian model, and computational process are greatly simplified.

B. Quantitative Steganalysis Based on PLSR

PLS (Partial Least Square) is functionally similar to CCA (Canonical Correlation Analysis) and realizes modeling through orthogonal score vectors [13, 14]. In recent years, PLS has been developed rapidly on account of the development of theory and the need of application, and it has a wide range of applications, including modeling, regression and classification.

PLSR is an important branch of PLS applications, which combines the advantages of CCA, PCA (principal component analysis), and multiple linear regression. PLSR is a multivariate statistical analysis method to construct model between independent variables and dependent variables, and it is applied to model a relation between parameter group θ=$\{\theta_{LH},\theta_{HL},\theta_{HH}\}$ and message embedding rates in this paper. Suppose the independent variables are stored in a matrix X, and the dependent variables are collected in another matrix Y, PLS decomposes the matrixes X and Y as (7) and (8) respectively [13]:

$$X = TP^T + E \tag{7}$$

$$Y = UQ^T + F \tag{8}$$

where T and U are score vectors, and $T^TT = I, U^TU = I$, I is an identity matrix, P and Q are the matrixes of loadings, E and F are the matrixes of residuals. The classical PLS method finds weight vectors w and c as formula (9):

$$[Cov(t,u)]^2 = [\mathrm{cov}(Xw, Yc)]^2 = \max_{|r|=1,|s|=1} [\mathrm{cov}(Xr, Ys)]^2 \tag{9}$$

where $Cov(t,u) = t^Tu$ denotes the sample covariance between the score vector t and u. Larger $Cov(t,u)$ corresponds to more correlation between X and Y. loading vectors p,q can be calculated as (10) :

$$p = X^Tt/(t^Tt), \quad q = Y^Tu/(u^Tu) \tag{10}$$

X and Y are deflated as formula (11) until convergence:

$$X = X - tp^T \qquad Y = Y - uq^T \tag{11}$$

Suppose the input matrix is X_{test}, and the corresponding output of predict model is:

$$\tilde{Y}_{test} = X_{test}X^TU(T^TXX^T)^{-1}T^TY \tag{12}$$

In this paper, image features are consider as input variables and corresponding message embedding rates are considered as output variables. PLSR is utilized to construct quantitative steganalyzer as follows:

1) Embed secret message into cover image by using steganographic tools and obtain stego images, a half of stego images are selected for training and the remaining half of stego images are used for testing.

2) Decompose images into three-level of Haar wavelet, calculate parameter group θ= $\{\theta_{LH}, \theta_{HL}, \theta_{HH}\}$ in wavelet domain HMT as image features.

3) Input the training features and the corresponding embedding rates into PLSR and construct quantitative steganalyzer.

4) Input the testing features into the quantitative steganalyzer, which is obtained at step *3)*, and output predictive values.

The modeling and predicting processes of quantitative steanalyzer are showed in Fig. 2:

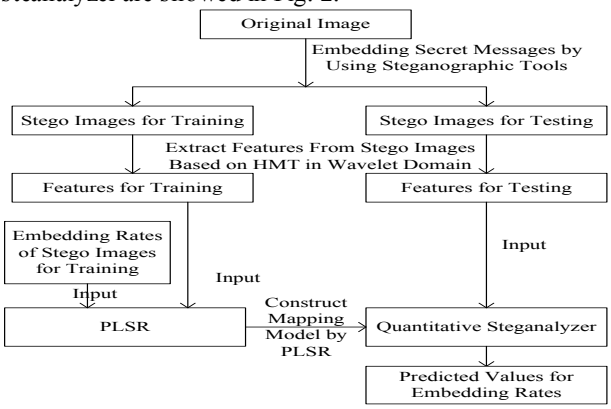


Figure 2. the modeling and predicting processes of quantitative steanalyzer

The accuracy of estimation is not linear in the number of features, besides, too many features sometimes can cause over fitting-phenomena. PLSR realizes dimension reduction like PCA based on K-L (Karhunen Loeve) Transform. In order to avoid the problem of over-fitting, CV (cross validation) is also used in this algorithm, and the results will be more credible.

C. Accuracy Evaluation Methods

In order to validate the accuracy of estimation, absolute error (13), variance (14) and median absolute error (15) are adopted to be the measure of the estimation accuracy:

$$E_1 = \frac{1}{l}\sum_{i=1}^{l}|y_i - \hat{y}_i| \tag{13}$$

$$E_2 = \frac{1}{l}\sum_{i=1}^{l}(y_i - \hat{y}_i)^2 \tag{14}$$

$$E_3 = mediam(abs(\hat{y} - mediam(\hat{y}))) \tag{15}$$

where y_i is the actual embedding rate, and $\hat{y}_i$ is the obtained estimation, and l is the sample number.

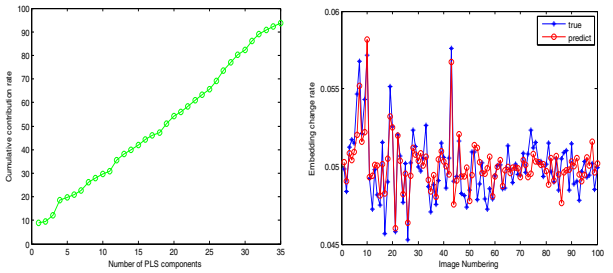
III. SIMULATION RESULTS

200 original images are randomly selected from the acknowledged JPEG database [14]. According to the actual requirements of the proposed algorithm, these selected images are readjusted to the size 512×512. Three type steganographic tools: F5 [16], outguess [17] and MB [18] are utilized to embed messages into images. Because the image features are more sensitive to embedding changes rather than the length of the message, the embedding rate means the ratio between the number embedding changes and the element number can be used for embedding. F5 and MB1 are forced to produce a predefined set 10 embedding rates $\{0.05, 0.10, \cdots, 0.50\}$ and outguess is forced to produce a predefined set 3 embedding rates $\{0.05, 0.10, 0.15\}$, then

23 groups of stego image are obtained, and each group has 200 images.

The image is decomposed into 4-level of Haar wavelet at first, and we can obtain 4 wavelet subbands in each high-frequency direction. This paper models wavelet coefficients as a two-state mixture model, that is to say, wavelet coefficients are divided into two states: “high” and “low”. Then, a 66-dimensional image feature vector based on wavelet domain HMT is obtained. Finally, PLSR is adopted to construct the mapping model between image features and embedding rates.

In the modeling process, PLSR achieves feature reduction too. Fully consider the test data volume and computational complexity, this paper adopts 5-fold cross-validation and selects 92% as CCR (accumulative contribution rate) threshold of principle component. Take a stego image set as an example, which was embedded by F5 steganography and the embedding rate was 0.05, Fig. 3(a) shows the relationship between the number of components and CCR in the process of modeling, and Fig. 3(b) shows the curves of predictive embedding rates and true embedding rates of testing images.



(a) (b)

Figure 3. (a) the relationship between the number of components and accumulative contribution rate in the process of modeling for 5% F5 stego images. (b) the curves of predictive embedding rates and true embedding rates of 5% F5 testing stego images

As showed in Fig. 3(a), the feature dimension is reduced to 35 under the conditions stated above, and computational process is simplified too. Fig. 3(b) shows that the regression curve closely fit the true curve, and the corresponding absolute error, variance and median absolute error are 8.40e-03, 1.18e-06 and 9.71e-05. Those results demonstrate that the proposed quantitative steganalysis is reliable.

We perform the process of modeling for each algorithm and embedding rate under the same conditions. The absolute error (*E1*), variance (*E2*) and median absolute error (*E3*) are calculated and showed in table 1. Rate means the message embedding rate.

We also adopt SVR (support vector regression) as the regression analysis tool to construct quantitative steganalyzer [10] under all the same operating conditions, and the comparison of predictive results based on PLSR and SVR were showed in table 2. The RAT means the regression analysis tool method, $\bar{E}1$, $\bar{E}2$ and $\bar{E}3$ are the averages of *E1, E2* 和 *E3* respectively.

Table1 shows the proposed quantitative steganalysis in this paper performs well, their predictive values and real values are close. The results demonstrate that the proposed quantitative steganalyzer based on PLSR is feasible.

Table2 shows the accuracies of two kinds of regression estimation are very close. The median absolute errors of predictive values based on SVR are smaller than that based on PLSR. However, the variance of predictive values based on PLSR is smaller. In time, PLSR reflects greater advantage. Under the same conditions, SVR as the regression model parser spend more than 35 minutes, while PLSR as the regression analyzer needs no more than 40 seconds. That is because SVR transforms the mapping relationship of the image features and embedding rate to the linear problems of high dimensional space, and the computation of linear solving in high dimensional space is large, PLSR iteratively solves the mapping relationship between image features and the information embedding rate in the feature space directly, and reduce the dimension of features, the amount of computation is greatly reduced. In summary, the proposed quantitative steganalysis method with fast operation speed, high accuracy.

TABLE I. THREE KINDS OF PREDICTION ERROR FOR EACH ALGORITHM AND EMBEDDING RATE

	MB steganography			F5 steganography			Outguess steganography		
Rate	*E1*	*E2*	*E3*	*E1*	*E2*	*E3*	*E1*	*E2*	*E3*
0.05	1.58e-03	4.89e-06	1.18e-04	8.40e-03	1.18e-06	9.71e-05	6.24e-03	6.77e-07	7.29e-05
0.10	1.01e-02	2.05e-06	8.67e-05	9.94e-03	2.09e-06	9.64e-05	3.98e-03	3.09e-07	3.89e-05
0.15	4.77e-03	3.85e-07	4.65e-05	3.68e-03	2.20e-07	5.43e-05	9.19e-03	2.84e-06	7.32e-05
0.20	6.17e-03	6.79e-07	6.42e-05	5.90e-03	5.94e-07	5.35e-05	--	--	--
0.25	7.51e-03	1.14e-06	4.41e-04	1.77e-02	1.75e-05	5.94e-04	--	--	--
0.30	7.41e-03	1.02e-06	7.49e-05	7.14e-03	9.24e-07	6.93e-05	--	--	--
0.35	4.43e-03	3.73e-07	3.89e-05	4.73e-03	4.16e-07	3.91e-05	--	--	--
0.40	5.54e-03	5.28e-07	5.41e-05	5.33e-03	4.71e-07	4.50e-05	--	--	--
0.45	4.18e-03	3.08e-07	4.71e-05	3.69e-03	2.68e-07	3.94e-05	--	--	--
0.50	5.18e-03	4.57e-07	5.07e-05	4.57e-03	3.94e-07	6.27e-05	--	--	--

TABLE II. THE COMPARISON OF PREDICTIVE RESULTS BASED ON PLSR AND SVR

	MB steganography			F5 steganography			Outguess steganography		
RAT	$\bar{E}1$	$\bar{E}2$	$\bar{E}3$	$\bar{E}1$	$\bar{E}2$	$\bar{E}3$	$\bar{E}1$	$\bar{E}2$	$\bar{E}3$
SVR	2.09e-02	3.74e-06	3.77e-05	2.22e-03	5.33e-05	4.50e-06	8.80e-03	3.71e-06	3.54e-05
PLSR	5.68e-02	1.18e-06	1.02e-04	7.11e-03	2.41e-06	1.15e-05	6.47e-03	1.28e-06	6.17e-05

IV. CONCLUSION

The problem of secret message length estimation is an important stage in steganalysis field. The modeling method is applied to the quantitative steganalysis in this paper, and a quantitative steganalysis algorithm based on wavelet domain HMT and PLSR is proposed. Base on the clustering and persistence of wavelet coefficients, the parameter set of Gaussian mixture model with HMT is calculated as the features, PLSR is utilized to construct the mapping model between image features and embedding rates. The results show that this algorithm is not only fast, but also more accurate estimates. However, the wavelet domain HMT is obtained through the iterative EM algorithm for parameters set, extracting features costs long time, the image embedded change rate is in the vicinity of a fixed value, and the scalability is poor. Future research priorities are: (1) Find more effective feature, called for the change after the embedded information is sensitive and extraction speed is fast ; (2) Not only estimate the embedded change rate, also further identify the type of steganography algorithm, embedding position, even extract the secret message.

ACKNOWLEDGMENT

This paper is supported by "the Fundamental Research Funds for the Central Universities". (Project Number: JUSRP21131)

REFERENCES

[1] Y .Wang, J.F.Liu, W.M.Zhang. "Blind JPEG Steganalysis Based on Correlations of DCT Coefficients in Multi-Directions and Calibrations". Proc. 2009 International Conference on Multimedia Information Networking and Security(MINE'09), IEEE press, Nov. 2009, pp. 495-499. doi: 10.1109/MINES.2009.135

[2] Q.L.Deng, J.J.Lin. "A Universal Steganalysis Using Features Derived from the Differential Image Histogram in Frequency Domain", Proc.International Congress on Image and Signal Processing, 2009 (CISP 2009), IEEE press, Oct.2009, pp.1-4. doi: 10.1109/CISP.2009.5304562

[3] S.H.Zhan, H.B.Zhang. "Blind Steganalysis Using Wavelet Statistics and ANOVA". Proc. of the sixth International Conference on Machine Learning and Cybernetics, IEEE press, Aug.2007, pp.2515-2519. doi: 10.1109/ICMLC.2007.4370570.

[4] Z.W. Sun, H.Li, Z.J.Wu, Z.P.Zhou. "An Image Steganalysis Method Based on Characteristic Function Moments of Wavelet Subbands". Proc. 2009 International Conference on Artificial Intelligence and Computational Intelligence(AICI'09). IEEE press, Nov.2009, pp. 291-295.doi: 10.1109/AICI.2009.185

[5] H.Zong, J.F.Liu, X.Y.Luo. "A Wavelet-based Blind JPEG Image Steganalysis Using Co-occurrence Matrix". Proc.11th International Conference on Advanced Communication Technology, 2009 (ICACT 2009), IEEE press, Feb.2009, pp.1933-1936.

[6] Z. Khan, A.B. Mansoor. "An Evaluation of Wavelet Filters Performance for Steganalysis". Proc. 2009 2nd International Conference on Computer, Control and Communication (IC4 2009) IEEE Press, Feb.2009, doi:10.1109/IC4.2009.4909227.

[7] D.Sorina, X.L. Wu, Z.Wang. "Detection of LSB Steganography via Sample Pair Analysis". Proc Information Hiding, 5th International Workshop, Lecture Notes in Computer Science, IEEE press, July. 2003, pp.355-372. doi:10.1109/TSP.2003.812753.

[8] S. K. Jena, G.V.V. Krishna. "Blind Steganalysis: Estimation of Hidden Message Length". Proc of International Journal of Computers, Communications & Control. CCC press, 2007, pp. 149-158.doi: 10.1.1.134.4333

[9] X.Y.Yu, A.M.Wang. "Detection of Quantization Data Hiding". Proc. 2009 International Conference on Multimedia Information Networking and Security (MINES'09). Nov.2009, pp.45-47. doi: 10.1109/MINES.2009.272

[10] T. Pevny, J. Fridrich, A. D. Ker. "From Blind to Quantitative Steganalysis". Proc. SPIE, Electronic Imaging, Media Forensics and Security XI. SPIE press,. Jan.2009,pp.1-14.

[11] M. S. Crouse, R. G. Baraniuk. "Wavelet-Based Statistical signal Processing Using Hidden Markov Models". Proc.Transactions on Signal Processing, IEEE press, Apr.1998, pp.886-902, doi: 10.1109/78.668544

[12] J. K Romberg. "A Universal Hidden Markov Tree Image Model". A Thesis Submitted in Partial Fulfillment of The Requirements for The Degree Master of Science. Rice University, Houston, Texas.May, 1999.

[13] R. Rosipal, N. Kramer. "Overview and Recent Advances in Partial Least Squares". Proc. Statistical and Optimization Perspectives Workshop(SLSFS 2005). LNCS, vol.3940, 2006, pp.34-51. doi: 10.1.1.85.7735.

[14] H.Abdi. "Partial least squares regression and projection on latent structure regression (PLS Regression)". Wiley Interdisciplinary Reviews: Computational Statistics. Vol.2, Jan.2010, pp.97-106. doi:10.1002/wics.51.

[15] A.R.S.Goldenstein,W.S.T .Boult."The Unseen Challenge Data Sets". Computer Vision and Pattern Recognition Workshops (CVPRW'08) IEEE press, June.2008, pp.1-8, doi:10.1109/CVP RW.2008.4562987. http://www.liv.ic.unicamp.br/wvu/datasets.php.

[16] A. Westfeld. "F5-A Steganographic Algorithm High Capacity despite Better Steganalysis".Information Hiding,vol.2137,April.2001,pp.289-302, doi:10.1007/3-540-45496-9_21. http://www1.inf.tudresden.de/~aw4/ publikationen.html.

[17] N. Provos. "OutGuess-Universal Steganography tool". http://www. Outguess.-org/ 2001.

[18] P. Sallee. "Model-Based Steganography". Proc. International Workshop on Digital Watermarking(LNCS2939), 2004, pp.154-167.doi: 10.1007/978-3-540-24624-4_12.

Distributed Fusion Steganalysis Based on Combination System Likelihood Function

Ziwen SUN
School of Internet of Things Engineering
Jiangnan University
Wuxi, China
sunziwen@jiangnan.edu.cn

Jiajie LIU, Zhicheng JI
School of Internet of Things Engineering
Jiangnan University
Wuxi, China
liujiajie20064300@163.com,zcji@jiangnan.edu.cn

***Abstract*—This paper presents a distributed fusion decision-making steganalysis method for attacking image steganography. The fusion rule of combination system likelihood function is derived based on the smallest error rate Bayesian decision-making rule. According to information fusion theory, by introducing fusion decision-making theory into steganalysis research, it will achieve more effective, more reliable and more stable steganalysis detection effect than just adopting a single local decision-making. The image characters are first extracted from spatial domain, DCT domain and DWT domain, and then the final result is obtained by using fusion rule. The simulation results show the fusion steganalysis method significantly improves the detection performance of local decision-making through introducing the fusion rule of combination system likelihood function.**

Keywords- Steganalysis; Bayesian decision-making; fusion decision-making; combination system likelihood function

I. INTRODUCTION

Steganalysis is the art of discovering the presence of hidden data in cover objects. Various steganalysis methods can be divided into special steganalysis and universal steganalysis. Compared with specific steganalysis techniques, universal steganalysis techniques have become more attractive since they work independent of the embedding techniques. Universal steganalysis can be considered as two-class images pattern recognition problem. At present, the studies about steganalysis mainly focus on feature extraction and design (or selection) of classifier. A majority of literatures extract relative sensitive features from spatial domain, DCT (Discrete Cosine Transform) domain and DWT (Discrete wavelet transform) domain respectively. Features extracted from a data domain are sensitive to steganographic embedding techniques operated in the same data domain, and insensitive to steganographic embedding techniques operated in other domains. The generality of steganalysis would be hardly improved if we only consider the feature extraction methods, such as F5、JPHide and JSteg can resist the low-order statistical analysis while YASS [1] and its improved algorithm [2] can resist typical universal steganalysis methods.

In recent years, information fusion theory is used in steganalysis in a few literatures. In 2006, Kharrazi et al [3] used the max and mean fusion rules in the decision-level , and their experimental results showed the mean rule outperforms the max rule for steganalysis. In 2007, Rodriguez et al [4] used variance weighted fusion and Gaussian weighted fusion for steganalysis. In 2008, Rodriguez et al [5] applied multiple steganalysis systems and Bayesian model averaging fusion to identify the embedding algorithm used to create a stego JPEG image. And in the same year, Jing et al [6] detected six typical steganographic algorithms effectively by using BFS (Boosting Feature Selection) algorithm as the fusion tool to form fusion feature vector in the feature-level. In 2009, we[7] calculated feature vectors from difference arrays along three different directions from quantized block DCT coefficients, and made use of a weighted feature fusion method to fuse above feature vectors. In 2009 and 2010, Christian and Jana [8-9] used information fusion in audio Steganalysis. Obviously, the steganalysis methods which have used information fusion in decision-level or feature-level performed better than other steganalysis methods. Above literatures showed many different implementation methods of information fusion for steganalysis, but they all lacked strict theoretical derivation.

In this paper, the fusion rule of combination system likelihood function and its usage in steganalysis are proposed based on the smallest error rate Bayesian decision-making rule. And a distributed fusion decision-making steganalysis method is used in this paper. Different kinds of feature vectors are firstly extracted from each domain, e.g., spatial domain, DCT domain and DWT domain, then posterior probabilities of three local decision-making results were obtained by input above features to SVM respectively, and at last global decision-making rusult is obtained by fusion the three local results. As simulation results revealed, global fusion steganalysis, which can comprehensive local decision-making classification performance, has achieved better detection performance than local decision-making.

II. FUSION DECISION MAKING RULE

Decision-making is the process of obtaining conclusion from evidences. There are at least two decision-making cases: (1) same evidences and different decision-making functions; (2) different evidences and same decision-making functions. The two cases might obtain same or different results. For a given decision-making task, information fusion technique can be used to organize and utilize various information resources effectively. The decision-making results using information fusion technique would be more accurate, more reliable and more stable than those decided by partial information resources. The technology utilizing information fusion technique to make various decisions into a final decision is fusion decision-making.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.38

A. Smallest Error Rate Bayesian Decision- Making Based on Evidence Set

Suppose the state space Ω consists of c kinds of pattern classes: $\Omega = \{\omega_1, \omega_2, \cdots, \omega_c\}$. Discrete random variable ω_j is modelled by probability $P(\omega_j)$ defined in random variable set Ω .Feature space X consists of d-dimension vectors: $X = [x_1, x_2, \cdots, x_d]^T$. Suppose that the prior probability $P(\omega_j)$ corresponding to class ω_j is known.

According to the Bayesian formula, the posterior probability of the state of vector X belonging to class ω_j, which is observed from feature space , is defined as:

$$P(\omega_j \mid X) = \frac{p(X \mid \omega_j)P(\omega_j)}{p(X)}. \tag{1}$$

Where $p(X) = \sum_{j=1}^{c} p(X \mid \omega_j)P(\omega_j)$ can be deduced from total probability formula.

Assume that each evidence $z_i (i = 1,2,\cdots,n)$ is consists of feature vector $X = [x_1, x_2, \cdots, x_d]^T$ with different dimensions. According to formula (1), the posterior probability distribution of occurrence of class ω_j under the evidence set $Z^n \equiv \{z_1, z_2 \cdots, z_n\}$ is $P(\omega_j \mid Z^n)$ as shown in formula (2) :

$$P(\omega_j \mid Z^n) = \frac{p(Z^n \mid \omega_j)P(\omega_j)}{p(Z^n)} = \frac{p(z_1, \cdots, z_n \mid \omega_j)P(\omega_j)}{p(z_1, \cdots, z_n)}. \tag{2}$$

Suppose evidences z_i in set $Z^n \equiv \{z_1, z_2 \cdots, z_n\}$ under known class ω_j are conditional independent, and then the joint probability of the evidence set Z^n which appearance under class ω_j can be described as:

$$\begin{aligned} & p(z_n \mid \omega_j) = p(z_1, \cdots, z_n \mid \omega_j) \\ & = p(z_1 \mid \omega_j) \cdots p(z_n \mid \omega_j) = \prod_{i=1}^{n} p(z_i \mid \omega_j). \end{aligned} \tag{3}$$

Then:

$$\begin{aligned} p(Z^n) &= \sum_{j=1}^{c} p(Z^n \mid \omega_j) = \sum_{j=1}^{c} (\prod_{i=1}^{n} p(z_i \mid \omega_j))F(\omega_j) \\ &= \sum_{j=1}^{c} F(\omega_j) \prod_{i=1}^{n} p(z_i \mid \omega_j). \end{aligned} \tag{4}$$

Further convert the posterior probability cdistribution formula (2) to formula (5):

$$\mathrm{p}(\omega_\mathrm{j} \mid \mathrm{z}^\mathrm{n}) = \frac{p(\omega_\mathrm{j}) \prod_{i=1}^{n} p(z_i \mid \omega_\mathrm{j})}{\sum_{j=1}^{c} p(\omega_\mathrm{j}) \prod_{i=1}^{n} p(z_i \mid \omega_\mathrm{j})}. \tag{5}$$

We can define the smallest error rate Bayesian decision-making rule base on the various evidences according to the Bayesian formula (5).

Definition 1: The smallest error rate Bayesian decision-making rule based on the evidence set Z^n is defined as: evidence set Z^n would be classified as class ω_k if the posterior probability of class ω_k under evidence set Z^n reaches maximum. That is to say:

$$k = \arg(\max_{j=1,2,\cdots,c} (P(\omega_j \mid Z^n)), q(\Omega) \in \omega_k. \tag{6}$$

B. The Fusion Rule of Combination System Likelihood Function

According to the smallest error rate Bayesian decision-making rule based on the evidence set Z^n, many kinds of fusion decision-making functions that based on the posterior probability can be derived. There are some typical fusion decision-making rules, e.g., product fusion rule, mean fusion rule and vote fusion rule et al. This paper emphatically proposes the fusion rule of combination system likelihood function.

Definition 2: Let likelihood matrix $P(\Omega \mid z_i) = (P(\omega_1 \mid z_i), P(\omega_2 \mid z_i), \cdots, P(\omega_c \mid z_i))$ denotes the posterior probability distribution of the decision-making results of evidence z_i, and $\sum_{j=1}^{c} p(\omega_j \mid z_i) = 1$.The combination system likelihood function about evidence set Z^n is defined as formula (7):

$$\begin{aligned} P(\Omega \mid Z^n) &= \lambda \cdot \overset{n}{\underset{i=1}{\otimes}} P(\Omega \mid z_i) \\ &= \lambda \cdot \overset{n}{\underset{i=1}{\otimes}} (P(\omega_1 \mid z_i), P(\omega_2 \mid z_i), \cdots, P(\omega_c \mid z_i)) \\ &= (\lambda \cdot \prod_{i=1}^{n} P(\omega_1 \mid z_i), \lambda \cdot \prod_{i=1}^{n} P(\omega_2 \mid z_i), \cdots, \lambda \cdot \prod_{i=1}^{n} P(\omega_c \mid z_i)). \end{aligned} \tag{7}$$

Where λ is a normalization constant, which ensures the sum of posterior probabilities is 1, so $\sum_{\mathrm{j}=1}^{\mathrm{c}} \lambda \cdot \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{p}(\omega_\mathrm{j} \mid \mathrm{z}_\mathrm{i}) = 1$.

Definition 3: Fusion rule of combination system likelihood function is that fusion decision-making result is the class whose combination system likelihood function is the maximum. It is described as formula (8):

$$\mathrm{k} = \arg_{j}(\max(\lambda \cdot \prod_{i=1}^{n} P(\omega_j \mid z_i)), q(\Omega) = \omega_k). \tag{8}$$

The fusion rule of combination system likelihood function fuses the results of the smallest error Bayesian decision-making of evidences in Z^n.

III. FUSION STEGANALYSIS

A. Fusion Steganalysis Based on the Combination System Likelihood Function

Steganalysis could be considered as two-class pattern recognition problem. There are two kinds of classes in the state space, cover image and stego image, which are described as $\omega_1 = \omega_{cover}, \omega_2 = \omega_{stego}$ respectively.A group of feature sets extracted from stego images and cover images are used for training classifier, then the trained classifier could be consider as a function *fclass* providing the decision value to classify a image to either cover or stego [3].

$$V_{Z^n} = f_{class}(Z_1^n) = f_{class}(z_1, z_2, \ldots z_n). \tag{9}$$

The obtained decision value is converted to a conditional probability distribution and normalized:

$$P(\omega_{stego} \mid Z_1^n) = f_{norm}(V_{Z^n}) = f_{norm}(f_{class}(Z_1^n)). \tag{10}$$

An input image I is decided to the class whose probability is the maximum value:

$$\begin{aligned} &P(\omega_{stego} \mid Z_1^n) > 0.5 \Rightarrow I \in \omega_{stego}, \\ &P(\omega_{stego} \mid Z_1^n) < 0.5 \Rightarrow I \in \omega_{cover} \end{aligned}. \tag{11}$$

Different fusion steganalysis rules can be derived from different fusion rules. According to definition 3, here we derive a fusion steganalysis rule of combination system likelihood function. Suppose a test image is detected by *n* evidences, and likelihood matrix of posterior probability distribution of each evidence is:

$$P(\Omega \mid z_i) = (P(\omega_{cover} \mid z_i), P(\omega_{stego} \mid z_i)). \tag{12}$$

where $i = 1, 2, \cdots, n$.

Then the combination system likelihood function for steganalysis is:

$$P(\Omega \mid Z_1^n) = \lambda \cdot \bigotimes_{i=1}^{n} (P(\omega_{cover} \mid z_i), P(\omega_{stego} \mid z_i)). \tag{13}$$

Definition 4: Fusion steganalysis rule of combination system likelihood function: The class which has maximum value of the posterior probability of combination system likelihood function for steganalysis should be considered as the final classification result.

$$\begin{aligned} &k = \arg_j \max_{j=cover,stego} (\lambda \cdot \prod_{i=1}^{n} P(\omega_{cover} \mid z_i), \lambda \cdot \prod_{i=1}^{n} P(\omega_{stego} \mid z_i)) \\ &q(\Omega) = \omega_k. \end{aligned} \tag{14}$$

B. Distributed Fusion Decision-Making Steganalysis

A distributed fusion decision-making steganalysis method is proposed as showed in Figure1. Distributed fusion decision-making contains two steps: local decision-making and global decision-making.

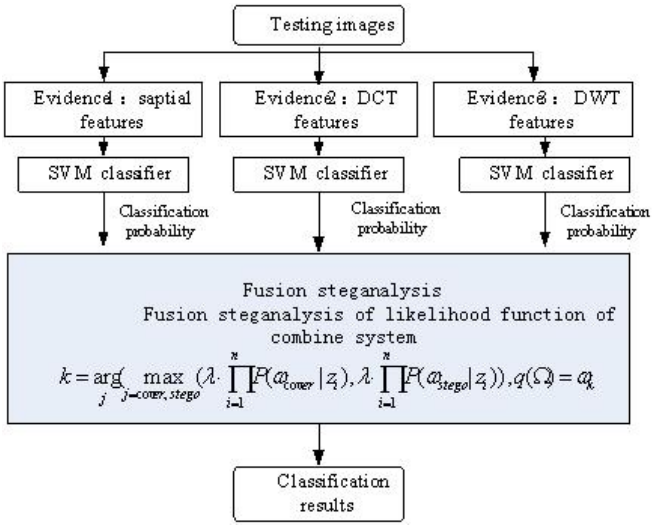


Figure 1. Distributed fusion decision-making steganalysis

1) Evidence Selection

The evidences for steganalysis are extracted from various data domains. Three feature generation methods are used in this paper.

In the spatial domain, the co-occurrence matrices of three directional differential images are used as features for steganalysis [10]. These features are sensitive to the steganographic algorithms based on spatial domain. We calculate the forward difference between neighbouring pixels in three directions: horizontal, vertical and diagonal. The value range of the differential image is limited to [0, 16]. Extract total $153 \times 3 = 459$ dimension feature vector from upper triangular co-occurrence matrixes of thresholded differential images.

In the DCT domain, TPM (the transition probability matrices) of intra- and inter-block difference arrays along zigzag direction are used as features for steganalysis [11]. Image is transformed into 8×8 DCT blocks, two dimensional intra- and inter-block difference arrays are calculated along zigzag direction, and all values of difference arrays are limited to [-5, 5].Extract total 132 dimension feature vector from upper triangular of intra- and inter- block difference arrays.

In the DWT domain, the first three statistical moments of test image and its corresponding predicted error image are selected as features for steganalysis [12]. A test image and its corresponding predicted error image are decomposed using three-level Haar wavelet, and their first scale diagonal subbands are further decomposed. The first three order statistical moments of each band are utilized to form a 102 dimension feature vector for steganalysis. By using K-L (Karhunen-Loeve) conversation, the principle components whose accumulative contribution rates are no less than 92% are used to reduce feature dimension.

2) Fusion Decision-making

Local decision-making is the first step.Each feature set reflect changes caused by information embedding from each viewpoint, therefore the detection performances of each

feature set to different steganographic methods are different from each other. To the same object, classification results in different domain are different, even sometimes conflict. Features extracted from spatial domain, DCT domain, DWT domain are input into SVM classier respectively to make the pre-classifying, and classier output three results of likelihood matrix of posterior probability distribution.

Global decision-making is the second step. In order to improve finale detection rate, based on the combination system likelihood function, the global decision-making result is obtained according to the fusion steganalysis rule of combination system likelihood function. Combination system likelihood functions are calculated by formula (13), and the final global decision-making classification result would be calculated by formula (14).

IV. IMULATION RESULTS AND ANALYSIS

A. Simulation Results

The images used for simulation are JPEG image testing data set [13], and the classifier is LIBSVM [14].

Simulation 1: detect 3 typical steganographic methods (F5, JPHide and JSteg). For each kind of steganographic methods, there are 4 data embedding sizes: tiny(less than 5%), small (between 5% and 15%), medium (between 15% and 40%), and large (more than 40%). 100 cover images and total 1200 stego images from each stego image set are selected, so 1300 images are selected in total. 60 pair images are used for training, and the remaining 40 pair images are used for testing.

Simulation 2: detect YASS. 500 clean JPEG images are selected from testing data. Parameters of YASS are set as follows: big block size chooses 10, 12 and 14; JPEG output quality factor QF_a and design quality factor QF_h select two schemes: 50/50 and 75/50, and the quantization index step length is 1. And so 3000 YASS stego-images are generated in total. 300 pairs are used for training and the remaining 200 pairs are used for testing.

For examine the performance of the proposed steganalysis method, we also do two other typical fusion methods which are mean fusion[3] and vote fusion.

Results are showed in TABLE 1 and TABLE 2, where FNR is false negative rate, FPR is false positive rate, A is accuracy, and L, M, S, T mean embedding sizes of large, medium, small and tiny respectively.

TABLE I. THE EXPERIMENTAL RESULTS OF DETECTING F5、JPHIDE AND JSTEG

Embed method	Embed ingrate	Spatial detection			DCT detection			DWT detection			Mean fusion [3]			Vote fusion			Proposed fusion		
		FNR	*FPR*	*A*	*FNR*	*FPR*	*A*	*FNR*	*FPR*	*A*	*FNR*	*FPR*	*A*	*FNR*	*FPR*	*A*	*FNR*	*FPR*	*A*
F5	L	47.5	5.0	51.3	0.0	7.5	96.3	4.5.0	22.5	66.3	2.5	7.5	95.0	72.5	2.5	62.5	0.0	7.5	96.3
	M	35.0	40.0	62.5	5.0	7.5	93.8	42.5	37.5	60.0	7.5	7.5	92.5	65.0	0.0	67.5	7.5	7.5	92.5
	S	57.5	45.0	48.8	2.5	5.0	96.3	40.0	42.5	58.8	5.0	7.5	93.8	72.5	2.5	62.5	5.0	7.5	93.8
	T	47.5	62.5	45.0	2.5	10.0	93.8	35.0	42.5	61.3	2.5	7.5	95.0	67.5	5.0	63.8	2.5	7.5	95.0
Average results		46.9	49.4	51.9	2.5	7.5	95.0	40.6	36.3	61.6	4.4	7.5	94.1	69.4	2.5	64.1	3.8	7.5	94.4
JPHide	L	7.5	15.0	88.8	7.5	20.0	86.3	0.0	10.0	95.0	2.5	10.0	93.8	12.5	7.5	90.0	0.0	10.0	95.0
	M	10.0	40.0	75.0	27.5	22.5	75.0	10.0	17.5	86.3	2.5	25.0	86.3	45.0	7.5	73.8	2.5	22.5	87.5
	S	40.0	22.5	68.8	20.0	15.0	82.5	12.5	17.5	85.0	15.0	20.0	82.5	52.5	2.5	72.5	15.0	22.5	81.3
	T	32.5	25.0	71.3	20.0	30.0	75.0	40.0	25.0	67.5	22.5	25.0	76.3	60.0	2.5	68.8	22.5	25.0	76.3
Average results		22.5	25.6	75.9	18.8	21.9	79.7	15.6	17.5	83.4	10.6	20.0	84.7	42.5	5.0	76.3	10.0	20.0	85.0
Jsteg	L	32.5	25.0	71.3	12.5	22.5	82.5	22.5	22.5	77.5	20.0	20.0	80.0	52.5	0.0	73.8	20.0	20.0	80.0
	M	25.0	25.0	75.0	27.5	25.0	73.8	20.0	17.5	81.3	20.0	17.5	81.3	40.0	5.0	77.5	20.0	17.5	81.3
	S	30.0	35.0	67.5	20.0	27.5	76.3	32.5	22.5	72.5	22.5	30.0	73.8	50.0	12.5	68.8	22.5	30.0	73.8
	T	25.0	25.0	75.0	22.5	25.0	76.3	17.5	37.5	72.5	15.0	32.5	76.3	50.0	12.5	68.8	15.0	32.5	76.3
Average results		28.1	27.5	72.2	20.6	25.0	77.2	23.1	25.0	75.9	19.4	25.0	77.8	48.1	7.5	72.2	19.4	25.0	77.8

TABLE II. THE EXPERIMENTAL RESULTS OF DETECTING YASS

Big Block	QF_h/QF_a	Spatial detection			DCT detection			DWT detection			Proposed Fusion		
		FNR	*FPR*	*A*	*FNR*	*FPR*	*A*	*FNR*	*FPR*	*A*	*FNR*	*FPR*	*A*
10	50/50	8.50	5.00	93.25	26.50	27.50	73.00	25.50	18.50	78.00	5.00	4.50	95.25
	75/50	9.00	5.00	93.00	66.50	50.50	41.50	26.50	23.50	75.00	9.00	3.50	93.75
12	50/50	8.50	5.00	93.25	65.00	58.00	38.50	26.50	22.50	75.50	8.00	2.00	95.00
	75/50	8.00	5.00	93.50	68.00	54.00	39.00	25.00	22.50	76.25	9.50	4.00	93.25
14	50/50	8.50	5.00	93.25	66.50	51.00	41.25	23.50	20.50	78.00	7.50	2.00	95.25
	75/50	9.50	5.00	92.75	66.50	56.50	38.50	22.50	23.50	77.00	8.50	3.50	94.00
Average detection results		8.67	5.00	93.17	59.83	49.58	45.29	24.92	21.83	76.63	7.92	3.25	94.42

B. *Analysis of Fusion Results*

The proposed method performs better than vote fusion detection and mean fusion detection. From table I , it can be see the effect of vote fusion detection is the poorest one. Because in vote fusion, the global fusion detection result is simply decided by the majority without considering the support from different evidence, and it has high missing alarm rate and low false alarm rate. In mean fusion detection, according to the mean of local decision making probabilities, the class whose mean value reach the maximum value is considered as the final decision, and it performs better than vote fusion detection. Fusion steganalysis of combination system likelihood function considers overall of the posterior probabilities of patter classes of all evidences. This fusion method combines all of the posterior probabilities of evidences to strengthen support to results, it is better than mean fusion[3].

The best fusion performance appears when a local decision is better than other local decisions, which can be proved in Table I and Table II. For attacking YASS, the detection performance of spatial detection is better than that of the other two domain detections significantly, so that the fusion steganalysis detecting YASS performes well. For attacking F5, DCT detection obtains better performance than that of the other two domain detections, and result in fusion decision-making obtained the best results. For attacking JPHide, the DWT detection performs well but does not show any decided advantage over the other methods, though the best fusion performance is not achieved, but the whole detection rate is improved a bit. For attacking Jsteg, there's no best method in all of the three domains, and the improvement of fusion steganalysis is not obvious.

CONCLUSION

Comprehensive contributions of each domain features to reveal the steganography, this paper proposes a distributed decision-making fusion steganalysis method. Universal steganalysis is to solute steganalysis problem with no stegangraphic method information, so we construct feature sets from three domains. Simulation results show that global decision-making performances are superior to local decision-making performances, especially the proposed fusion steganalysis of combination system likelihood function comprehensiving local decision-making classification performance, can reduce missing alarm rate and false alarm rate obviously, and obtain better global detection performance.

ACKNOWLEDGMENT

This paper is supported by "the Fundamental Research Funds for the Central Universities". (Project Number: JUSRP21131）.

REFERENCES

[1] K. Solanki, A. Sarkar, and B. S. Manjunath, YASS: Yet another steganographic scheme that resists blind steganalysis, LNCS 4567.Berlin Heidelberg: Springer, 2007 , pp . 16–31,doi: 10.1007/978-3-540-77370-2_2.

[2] A. Sarkar, K. Solanki, and B. S. Manjunath, Further study on YASS: Steganography based on randomized embedding to resist blind steganalysis, Security, Forensics, Steganography, and Watermarking of Multimedia Contents X, San Jose CA: SPIE, 2008, pp. 681917-681917-11,doi:10.1.1.139.6691.

[3] M. Kharrazi, T. H. Sencar, N. Memon, Improving Steganalysis by Fusion Techniques: A Case Study with Image Steganography, In: Proceedings of SPIE, Volume 6072 -- Security, Steganography, and Watermarking of Multimedia Contents VIII, San Jose, CA,USA: SPIE, 2006, pp. 123-137,doi:10.1117/12.645510.

[4] B. M. Rodriguez, G. L. Peterson, S. S. Agaian, Multi-class classification averaging fusion for detecting steganography. Proceedings of IEEE International Conference on System of Systems Engineering, San Antonio, Publisher IEEE Computer Society Washington, DC, USA. 2007, pp. 1-5, doi: 10.1109/SYSOSE.2007.4304292.

[5] B. Rodriguez, G. Peterson, K. Bauer, Fusion of Steganalysis Systems Using bayesian model averaging, IFIP International Federation for Information Processing, Boston: Springer, 2008, pp. 345–355,doi:10.1007/978-0-387-84927-0_27.

[6] Dong. Jing, Chuan Chen. Xiao, Guo. Lei, and Niu Tan. Tie, Fusion Based Blind Image Steganalysis by Boosting Feature Selection, LNCS 5041. Berlin Heidelberg: Springer,2008, pp. 87–98,doi:10.1007/978-3-540-92238-4_8.

[7] Z W. SUN, Z C. JI, Image steganalysis based on Markov model and feature fusion, Control and Decision, 2009,24(8), pp. 1039-1242,doi: CNKI:SUN:KZYC.0.2009-08-022.

[8] K. Christian and D. Jana, The impact of information fusion in steganalysis on the example of audio steganaysis, Proceeding of SPIE-IS & T electronic imaging, 2009,SPIC Vol.7254, pp. 725409-1-725409-12.

[9] K. Christian and D. Jana, Improvement of Information Fusion Based Audio Steganalysis, Proc. of SPIE-IS&T Electronic Imaging,2010, SPIE-IS&T Vol. 7542, pp. 75420A-1-75420A-11.

[10] Z W. SUN, M M. HUI,C. GUAN, Steganalysis based on co-occu rrence matrix of differential image, IIH-MSP. Washington DC USA: IEEE, 2008, pp. 1097-1100,doi: 10.1109/IIH-MSP.2008.176.

[11] Z W. SUN, Z C. JI, Image stganalysis based on block correlation and Markov model in DCT domain, Information and Control, 2009,38(5), pp. 60-607.

[12] Z W. SUN, Z P. ZHOU,H. LI, A Image Steganalysis Method Based on Characteristic Function Moments of Wavelet Subbands and PCA, Information Network Security, 2009(7), pp. 41-43, 46,doi: 10.3969/j.issn.1671-1122.2009.07.013.

[13] A. Rocha, S. Goldenstein, W. Scheirer, T. Boult, The Unseen Challenge Data Sets, IEEE Computer Society Conference on Computer Vision and Recognition Workshop. Anchorage, Alaska: IEEEComputerSociety,2008,pp.1-8,doi: 10.1109/CVPRW.2008.4562987.

[14] C.C. Chang , C.J. Lin, SVM: a Library for Support Vector Machines.[Online],available:http://www.csie.ntu.edu.tw/~cjlin/libsvm

THE RESEARCH ON THE ALGORITHM OF LATTICE LCD DRAWING CONTROLLED BY PIC

Hua Fang[1, 2]
1.Department of Management & Information Shandong Transport Vocational College
Weifang ,Shandong province 261206,China
2.Department of Information Shandong Science Technology Vocational College
Weifang ,Shandong province 261053,China
Email: sdfanghua@sina.com

Hua Jiang
Department of Information Shandong Science Technology Vocational College
Weifang ,Shandong province 261053,China
Email: wf-jianghua@163.com

***Abstract*—With the instruments become more and more intelligent, multifunctional, the pointer display way can not meet this kind of development. LCD, especially the lattice LCD with great display screen makes it possible to produce instruments with much more beautiful interface and much more information on the display screen. In this paper some algorithms on drawing dot, line, circle, and the averagely partition of circle are introduced, which fit the applications of MCU. The actual application proves that these algorithms work well in the MCU applications. They can draw all the pictures that are needed for the instruments.**

Keywords- instruments; lattice LCD; MCU; drawing

In a variety of real-time automatic control and intelligent instruments, the human-computer interaction is an indispensable part. Generally, the human-computer interaction consists of two parts. The first part is to accept external control commands. The other part is to provide a variety of real-time information of equipment. In previous devices, the information was provided mainly by mechanical pointer, indicator, etc. With the instruments become more and more intelligent, multifunctional, more and more information need to be provided. These traditional methods can not meet these requirements.[1] LCD solves this problem. It can display rich content. The modern lattice LCD with great display screen can not only display a variety of characters, but also draw a variety of graphics. Displayed contents are greatly enriched. So this type of LCD has been widely used in the instruments.

At the same time the modern MCU technology has developed to a very high level. It can carry on a lot of computing operations, while it integrates a lot of hardware features. Using MCU to control the LCD display has been widely used in the intelligent instruments.[1] In the following it will introduce some algorithm theories to draw on LCD under the control of MCU.

I. THE HARDWARE INTERFACE OF SYSTEM

LM64P83L produced by the SHARP is a large-screen LCD display, which lattice is 640*480. With higher display resolution and very small pixels on its screen is, only 0.27 * 0.27, it is suitable for displaying large amounts of information. Therefore, this study uses this kind of LCD display. The QPYD-01 controller which is produced by Varitronix Company of Hong Kong is used, and the interface with the controller is PIC MCU in this system.

The LCD display system in this paper is a sub-module of a new vehicle digital instrument system. The system includes signal detection module, signal display module and other parts. The signal detection module will not be introduced in this paper. The application interface of LCD display and MCU is shown in Figure 1.

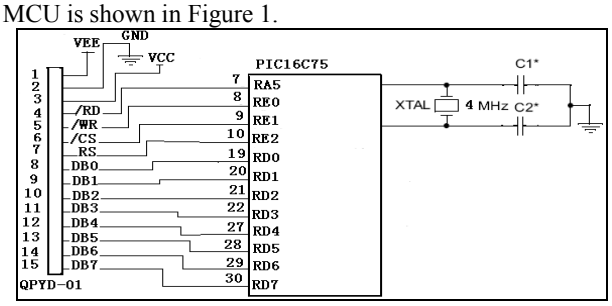


Figure 1. Hardware Interface Diagrams of PIC16C75 and QPYD-01

II. DETERMINE THE SCREEN COORDINATES

Drawing in the lattice LCD is achieved by lit the corresponding pixel in the LCD screen. The pixel on the LCD screen corresponds with the storage unit of the LCD controller. One byte in the controller's storage unit corresponds with eight neighboring pixels of the LCD screen. The controller controls the LCD display in the unit of byte. To control the LCD to display, the address of the controller's storage unit must be determined, which is corresponded with the LCD screen pixels.[2] [3] Therefore, in order to use the LCD screen, the corresponding relationship between all the pixels of the LCD screen and the storage location of the controller must be clear.

The storage unit address of the control module storage unit, which is corresponded with the pixels of the lattice LCD screen, is not arranged according to the axis way. Therefore, the Cartesian coordinate system on the screen must be established firstly. The corresponding relationship between the pixels on the LM64P83L LCD screen and the

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.99

storage unit is column-major. Shown in figure 2, the LM64P83L is controlled by the QPYD-01 controller, and the vertex of the upper left corner of the screen is the origin of its Cartesian coordinates system, and then follow from left to right, each of eight points are one byte. There are 80 bytes per row, and there are 480 rows in total. Therefore, there are 480*80=38400 bytes in the storage unit, which is corresponded with the LCD screen. The arrangement of each byte is started from high, that is, they are 7,6,5,4,3,2,1,0 from left to right.[2] This is shown in figure 3.

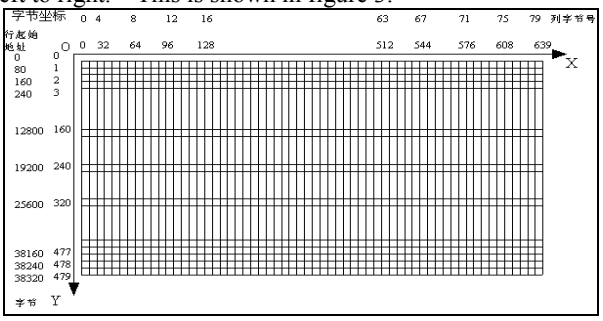


Figure 2. Determine of Screen Position and Axis

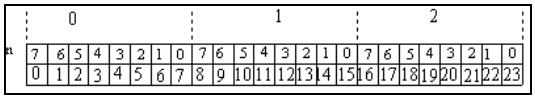


Figure 3. the Relationship between the Storage Unit and the Bit of Storage Byte

In addition to specify the byte address which is corresponded with the pixel in the controller storage unit, it is necessary to assigned value to the address storage unit.

Take Pixel A(x,y) on the screen as an example. First, its corresponding storage unit address is determined according to its Cartesian coordinates. The formula is $ADD = y * 80 + [x/8]$. The character y represents the row where the pixel is, and the character x represents the column where the pixel is. [x/8] is the integer part of the results of x/8 which represents the byte position in this line where the pixel is. And the remainder represents the bit of this byte. Then with this remainder, the value to be assigned to the corresponding address byte for lighting the pixels can be obtained. The specific method is that, first assigned to a register 80H, and then loop move the value to the right, loop times are the remainder of the division (if the remainder is zero, then no loop). For example, the pixel (201, 50), its corresponding storage unit is 50*80+[201 ÷ 8]=4025. The remainder of 201/8 is 1, so move one bit to the right, and get the result: 40H. Therefore, for lighting the pixel, the 4025 storage unit should be assigned 40H. Of course, the bytes arrangement of different LCD may be different, because the controllers of these LCD are different. For example, the LCD has built-in HD6202U graphic LCD module, which is produced by Varitronix Company. The bytes arrangement of the LCD is row-major. But their basic idea is same. That is 8 pixels on the screen correspond with a byte of the controller storage unit.

The dot in the plane graph is a basic element. The way of lighting up a certain pixel of the LCD screen has already been described in the previous section. A complete graph is composed by a number of dots which satisfy certain rules. These rules are called algorithm. Through the algorithm, the coordinates of each dot is obtained. Then, by the drawing methods described above, each corresponding dot is drawn. This completes the graphics drawing. In the following some drawing algorithms for commonly used graphic elements are introduced.

III. LINE DRAWING ALGORITHM

In the plane graph, in addition to the dot, the line is another basic element. Here a line drawing algorithm suitable for MCU is introduced.

There are many algorithms of slope k=(Ye-Ys)/(Xe-Xs) of the straight line whose starting dot is (Xs, Ys) and end dot is (Xe, Ye). In order to avoid the complex floating-point operations, the integral digital differential analysis (Integer DDA) is commonly used in MCU system. The basic idea of the Integer DDA is to avoid the division in the process of calculating the slope K. By ensuring$\Delta Y \geq 0$, and applying the relationship between $\Delta Y=Ye - Ys$ and $\Delta X=Xe–Xs$, divide the slope K into four areas. In each area, using ERR parameters it specifies the error between the location of the established dot and the true line dot. Then with the error the corresponding rule of the drawing dots is determined. The situation of the end dot is also considered in this rule. The four areas are described in the following:

1) K=0 ~ 1

It's $\Delta Y \geqslant 0$, $\Delta X \geqslant 0$ and $\Delta X \geqslant \Delta Y$. The rule of the drawing dots is as the following: at the current dot (X, Y): if it's ERR <0, then draw the dot at the coordinates (X +1, Y), and ERR = ERR + ΔY;

If it's ERR$\geqslant$0, then draw the dot at the coordinates (X +1, Y+1), and the ERR = ERR + ΔY- Δ X;

Then the new dot coordinate is taken as the current dot (X, Y), and the resulting ERR is the deviation of the dot coordinates. Repeat above steps until X = Xe. The initial coordinates (X, Y) is (X, Y), and ERR=0.

2) K > 1

It's $\Delta Y \geqslant 0$, $\Delta X \geqslant 0$ and $\Delta Y \geqslant \Delta X$. The rule of drawing dots is as the following: at the current dot (X, Y): if it's ERR <0, then draw the dot at the coordinates (X +1, Y+1), and the ERR = ERR + ΔY– Δ X;

If it's ERR$\geqslant$0, then draw the dot at the coordinates (X, Y+1), and the ERR = ERR - Δ X.

Then the new dot coordinate is taken as the current dot (X, Y), and the resulting ERR is the deviation of the dot coordinates. Repeat above steps until Y = Ye. The initial coordinates (X, Y) is (X, Y), and ERR=0.

3) K=-1 ~ 0

It's $\Delta Y \geqslant 0$, $\Delta X<0$ and $|\Delta X| \geqslant \Delta Y$. The rule of drawing dots is as the following: at the current dot (X, Y): if ERR <0, then draw the dot at the coordinates (X-1, Y), and the ERR = ERR + ΔY;

If it's ERR≥0, then draw the dot at the coordinates (X-1, Y+1), and the ERR = ERR + Δ Y - Δ X.

Then the new dot coordinate is taken as the current dot (X, Y), and the resulting ERR is the deviation of the dot coordinates. Repeat the above steps until X = Xe. The initial coordinates (X, Y) is (X, Y), and ERR=0.

4) K < 1

It's Δ Y ≥ 0, Δ X<0 and Δ Y ≥ | Δ X|. The rule of drawing dots is as the following: at the current dot (X, Y): if ERR <0, then draw the dot at the coordinates (X-1, Y+1), and the ERR = ERR + ΔY– Δ X;

If it's ERR≥0, then draw the dot at the coordinates (X, Y+1), and the ERR = ERR - Δ X.

Then the new dot coordinate is taken as the current dot (X, Y), and the resulting ERR is the deviation of the dot coordinates. Repeat the above steps until Y = Ye. The initial coordinates (X, Y) is (X, Y), and ERR=0.[2]

Actually, in this algorithm line drawing is divided into four cases according its slope. The program developped according to this algorithm can complete drawing line (the line also includes the slash) well. In figure 5, it can be seen, except for the horizontal line or vertical line, the other lines are actually composed by many short horizontal lines or short vertical lines. This is because of the discontinuity of the dots on the LCD screen and the principle of this algorithm. This algorithm is suitable for MCU application.

IV. CIRCLE DRAWING ALGORITHM

The definition of a circle in the plane geometry is the set of the dots which are equidistant from a fixed dot in the plane. Therefore, a lot of drawing algorithms about the circle can be developped based on this definition. Because othe the computing power limit of MCU is limited, it will take long time to perform calculation, especially the calculation of floating-point. In real-time system, this is not allowed. Therefore, for MCU application the calculation operation of the algorithm for drawing circles or arcs should be less. And floating-point operations are not allowed.

Based on above analysis, this paper presents an algorithm for drawing circles, which is called Senham's Circle Algorithm. The idea of this algorithm is that the circle, whose center is the coordinate origin, can be drawn by calculating the arc between 0^o and 45^o. That is, the radius is increased from 0 to $r/\sqrt{2}$, the solid line shown in Figure 4. And then calculate the remained 7 arcs by using the symmetry. By comparing the distance between the adjacent dots to the dot of the arc of the circle, the dot with the smaller error is selected as the drawing dot. Take the solid line in Figure 4 arc as an example. According to the definition of the circle, the dot (X1, Y1) should be a dot on the arc. But in the algorithm, only the integer coordinates dots can be used. So according to the algorithm, the dot A or the dot B should be taken. In the figure, dot A is closer to the arc. Therefore, the dot A is taken as the drawing dot to indicated the dot (X1,Y1). For other dots on the arc, the drawing way is same.

In this algorithm, $r/\sqrt{2}$ needs to be calculated through floating-point calculations. Therefore, MCU programming becomes complicated. To avoid this calculation, the dotted line of the arc in the Figure 4 is considered. If dot (r, 0) is taken as the staring dot, then the value of X is decreased from r to $r/\sqrt{2}$ while the value of Y is increased from 0 to $r/\sqrt{2}$. Thus a 45 º arc is obtained.[3]

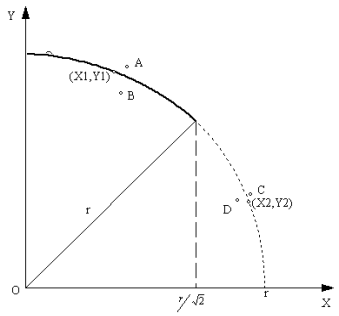


Figure 4. Using Senham's Circle Algorithm to Draw a Circle

The core of this algorithm is the symmetry of circles. The work of drawing circle is finished through drawing two one-eighth arc. With this algorithm, the calculations of MCU are greatly reduced. Therefore it is very suitable for the MCU applications.

V. DIAL DRAWING ALGORITHM

By applying this algorithm, the interface of a new vehicle digital cluster instrument is drawn out. It is shown in Figure 5. The LCD screen is the LM64P83L which is described above. The controller of the LCD is the QPYD-01. In this new vehicle digital cluster instrument, besides various operating parameters of the vehicles, it also displayes graph, dials, and pointers similar as traditional cluster instrument. During vehicle running, parameters digit keep changing. It is very inconvenience for the drivers to view. Therefore, to facilitate reading the driving parameters, the system not only displays the digit on the LCD screen, but also displays graphes. Referring to traditional cluster instrument, the speedometer, engine tachometer dial, pointer, are kept. And the fuel quantity, water temperature and others are shown through the column graphics. All of these graphics are drawn out through the LCD screen, as shown in Figure 5.

The problem of the division of the scale arc is involved when drawing the speedometer and the engine revolution meter. This problem will be encountered in the instruments which need the LCD to display. After searching a lot of files, it's failed to find the algorithm which is related to the circular scale division. In the following a self developped algorithm is described.

According to the principle of circle, it can be seen that in the same circle, the length of the arc, which is corresponded with the equal central angle or the equal circumferential angle, is equal. Based on this principle, we can get a lot of methods of the division arc, such as the method according to

the arc central angle, and the method according to the actual length of the arc. The algorithm used to division arc in this paper is based on the actual length of the arc. According to the Senham's Circle Algorithm it can be seen that to draw a circle is actually done by drawing dots. Each dot can be taken as an equal small arc of the large arc. Thus it is the basis of dividing the arc. Firstly, the total number of the dots which are used to draw the 1/8 arc is got through the experiments. Then this total number divided by the number of the arc segments which are the 1/8 arc is divided into. And the result is the number of the dots which are needed to draw one arc. The arc can be divided evenly. A register should be set up to count the number of the drawing dots, while the system is programming. When the number of the drawing dots is equal to the number of the dots which each arc segment needs, the system can begin to draw the scale, according to the coordinate of the current dot, the coordinate of the center of the circle and the line drawing algorithm described above.

Figure 5. the Interface of the new vehicle digital instrument interface

Take the speedometer in Figure 5 as an example, the arc should be drawn firstly, and the scale line should be drawn during the process of drawing the arc. According to the experiments, to draw the 1/8 arc of the speedometer in this paper, 150 dots are needed. According to the design of the speedometer size, this 1/8 arc need to be divided into 6 segments. Therefore, there are 25 dots in the each arc. When the number of the actual drawing dots is equal to the number of scale dots, it should start to draw the scale. The graphic which is actually drawn out is shown in Figure 5. In this figure, the length of each arc is equal, and the division of the arc is very average. This shows that the algorithm is correct.

VI. CONCLUSION

The LCD has many advantages which the other monitors do not have, such as low power, small size, light weight, ultra-thin displays, etc. The lattice LCD can display not only the characters and the numbers, but also the graphics, the curves and the Chinese characters. These advantages make the man-machine interface more friendly and the operation more flexible and convenient. Therefore the LCD is widely used in a variety of instruments, meters, electronic equipments and other low-power products. The key of the LCD display controlled by the MCU is to find the right algorithm.[1] And then the practical results can be achieved. The result proves that the algorithm described above is very suitable for the MCU applications, and the effect is very good.

REFERENCES

[1] Hu Aihua, Yang Yuchi, Liu Yuanying, Wang Yu. LCD Module and Its Application in Intelligent Instrument. Computer Measurement & Control. 2007-02：34-37

[2] Chen Jianqun, Duan Chendong, Dong Penghui, Lu Wei. Display and Drawing Methods for LCD Module. Computer Measurement & Control. 2009-04：41-43

[3] Huang ziqiang. Princeple of Liquid Crestal Display. Beijing , National Defence Industry Press

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Image Retrieval Based on PCA-LPP

Zhenhua Zhang, Xinzhong Zhu*
College of Mathematics, Physics and Information Engineering, Zhejiang Normal University
Jinhua, China
zzhqq0000@163.com, zxz@zjnu.cn

Jianmin Zhao, Huiying Xu
College of Mathematics, Physics and Information Engineering, Zhejiang Normal University
Jinhua, China
zjm@zjnu.cn, xhy@zjnu.cn

***Abstract*—In many areas of image retrieval, pattern recognition and data mining, one is ofen confronted with some form of dimensionality reduction. In this paper, we introduce a algorithm of Principal Components Analysis-Locality Preserving Projections (PCA-LPP). This algorithm can compute the principal components of a sequence of data set and at the time optimally preserves the neighborhood structure of the data set. This procedure is done by merging the runs of two algorithms based on principal component analysis (PCA) and Locality Preserving projections (LPP) running sequentially. This algorithm is applied to image retrieval problem. Simulation results on Corel image database showed better accuracy rate of this algorithm compared to PCA and LPP algorithms.**

Keywords-image retrieval; dimension reduction; Principal Components Analysis; Locality Preserving Projections (key words)

I. INTRODUCTION

Suppose we have a large number of data points of $n-$ dimensional real vectors ,for example, gray scale $n\times n$ images yield data points in R^{n^2}. One is confronted with the situation where n is *very large*. This leads one to consider methods of dimensionality reduction that allow one to represent the data in a lower dimensional space. Popular dimensionality reduction techniques like PCA[1-3], LLE[4], Laplacian Eigenmaps [5], ISOMAP[6], LPP[7], however, LLE, Laplacian Eigenmaps, ISOMAP are defined only on the training data points and they are unclear how to evaluate the map for new test points. In contrast, PCA and LPP are defined everywhere. In this paper, We analysis the advantage of PCA and LPP techniques.

The first technique is called Principal Components Analysis[1-3](PCA)which is a popular principal component technique. The algorithm computes eigenvectors and eigenvalues for a sample covariance matrix derived from a well known given image data matrix, by solving an eigenvalue system problem. It has the following advantages: 1). Preserving Principal Components of image data set and denoising. This make it removing noise from a image signal and suitable for large data set. 2). Linear dimensionality reduction technique. This make it can find a transformation matrix A that maps these m points from R^n to R^l ($l \ll n$). However, It can not preserve the neighborhood structure of the data set.

The second technique is called Locality Preserving Projections[7](LPP). It builds a graph incorporating neighborhood information of the data set. Using the notion of the Laplacian of the graph, we then compute a transformation matrix which maps the data points to a subspace. This linear transformation optimally preserves local neighborhood information in a certain sense. The algorithm has some advantages: 1). LPP is defined everywhere. The LPP may be simply applied to any new data point to locate it in the reduced representation space. 2). The locality preserving quality of LPP is likely to be of particular use in image retrieval applications. Since LPP is designed for preserving local structure, it is likely that a nearest neighbor search in the low dimensional space will yield similar results to that in the high dimensional space. This makes for an indexing scheme that would allow quick retrieval. However, It can not neglect noise from data set, because LPP basis on preserving the local geometry structure, it can not judge a whole the overall structure of manifolds.

In this paper, after analyze the advantages of PCA and LPP algorithms, we propose a new linear dimensionality reduction algorithm, called Principal Components Analysis-Locality Preserving Projections (PCA-LPP). It merges sequentially two techniques based on PCA and LPP.

The algorithm is interesting from a number of perspectives.

1） It inherits the advantages of denoising and preserving Principal Components of image data set.

2） It can preserve the neighborhood structure of the data set.

3） PCA-LPP is linear. It is clear how to map a new data point.

II. PCA-LPP ALGORITHM

Given that large number of $n-$ dimensional vectors, $x(1);x(2);...,$ which are the observations from a certain given image data, are received sequentially. Without loss of generality, a mean image m is calculated in the beginning of the algorithm. It should be noted that the sample mean image is representation of all the images . The mean image can be subtracted from each vector $x(n)$ in order to obtain a normalization vector of approximately zero mean. Let $E((x-m)(x-m)^t)$ be a $n\times n$ scatter matrix.

* Corresponding Author. E-mail address: zxz@zjnu.cn

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.52

The feature vectors which are obtained from S matrix will form a basis which describes the original data set without loss of information. We then take the number of the data points which reduced dimensionality as inputs of LPP algorithm. LPP constructs a neighborhood graph and chooses suitable weights. Finally, we obtained a transformation matrix which maps the data points to another subspace again. It works like a linear system that processing sequentially the PCA and LLP algorithms.

The image retrieval can be done by projecting the input test image onto this subspace and comparing the resulting coordinates with those of the training images in order to find the nearest appropriate image. The algorithm:

1. Subtracting the sample mean from the each data vector : $x_i - m \quad i = 1,2,....,n$, m represent image mean.
2. Computing the scatter matrix S : $S = \sum_{k=1}^{n} (x_k - m)(x_k - m)^t$, The scatter matrix should look familiar—it is merely $n-1$ times the covariance matrix and find the best direction.
3. Computing eigenvectors $\{a_1, a_2, \ldots a_k\}$ corresponding to the largest k eigenvalues of the scatter matrix S. Compute equations : $Sa = \lambda a$
4. Let $\{a_1, a_2, \ldots a_k\}$ be the columns of eigenvector matrix $A = \{a_1, a_2, \ldots a_k\}$.
5. Obtaining LPP data from the project: $Y = Ax_i \quad i = 1,2,\ldots n;$ such that y_i represents x_i
6. Seeking k nearest neighbors and Constructing the adjacency graph G : We put an edge between nodes i and j if y_i is among k nearest neighbors of y_j or y_j is among k nearest neighbors of y_i .
7. Computing the weights and Constructing the weight matrix W : $W_{ij} = 1$ if and only if vertices i and j are connected by an edge. $W_{ij} = 0$ if there is no such edge.
8. Obtaining Eigenmaps : Compute the eigenvectors and eigenvalues(1):

$$YLY^T b = \lambda YDY^T b \tag{1}$$

Where D is a diagonal matrix whose entries are column or row sums W , $D_{ii} = \sum_j W_{ji}$. $L = D - W$ is the Laplacian matrix. Y is a result from reduction dimensionlity by PCA. Let the column vectors $b_0, b_1 \ldots b_{l-1}$ be the solutions of equation, sorted according to their eigenvalues, $\lambda_0 < \lambda_1 < \ldots \lambda_{l-1}$. We then sorted eigenvectors and constructed the eigenvectors matrix, where $B = \{b_0, b_1 \ldots b_{l-1}\}$. Thus, the embedding is as follows: $y_i \to z_i = B^T y_i$, where $y_i = A x_i \quad i = 1, 2, \ldots n;$ from PCA. A is a eigenvector matrix from PCA. X is a matrix from image data, where $x_1, x_2, \ldots x_n \in X$.

III. EQUATIONS AND JUSTIFICATIUON

Fist, Recall Step 1-4 that given a image data set we construct a scatter matrix S to preserve principal components(2). Then, we compute eigenvector of the scatter matrix to finding the best direction (3).

$$S = \sum_{k=1}^{n} (x_k - m)(x_k - m)^t \tag{2}$$

$$Sa = \lambda a \tag{3}$$

We should select the eigenvectors corresponding to the largest eigenvalues of the scatter matrix. In other words, the project is best when the vectors $\{a_1, a_2, \ldots a_k\}$ are the k eigenvectors corresponding to the largest eigenvalues of S . Because S is real and symmetric, these eigenvectors are orthogonal. They form a natural set of basis vectors for representing any feature image vector X , where X is a matrix from image database. a_i are the components of X in that basis, and called principal components. Principal component analysis reduces the dimensionality of feature space by restricting attention to those directions along which the scatter matrix is greatest.

Second, Recall 5-8 that given a data set from $Y = Ax_i$ we construct a weighted graph $G = (V, E)$ with edges connecting $k-$neighborhood points to each other. And then map a weighted graph G with to a line so that connected points stay as close together as possible. Let $z = (z_1, z_2 \ldots z_m)^T$ be such a map. Choosing $z_i \in R$ to minimize the following objective function(4).

$$\sum_{ij} (z_i - z_j)^2 W_{ij} \tag{4}$$

The objective function with our choice of W_{ij} incurs a heavy penalty if neighboring points y_i and y_j are mapped far apart. Therefore, minimizing it is an attempt to ensure that if y_i and y_j are "close" then z_i and z_j are close as well.

Suppose a is a transformation vector, that is $y = a^T X$, where the i^{th} column vector of X is x_i . b is a transformation vector, that is $z = b^T Y$, where the i^{th} column vector of Y is y_i . By simple algebra formulation, the objective function can be reduced to

$$\frac{1}{2}\sum_{ij}(z_i - z_j)^2 W_{ij} = \frac{1}{2}\sum_{ij}(b^T y_i - b^T y_j)^2 W_{ij}$$
$$= \frac{1}{2}\sum_{ij}(b^T a^T x_i - b^T a^T x_j)^2 W_{ij}$$
$$= \sum_i b^T a^T x_i D_{ii} x_i ab - \sum_{ij} b^T a^T x_i W_{ij} x_j ab$$
$$= b^T a^T X(D-W)X^T ab$$
$$= b^T a^T XLX^T ab = b^T YLY^T b$$

where D is a diagonal matrix; $D_{ii} = \sum_j W_{ij}$. $L = D - W$ is the Laplacian matrix. Matrix D provides a natural measure on data points. We impose a constraint as follows:

$$b^T YDY^T b = 1$$

The transformation vector b that minimizes the objective function is given by the minimum eigenvalue solution to the generalized eigenvalue problem:

$$YLY^T b = \lambda YDY^T b \qquad (1)$$

IV. EXPERIMENT

In this section, we apply PCA-LPP algorithm to image retrieval and compare with PCA and LPP algorithms, by experiments on the data sets from COREL image database. The image dataset we use consists of 1,000 images of 10 semantic. Each semantic category contains 100 images. The 1,000 images are divided into two subsets. The first subset consists of 900 images, and each semantic category contains 90 images, we called this subset "training set". The second subset consists of 100 images, and each semantic category contains 10 images, we called this subset "test set". To reduce the computational, each image in COREL database is scaled into (96×64), original image size is 384×256 .Figure 1 show a sample of flower images from COREL database.

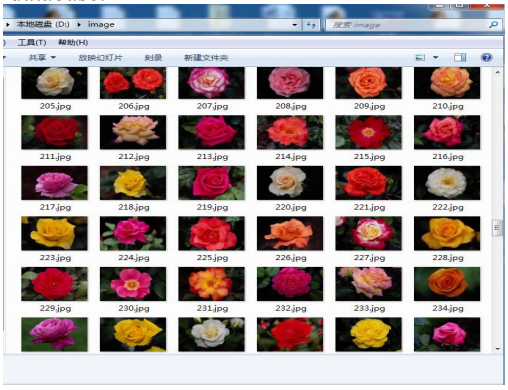


Figure 1. some sample images of flower in the COREL database

A retrieved image is considered correct if it belongs to the same category of the query image. To facilitatly detection efficiency of retrieval, The retrieval accuracy is defined as follows (5):

$$Accuracy = \frac{relevant\ \ K\ images\ in\ retrieved\ \ results}{N} \qquad (5)$$

Figure 2 show the results of PCA-LPP. We applied the PCA-LPP to images of Flower semantic. The image results of retrieval show that fist row note "Query image" and other note image results which applied a algorithm. We suppose $N = 36$, and reduction space is 30-dimensional.Figure 3 show the results of PCA. Figure 4 show the results of LPP.

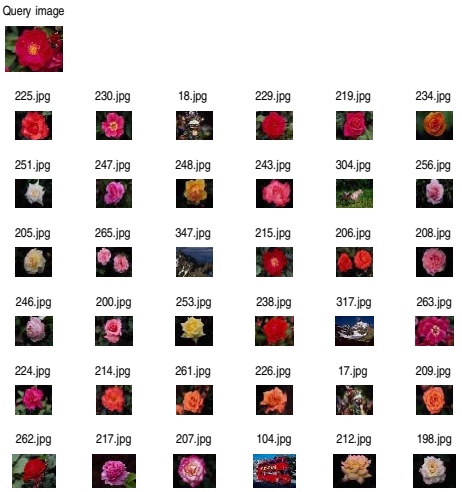


Figure 2 . the results of PCA-LPP

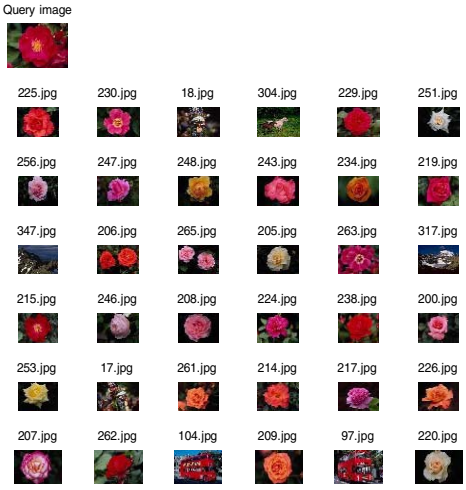


Figure 3. the results of PCA

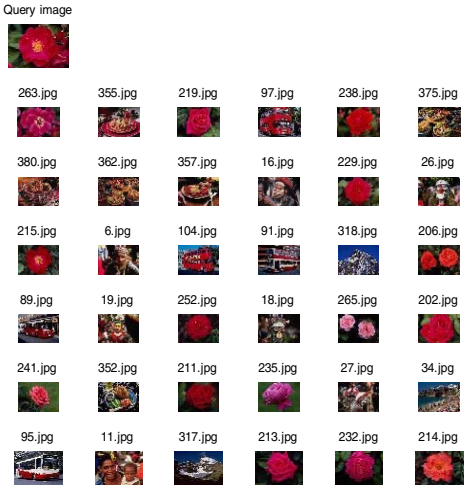


Figure 4. the results of LPP

We applied the PCA-LPP to images of Flower, Elephant and Dinosaur semantic. These image data set are mapped to a 30-dimensional space using different dimensionality reduction algorithm, PCA-LPP, PCA and LPP. In this subsection, we chose $N = 36$, K notes relevant image numbers. For example, there are 30 relevant images in Flower semantic, so $K = 30$, $Accuracy = 30/36$. The experimental results are shown in Table 1, As can be seen, PCA-LPP extends the advantage of PCA and LPP , and it performs much better PCA and LPP.

Algorithm / category	PCA-LPP	PCA	LPP
Elephant	$20/36$	$17/36$	$12/36$
Dinosaur	$29/36$	$29/36$	$16/36$
Flower	$30/36$	$29/36$	$14/36$

Table 1. Comparison for the Three semantic category

V. CONCLUSIONS

In this paper, we introduce a linear dimensionality reduction algorithm called PCA-LPP. It extends the advantage of the PCA and LPP algorithms. The method concentrates on a issue of reduced dimensionality. The algorithm can be applied to image retrieval. And it gives an acceptable image retrieval success rate in comparison with very famous algorithms such as PCA and LPP.

ACKNOWLEGEMENT

Thanks to the financial support from Zhejiang Provincial Natural Science Foundation of China under Grant No. Y1101269.

REFERENCES

[1] Ian T. jollife .Principal Component Analysis[M]. Apringerverlag,New York,1986

[2] J. Karhunen and J. Joutsensalo," Representation and separation of signals using non linear PCA type learning", Neural Networks, 7(1),1994.

[3] Richard O.Duda,Peter E.Hart and David G.Stork.Pattern Classification,Second Edition[M].beijing,2004.2

[4] Roweis S, Saul L. Nonlinear dimensionality reduction by locally linear embedding [J].*Science*, 2000, 290 (5500) :2323 - 2326.

[5] M. Belkin and P. Niyogi.Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering[C],In Proceedings of NIPS, 2001:585-591.

[6] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol 290, 22 December 2000.

[7] He Xiaofei , Niyogi P. Locality preserving projections [C] PPProceedings of Advances in Neural Information Processing Systems 161 MA : Cambridge , MIT Press , 2004 : 153 – 160

The Research of Multi-hop Routing Algorithm in the Field of Distributed Wireless Sensor Network

Mingming Wu
School of Internet of Things Engineering
Jiangnan University
Wu Xi, Jiang Su, China
E-mail: chu218@163.com

Wenbo Xu
School of Internet of Things Engineering
Jiangnan University
Wu Xi, Jiang Su, China
E-mail: xwb@jiangnan.edu.cn

***Abstract**: To improve the energy efficiency of nodes in Wireless sensor network, A multi-hop clustering routing algorithm is presented which is based on distributed wireless sensor network ,which can improve itself in the process of use according to a kind of protocol called low energy adaptive clustering hierarchy (Low Energy Adaptive Clustering Hierarchy, LEACH).*

Keywords: LEACH; Wireless Sensor Network; Multi-hop; Distributed

Distributed wireless sensor network (DWSN) is constituted by a large number of nodes which is limited of the energy and resources, and have self-organizing form of data collection, detection, strong control. It is always used in military affairs, medical science, industry, agriculture and other fields. Constrained in the limited energy， reducing energy consumption from the wireless touch, extending the network lifetime and network transmission reliability and stability is one of the hot spots in the study and application of wireless sensor network. So, design an efficient routing algorithm is a focus of WSN research.

With further research, a variety of routing algorithms have been proposed for the effective use of node energy, such as LEACH [1], SEP, HEED, etc. LEACH is the first hierarchical WSN routing algorithm proposed based on the structure of a cluster. Its basic idea is selecting cluster head by randomly cycle, evenly distributing the energy load of the entire network to each sensor node, which can reduce network consumption and increase the lifetime of the overall WSN [2]. But this algorithm requires that the sensor node and Sink nodes can communicate directly, so the expansion of the network is not strong, and does not apply to large network.

According to the "hot spots" of LEACH algorithm, combining the mechanism of multi-hop routing, this paper puts forward an improved algorithm, obtains better energy efficiency and higher network living space by comparing with the simulation and original LEACH

I. TOPOLOGY OF DISTRIBUTED WIRELESS SENSOR NEYWORK

Distributed sensor network [2] is one of the network application systems which is consisted by large sensor nodes deployed intensive in the monitored area. As the sensor node is numerous, we can only deploy them by the way of randomly launch, which cause that the position of sensor nodes cannot determine beforehand; at any moment, the nodes connected through wireless channel, and self-organizing network topology; the sensor nodes have a very strong coordinating ability, completing the overall situation task by the partial data collection, the pretreatment, as well as the data interaction between nodes.

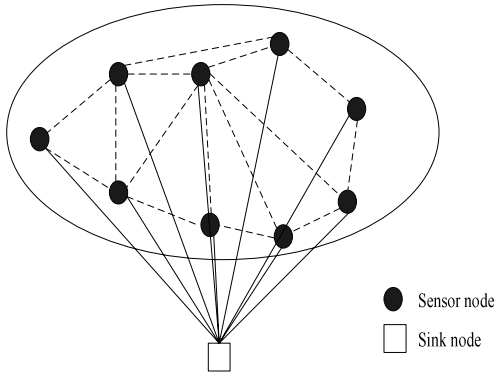


Figure (1)

In the topology of wireless sensing network, each node can exchange the data directly, transport data to the sink node [3]. Then sink node analyze the data received, and transform it to the data of entire wireless network, which is similar with Ad hoc. The network architecture is shown in Figure (1). From the figure, we can find that data can be transported between every two nodes, even between every two sensor nodes and the sink nodes [4].

II. THE MODEL OF ENERGY TRANSPORTATION IN DWSN

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.96

With regard to the energy consumption of sensor nodes in wireless communication, we adopt the models of free-space transmission and multi-attenuation, according to the distance between wireless sensors, the distance of receiver and transmitter between wireless sensor and base station. If the distance between the receiver and transmitter is less than a certain threshold d_0, then choose free-space model; if the distance between the receiver and transmitter is more than this certain threshold d_0, then choose multi-attenuation model. Then, the wireless channel transport *l* Byte data to the receiving device which is d apart from it, the energy consumption is [5]:

$$E_{Tx}(l,d) = E_{Tx-elec}(l) + E_{Tx-amp}(l,d) = \begin{cases} lE_{elec} + l\varepsilon_{fs}d^2 & d < d_0 \\ lE_{elec} + l\varepsilon_{amp}d^4 & d \geq d_0 \end{cases} \quad (1)$$

The energy consumption of receiving *l* Byte data is:

$$E_{Rx}(l) = E_{Rx-elec}(l) = lE_{elec} \quad (2)$$

Among them, $E_{Tx}(l,d)$ stands for energy consumption of transmitter; $E_{Rx}(l)$ stands for energy consumption of receiver; E_{elec} stands for energy consumption of transmitting circuit; ε_{fs} and ε_{amp} are coefficients of the amplifier; d_0 is threshold；the reference values of them are：

E_{elec} =50nJ/b, ε_{fs} =10pJ/(bm2), ε_{amp}=0.0013pJ/(bm^4) , d_0=87m.

The energy consumption of getting together *m* byte information is:

$$E_{DA} = mlE_{da} \quad (3)$$

Among them, E_{da} stands for energy consumption of getting together one byte data.

In distributed wireless sensor networks, we can set up a single hop or multi-hop data transmission, as shown in Figure 2, the length of AB and the length of BC are both less than the length of AC.

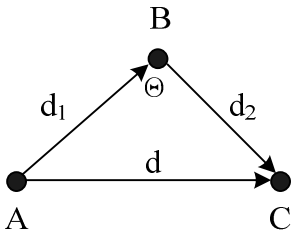


Figure (2)

Now, suppose to deliver *1* byte data from A to C:

According to the model of wireless network transmission, the energy consumption while transport from A to C directly is:

$$\begin{cases} 2lE_{elec} + l\varepsilon_{fs}d^2 & d < d_0 \\ 2lE_{elec} + l\varepsilon_{amp}d^4 & d \geq d_0 \end{cases} \quad (4)$$

The energy consumption while transferring from A to C through B is:

$$\begin{cases} 4lE_{elec} + l\varepsilon_{fs}(d_1^2 + d_2^2) & d_1, d_2 < d_0 \\ 4lE_{elec} + l\varepsilon_{amp}(d_1^4 + d_2^4) & d_1, d_2 \geq d_0 \end{cases} \quad (5)$$

According to the triangle relations knowable, we can get this:

$$d^2 = d_1^2 + d_2^2 - 2d_1^2 d_2^2 \cos(\theta) \quad (6)$$

So, θ is smaller than 90 degrees, and is smaller than 180 degree, that is $d^2 > d_1^2 + d_2^2$,

$$d^4 = d_1^4 + d_2^4 - 4(d_1^2 + d_2^2)d_1^2 d_2^2 \cos(\theta) + 2d_1^2 d_2^2 + 4d_1^4 d_2^4 \cos^2(\theta) \quad (7)$$

The same reason can derive: $d^4 > d_1^2 + d_2^2$.

Each parameter's reference value is substituted in formula (4), (5), then if $d < d_0$, we can get that the energy consumption of single-hop is:

$$E_1 = (10^{-7} + 10^{-11}d^2)l \quad (8)$$

If $d < d_0$, the energy consumption of multi-hop is：

$$E_2 = [2\times10^{-7} + 10^{-11}(d_1^2 + d_2^2)]l \quad (9)$$

Because $d^2 \le 87^2 = 7569$, Then we can get this：

$$10^{-11} d^2 < 10^{-7} + 10^{-11}(d_1^2 + d_2^2) \qquad (10)$$

So, while $d < d_0$, $E_1 < E_2$,that is the energy consumption of single-hop is less than multi-hop;

While $d > d_0$, the energy consumption of single-hop is:

$$E_1^{'} = (10^{-7} + 1.3 \times 10^{-13} d^4) l \qquad (11)$$

The energy consumption of multi-hop is:

$$E_2^{'} = [2 \times 10^{-7} + 1.3 \times 10^{-13}(d_1^4 + d_2^4)] l \qquad (12)$$

Because：

$$\begin{aligned} & d_4 - (d_1^2 + d_2^2) \\ & = -4(d_1^2 + d_2^2) d_1^2 d_2^2 \cos(\theta) + 2 d_1^2 d_2^2 + 4 d_1^4 d_2^4 \cos^2(\theta) \\ & < 2 d_1^2 d_2^2 < 2 \times 43.5^4 \approx 7.1 \times 10^6 \end{aligned} \qquad (13)$$

Then：

$$E_1^{'} - E_2^{'} > 0 \qquad (14)$$

So, While $d > d_0$, $E_1^{'} > E_2^{'}$, that is the energy consumption of single-hop is more than multi-hop;

Summary of the above, we can get a conclusion: while transporting distance is more than threshold d_0 , multi-hop way can use energy in wireless sensor network more efficient, and network lifetime can get extended.

III. THE IMPROVEMENT OF MULTI-HOP ROUTING PROTOCOL ALGORITHM IN DWSN

A. The Introduction of LEACH

We Select cluster nodes randomly by the way of cycle, evenly distributes the energy load of the entire network to each sensor node, in order to reduce network energy consumption and increase the lifetime of the entire network [6]. Simulation results show that LEACH can extend the life cycle of the network by 15% [7], comparing with the general plane multi-hop routing protocols and static hierarchical algorithm.

LEACH routing protocol can be divided into two stages mainly: the cluster building phase and stably operating phase. The sum time that cluster building phase and the stably operating phase lasting is calculated as one round. In order to reduce protocol's spending, the time of stably operating phase must be longer than the cluster building phase [8].

During the cluster stably operating phase, sensor node randomly generates a random number between 0 and 1, then compares it with the threshold T (n). If it is less than the threshold, the node will be selected as cluster head. T(n) can be calculated according to the formula (15) [1]:

$$T(n) = \begin{cases} \dfrac{P}{1 - P \times (r \bmod \frac{1}{P})} & n \in G \\ 0 & others \end{cases} \qquad (15)$$

In this formula: P stands for the percentage of how many nodes are selected to be cluster nodes, r stands for the current number of round, G is the node collection which was not selected to be cluster node in the last 1/p rounds. After the cluster nodes are chosen, they broadcast the news of becoming cluster head, then decide which group to join according to the intensity of news received, and inform the corresponding cluster head, complete the process of establishing clusters. Then cluster head nodes adopt the way of TDMA [9], to distribute the time of delivering data to the members of clusters.

In stabling stage, sensor nodes deliver the data they collected to cluster head node. Cluster head node transport this information to sink node after mixed the collected data. Then the sink node transmits data to the monitoring center, and the center will process all the data. After the stabling stage continues period of time, the network enters the cluster building stage again, carries on the next round of cluster's reconstruction, circulates like this uninterruptedly [10].

Through the simulation of Mat lab, like Figure (3) shows that there have 100 nodes in the 100*100 region.

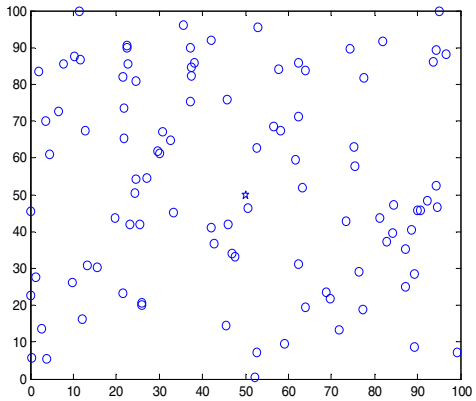


Figure (3)

Data can be transported between every two nodes. After transmitting the data, each node becomes a dead node which is showed in the figure (4). In this process, the number of times through the leach circulation is r=2632 times, Figure (5) is a simulation graph showed the number of nodes becoming cluster node and the surplus live nodes. During the cyclic process, I pay attention to the situation that the cluster head nodes appear expansible. According to this kind of problem, I proposes one kind of improved algorithm based on the LEACH.

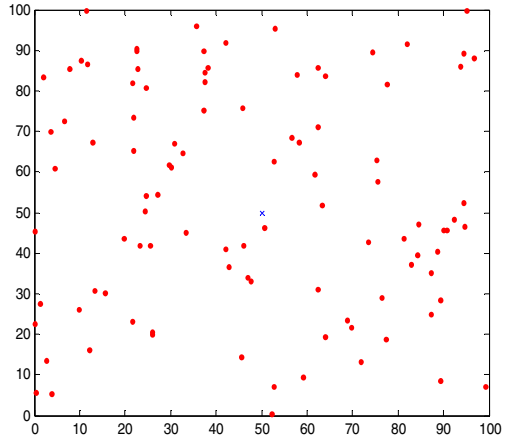


Figure (4)

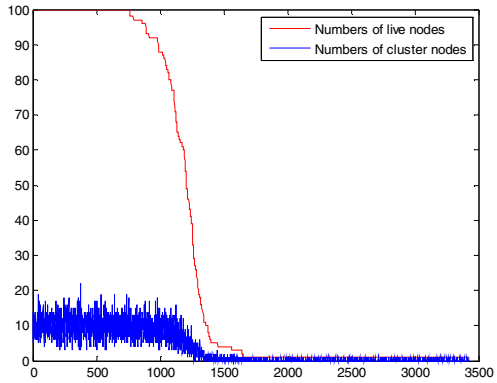


Figure (5)

B. *Improved Algorithm of LEACH*

1）selecting N sink nodes which are nearest away from the sink sensor in the region as the initial nodes.

2） the cluster head nodes broadcast to other nodes, select from N nodes which are nearest away from itself, follow by going to know all the nodes have been selected before.

3） divide nodes after the selection process into N most primitive areas of sink nodes, according to the largest selection of energy theorem, select the maximum energy in each regions of the node as cluster head, and communicate with the Sink node.

After the above improvements of the original LEACH algorithm, and survival through the MATLAB simulation compared cluster head nodes and show at figure (6)

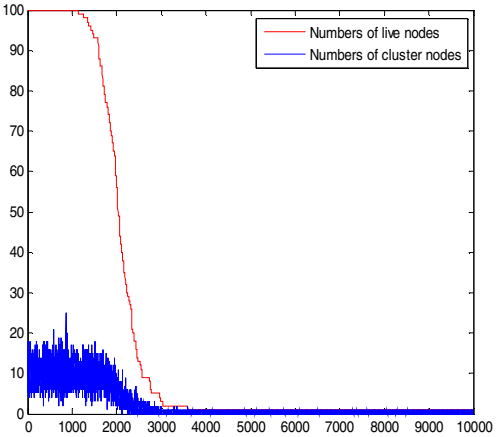


Figure (6)

C. *Conclusion*

From figure (5) and (6), we can see that the number of nodes in network begins to decrease from 1000 rounds in the improved routing algorithm, which starts to decrease from 800 rounds in the original LEACH algorithm. So when can get the conclusion that the improved algorithm can extend the network lifetime, and more efficient in use energy in net work than the original algorithm.

Ⅳ. REFERENCES

[1]HeinzelmanW,ChandrakasanA,BalakrishnanH. An Application Specific Protocol Architecture for Wireless Microsensor Networks[J].IEEE Transact ions on Wireless Communications, 2002, 1 (4) : 660 – 670

[2]SoummyaKar ,José M. F. Moura"Distributed Consensus Algorithms in SensorNetworks With Imperfect Communication: LinkFailures and Channel Noise" IEEE Transactions On Signal Processing, VOL. 57, NO. 1, JANUARY 2009

[3]Y.-J.Wen and A. M. Agogino, "Wireless networked lighting systems for optimizing energy savings and user satisfaction," in Proc.IEEE Wireless Hive Netw. Conf., Aug. 2008, pp. 1–7.

[4] LEI Yang,SHANG Feng jun,REN Yusell," Research Situation on Routing Protocols in Wireless Sensor Networks",Communications Technology,No.3,2009

[5] ZHOU Dongxin, JIN Wenguang,RONG Zhineng,"Layer Based Multi-Hop Clustering Routing Algorithm for W ireless Sensor Networks",Chnese Jouranal of Sensors and Actuators,No.1,Jan,2011

[6] T. Li and Y. Fujimoto, "Control system with high-speed and realtime communication links," IEEE Trans. Ind. Electron., vol. 55, no. 4, pp. 1548–1557, Apr. 2008.

[7] R.Vedantham, “Energy-efficient network protocols for wireless sensor and actor networks,” Ph.D. dissertation, School Elect. Comput Eng., Georgia Inst. Technol., Atlanta, GA, Dec., 2006.

[8] X.Cao, J.Chen, C.Gao, and Y.Sun, “An optimal control method for applications using wireless sensor/actuator networks,” Comput. Elect . Eng., vol. 35, no. 5, pp. 748–756, Sep. 2009.

[9]R.Olfati-Saber, “Distributed Kalman filtering for sensor networks,” in Proc. 46th IEEE Conf. Decision Control, Dec. 2007, pp. 5492–5498.

[10]P.Park,C.Fischione,A.Bonivento,K.H.Johansson,A.Sangiovanni-Vincentelli,“Breath:a self-adapting protocol forwireless sensor networks in control and automation”, in IEEE SECON, 2008.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

A New Anchor-based Localization Algorithm for Wireless Sensor Network

WANG Jianguo[1] WANG Zhongsheng[1]
[1] School of Computer Science & Engineering
Xi'an Technological University
Xi'an, China
e-mail:wjg_xit@126.com

ZHANG Ling[1] SHI Fei[1] SONG Guohua[2]
[2] Taxation Server Offerce
WuXi Municipal Bureau of State Taxation
WuXi,China
e-mail: wzhsh1681@163.com

Abstract—Anchor plays a very important role in range-based localization algorithm. In order to solve disadvantages of traditional localization algorithm based on anchors, this paper proposes a new anchor-based localization algorithm for wireless sensor network. It optimizes anchors and creates an optimized anchors database. Anchors are chosen from the optimized anchors database to measure the distance between the unknown node and the anchor. Trilateration localization algorithm or Euclidean algorithm is used to locate unknown nodes. After unknown nodes have been located, they are used as new anchors so as to reduce the dependency of localization algorithm on anchors and the requirement that anchors should be uniformly distributed in the network. Simulation experiments results show that the new anchor-based localization algorithm can solve disadvantages of traditional localization algorithm based on anchors and can improve the localization accuracy.

Keywords-optimized anchors databases; anchors-based localization algorithm; wireless sensor network; anchor

I. INTRODUCTION

At present, there are two positioning technologies for unknown nodes in WSN(Wireless Sensor Network) [1~3], range-based and range-free localization algorithms[4~5] according to whether measuring the range or not. There are also two positioning technologies, anchor-based and anchor-free localization algorithms[5~6] according to whether using anchors or not. In WSN, the location information of anchor is generally pre-deployed and set by hand or obtained by other positioning systems, for example GPS. Anchor-based positioning technology is widely used because it can improve the positioning accuracy and is also convenient.

However, anchor-based localization algorithms exist following disadvantages. 1) They have high dependence on anchors. They require that adequate anchors are deployed in WSN. But, anchors are expensive and have high energy consumption, so the number of anchors in the WSN should be controlled within an appropriate range[7]. 2) They require that anchors are uniformly distributed in the WSN. But many environments that people can not enter, such as the battlefield environment, natural disasters environment and so on, can not meet this requirement, therefore, in these environments, sensor nodes are often randomly deployed. 3) Anchors are randomly chosen to position unknown nodes. However different anchors have different influence on the positioning accuracy because of the instability of wireless communication and characteristics of anchors.

In order to solve above disadvantages and improve the positioning accuracy, this paper proposes a new anchor-based localization algorithm for WSN. Its basic idea is as follows. Firstly, some anchors are optimized, and an optimized anchors database is created. Secondly, the distance between unknown node and the anchor is measured by using anchors from the optimized anchors database. Lastly, the unknown node is located by using trilateration localization algorithm or Euclidean algorithm. In positioning phase, unknown nodes that have been located are used as new anchors so as to reduce the dependency of localization algorithm on anchors and the requirement that anchors should be uniformly distributed in the network.

II. NEW LOCALIZATION ALGORITHM BASED ON ANCHORS AND RANGE TECHNOLOGY

The main aim of this improved algorithm is to avoid randomly choosing and using anchors, reduce the dependency of localization algorithm on anchors and the requirement that anchors should be uniformly distributed in WSN, and improve the localization accuracy of WSN nodes.

A. Creating optimized anchors database

How to choose anchors has a great impact on positioning accuracy in anchor-based localization algorithm. Traditional anchor-based localization algorithms often randomly choose anchors. This paper tries to optimize anchors before unknown nodes are located, and optimized anchors databases are created.

The basic idea of creating optimized anchors databases is as follows. Firstly, some anchor is assumed as an unknown node, and it is located by using all anchors within its communication scope. Anchors are sorted and numbered according to their positioning accuracy, and then saved into optimized anchors databases.

For example, Fig.1 shows an irregular deployment of network nodes, assume that nodes 1 to 5 are anchors, and node 6 is an unknown node. These six nodes are within communication scope, thus they can communicate directly [8].

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.5

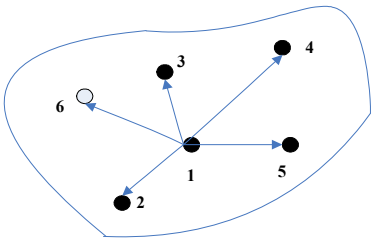


Figure 1 Deployment of network nodes

Firstly, assume that node 1 is an unknown node, then it can be positioned by using nodes 2 to 5. From range-based localization algorithm, we know that an unknown node can be positioned by using three different nodes within communication scope [8]. So, any three nodes from node 2 to 5 can be chosen to position node 1, i.e., the set of anchors (2,3,4), (2,3,5), (2,4,5) or (3,4,5) can be selected. In advance, the above four sets are respectively tried to use to position the node 1, and is numbered 1, 2, 3 or 4 according to the positioning accuracy from high to low. At the same time, they are saved into the optimized anchors database. When some unknown node need to be positioned in the future, its best set or the second best set can be chosen (when the best set of anchors is failure) for positioning. Then, according to the same method, the best positioning anchors set of other anchors can be archived. Therefore, the optimized anchors database of the whole network can be created.

B. *Choosing reference node*

When an actual unknown node is positioned, first of all, all of RSSI values from the unknown node to all of adjacent anchors are measured. Secondly, the anchor with the strongest RSSI value in all of adjacent anchors is chosen as a reference node. Finally, we can determine the optimized anchor database according to the reference node

For example, in fig.1 node 6 is an unknown node, other nodes are anchors. Firstly, node 1 is chosen as reference anchor according to the rule that the anchor with the strongest RSSI value is chosen, here, assume that anchor 1 has the strongest RSSI value in all of adjacent anchors from node 1 to 5. Then the optimized anchor database of anchor 1 is chosen to measure distances from node 6 to all nodes in the optimized anchor database.

C. *Ranging for unknown nodes*

RSSI is used to create optimized anchors databases. This ranging technology based on signal strength indication does not increase any additional hardware burden to the network[9~10], because sensor nodes usually have been fixed radio transceiver. However, the RSSI ranging technology has a rough and poor accuracy. Its measurement accuracy is affected by the distance between two nodes. If the distance is shorter, then the received signal strength will be stronger, so the measurement accuracy is better. On the contrary, the measurement accuracy is worse. Therefore, in this paper RSSI only be used to measure some distances, then other side distances are estimated by using the structure characteristics of the WSN.

The deployment of WSN can usually be treated as an undirected graph. Assume that an undirected graph is defined as G = (V, E) in which V means the set of sensor nodes in WSN, E means the set of distances between two communication nodes. Then, assume that G has N nodes which contain m anchors, and the rest are unknown nodes. If two nodes i and j can directly communicate each other, then which shows that there exists an edge e (i, j) between them. Each side in the network is associated with the RSSI value from node i to j.

Set p_{ij} as the RSSI value received by node i from j. Then in the corresponding undirected graph G, the edge e (i, j) can be expressed as (1):

$$p_{ij} = \Pi_0 - 10 n_p \lg(d_{ij} / \nabla_0) \quad (1)$$

In (1), the unit of P_{ij} is dbm, d_{ij} is physical distance from node i to j. Π_0 is the received energy relative to the reference distance ∇_0. n_p is an path loss exponent which associates with the environmental factor and usually takes value from 2 to 4.

Similarly, in the undirected graph G, another edge e (k, l) can be expressed as (2):

$$p_{kl} = \Pi_0 - 10 n_p \lg(d_{kl} / \nabla_0) \quad (2)$$

If set the value ∇_0 as 1m, (3) can be obtained with (1) and (2):

$$d_{ij} = 10^{\frac{p_{ij} - \Pi_0}{p_{kl} - \Pi_0} \lg d_{kl}} \quad (3)$$

According to (3), values of other sides can be estimated when the distance between any two nodes is measured. Therefore, calculations times of measuring distances can be reduced, and also the energy consumption of nodes can be reduced in the network. Thus, we can firstly measured distances between anchor and unknown node by using optimized anchors databases and RSSI ranging technology, then use the distance estimating formula((3)) to compute the distances. This method can not only improve the positioning accuracy but also reduce energy consumptions of sensor nodes.

D. *Positioning technology based on optimized anchors databases*

In positioning phase, we firstly position an unknown node by using trilateration algorithm, or using Euclidean positioning algorithm when the unknown node can not meet the requirement of the trilateration algorithm. Furthermore, unknown nodes can be used as new anchors after they have been positioned, and they are called of pan-anchors. In the case of anchors can not be uniformly distributed in the network, optimized pan-anchors databases are created according to above procedure, and pan-anchors are chosen from the relative optimized pan-anchors database to position the unknown node. Therefore, on the one hand, the requirement that anchors should be uniformly distributed in

the network can be reduced, on the other hand, with the positioning process forward, in theory the ratio of anchors is gradually increasing, positioning accuracy is also gradually improved.

1) Trilateration positioning algorithm

The location requirement of sensor nodes for trilateration positioning algorithm is shown as fig.2 in which nodes 2,3 and 4 are anchors, node1 is an unknown node. Node 1 is exactly located at circles with anchors 2,3 and 4 as centers, and in the point of intersection of three circles.

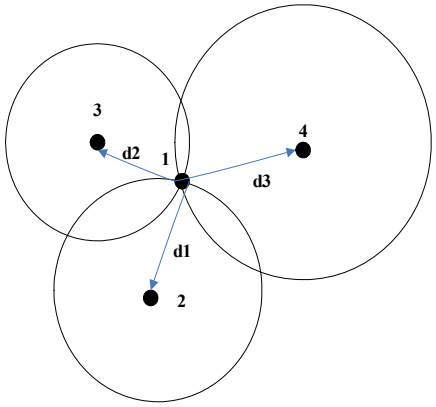


Figure2 location requirement for trilateration algorithm

Suppose that coordinates of anchors 2,3,and 4 are respectively (x1,y1), (x2,y2), (x3,y3), and corresponding distances to the unknown node (x,y) are respectively d1, d2 and d3, then, the coordinate (x,y) of the unknown node can be computed by using (4).

$$\begin{bmatrix} x \\ y \end{bmatrix} = \frac{1}{2} \times \begin{bmatrix} (x_2 - x_4) & (y_2 - y_4) \\ (x_3 - x_4) & (y_3 - y_4) \end{bmatrix}^{-1}$$

$$\times \begin{bmatrix} x_2^{\,2} - x_4^{\,2} + y_2^{\,2} - y_4^{\,2} + d1^2 - d3^2 \\ x_3^{\,2} - x_4^{\,2} + y_3^{\,2} - y_4^{\,2} + d2^2 - d3^2 \end{bmatrix} \qquad 4)$$

2) Euclidean positioning algorithm

Trilateration algorithm has a stringent requirement that the unknown node must be exactly located in circles with three anchors within its communication scope as centers, and in the point of intersection of three circles[11]. It is too strict to many unknown nodes meet this requirement. Therefore, in this paper, Euclidean algorithm is used when the unknown node can not meet this requirement.

The positioning principle of Euclidean is shown as fig.3 in which node A is an anchor, nodes B,C and D are unknown nodes. Set the communication radius of node A as R, nodes B and C are located within the communication scope, so they are neighbour nodes of node A, and can directly communicate with each other. Node D is out of the communication scope, so can not directly communicate with node A, but it can do this by using node B or C as the bridge.

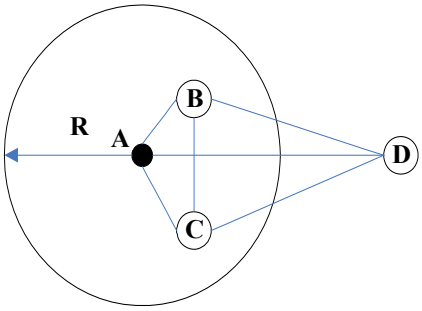


Figure3 The positioning principle of Euclidean

Distances from node A to B (AB) can be obtained by using RSSI，in the same way, AC,BC,BD and CD can be obtained. Then, the distance from A to D(AD) can be computed by using (5).

$$AD^{\,2} = AB^{\,2} + BD^{\,2} - 2 \times AB \times BD \times COS\ (\angle ABC + \angle CBD) \qquad (5)$$

III. SIMULATION EXPERIMENTS

A. Experimental environment

Network simulation software(NS2) is used to simulate experiments for the improved algorithm in this paper. Suppose that simulation scenarios and assumptions are as follows:

- 30 sensor nodes are randomly deployed in a $1000 \times 1000 m^2$ square area which is shown as Fig.4.
- Assume that all of unknown sensor nodes have same functions.
- Assume that probabilities of normal or failure communication between any two nodes are invariable.
- In deployed 30 nodes, there are four anchors, rest other nodes are unknown nodes. Locations of anchors are set in advance, for example the location of anchor 8 is (483,687).
- The average connectivity of the network is 5.

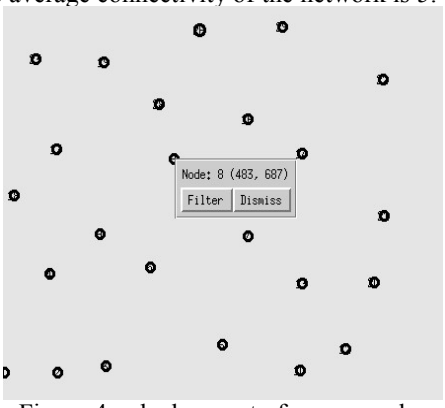
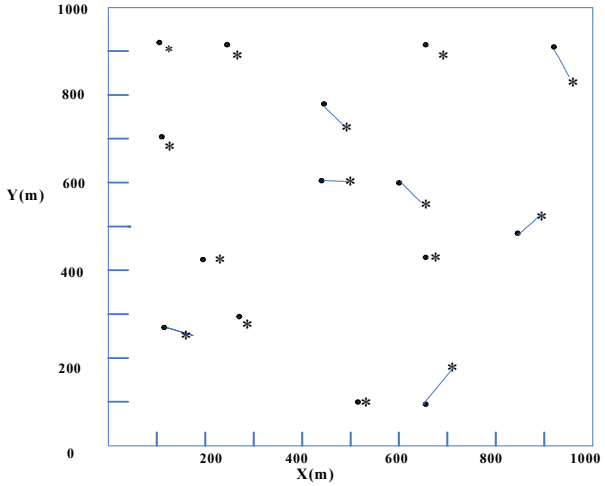


Figure 4 deployment of sensor nodes

B. Simulation analyzing

In this experimental, the communication radius is set to 50m. TwoRayGround acts as the radio propagation

model. In order to verify the effectiveness of the improved algorithm, two types of experiments have been implemented, one using traditional positioning algorithms and the positioning results is shown as Fig.5, another using improved localization algorithms and the results is shown as Fig. 6.

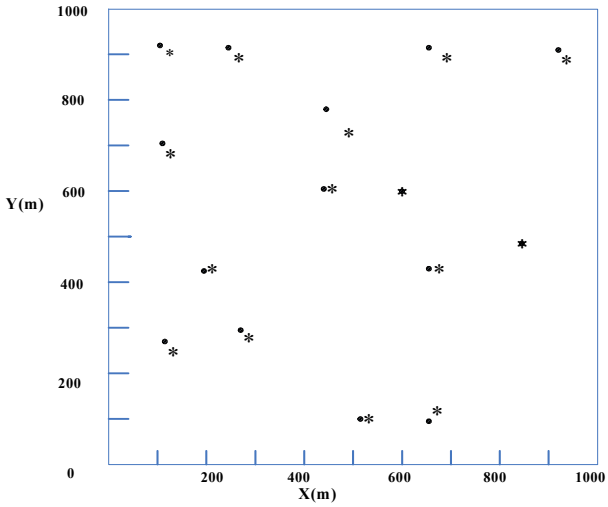


○ actual locations of unknown nodes

* Estimation locations of unknown nodes

Figure 5 Results of experiments with traditional algorithms

From fig.5 and fig.6, we know that there are large deviation between the estimation location and the actual location with traditional positioning algorithms, while there are less deviation between the estimation location and the actual location with improved localization algorithms.



○ actual locations of unknown nodes

* Estimation locations of unknown nodes

Figure 6 Results of experiments with improved algorithms

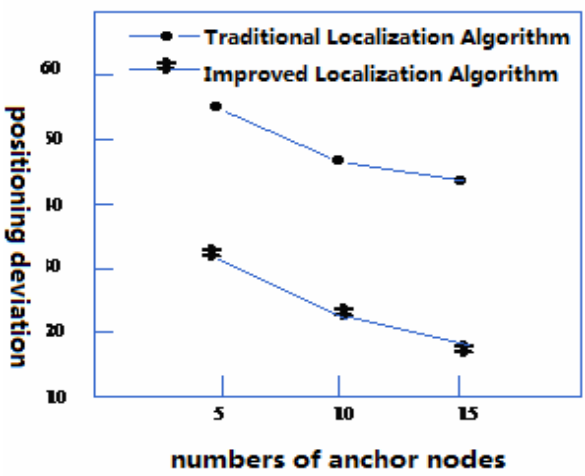


Figure 7 Results with different numbers of anchors

Fig.7 shows results that different numbers of anchors are chosen. From fig. 7, we can find that, with the increasing of numbers of anchors, traditional algorithm and improved algorithms can both improve the accuracy, but improved algorithms can improve more positioning accuracy. The reason is that, with the increasing of numbers of anchors, the effectiveness of optimized anchors database also increases, so it can decrease the deviation. In this paper, we need not add the number of true anchors, while use unknown nodes as new anchor after they have been positioned, and they are called of Pan-anchors. Therefore, with the positioning process forward, the ratio of anchors is gradually increasing, positioning accuracy is also gradually improved.

IV. CONCLUSION

In order to solve disadvantages of traditional anchor-based localization algorithms and further improve the positioning accuracy, this paper proposes a new anchor-based localization algorithm for wireless sensor network. It optimize anchors and create optimized anchors database so as to avoid randomly choosing and using anchors. Anchors are chosen from the optimized anchors database to measure the distance between unknown node and anchor. Unknown nodes are used as new anchors after they have been positioned, and they are called of pan-anchors. By using pan-anchors, on the one hand, the requirement that anchors should be uniformly distributed in the network can be reduced, on the other hand, with the positioning process forward, the ratio of anchors is gradually increasing, the positioning accuracy is also gradually improved. Simulation experiments results show that the localization algorithm proposed by this paper can solve disadvantages in traditional localization algorithm based on anchors and improve the localization accuracy. Simulation experiments results show that the localization algorithm proposed by this paper can solve disadvantages in traditional localization algorithm based on anchors and improve the localization accuracy.

REFERENCE

【1】REN Fengyuan, HUANG Haining, LIN Chuang, Wireless sensor networks,Journal of Software, 2003,14(2),PP:1148- 1157

【2】ZHANG Yihuang,The Studies on Wireless Sensor Network, Paper for Doctor degree,2009.2

【3】LI Jianzhong, LI Jinbao,SHI Shengfei, Concepts, Issues and Advance of Sensor Networks and Data Management of Sensor Networks, Journal of Software, 2003.10,PP:59-63.

【4】YU Ning, WAN Jiangwen, WU Yinfeng, Localization Algorithm in Wireless Sensor Networks，Chinese Journal of Sensors and Actuators，2007.1,PP:15-20.

【5】HAO Xiaohong,WANG Yinghui, KANG Yi,Study of localization algorithm for wireless sensor networks，Transducer and Microsystem Technologies，2008.6,PP:35-36

【6】LU Feng,Analysis on and Comparison of Localization Algorithm for Wireless Sensor Networks，Sci-Tech Information Development & Economy, 2008.7,PP:12-20.

【7】XU Benchong, CHEN Junjie. Localization Algorithms for Wireless Sensor Networks Based on Feasibility of Anchor Nodes Disposing, Measurement & Control Techniques,2007,8, P:46-51

【8】Azzedine Boukerche. Algorithms and Protocols for Wireless Sensor Networks . Wiley-IEEE Press.2008.

【9】XU Yan; SHI Jiang-hong; WU Xiao-fang，An Improved Localization Algorithm Based on RSSI-margin in WSN，Journal of Xiamen University(Natural Science)，2008. 5,PP:22-25.

【10】JIA Fu-li, LI Feng, ZHANG Ruihua，RSSI localization based on core in WSN，Computer Engineering and Applications, 2008.11,PP:2-6.

【11】GAO Lei, ZHENG Xiangquan, ZHANG Hong, A node localization algorithm for wireless sensor network based on trilateration and centroid algorithm, Journal of chongqing institute of technology (natural science) 2009.7, PP:26-31

【12】OU Yangyuu, WANG Chao,SHI Huichang, Euclidean Centralized Localization in Wireless Sensor Networks[J], Journal of shanghai university (natural science), 2008.2,PP:2-6

【13】YANG Qiang; WANG Zhong-jie，Study on localization algorithms for large-scale wireless sensor networks，Transducer and Microsystem Technologies，2007.2, PP:25-29.

Study on the numerical simulation of Heat transfer performance for multi-channel tubes based on fluid-solid coupled

Ying Yang[1,a], Wei Deng[1,b], Zhu Zhao[1,c]

[1]School of mechanical engineering and automation, Northeastern University, Shenyang 110819, PR China

[a]yangyingsy@163.com,[b]wwd824@126.com,[c]zz09074053@163.com

Abstract*—The author used the FLUENT software, established fluid-solid coupled finite element models of the multi-channel flat tubes in the parallel flow air-air intercooler, adopted three-dimensional numerical analysis for heat transfer process of hot side based on the SIMPLE algorithm, And selected the different channel numbers, hydraulic diameter, respectively, and compared the heat transfer rate, surface heat transfer coefficient and the friction factor. the Calculation results showed that the heat transfer performance and the resistance performance of multi-channel flat tubes, By comparing the heat transfer coefficient h with the theoretical calculation value ,the friction coefficient f and the correlation equation results, respectively, the errors are uniformly in six percentage and six-thirteen percentage, respectively, these data were well correlated. Consequently, this method be able to provide important reference for improving the heat transfer performance of the cooling tubes, and optimizing its structural design*

Keywords*—multi-channels; Fluent; Fluid-solid coupled; Numerical analysis*

1. Introduction

With the development of computational fluid dynamics and improvement of computer simulation technology and finite element analysis technology, in engineering analysis, CFD software used to study complex fluid heat and mass transfer. it has many advantages，such as，shorten the development cycle, save the products cost, improve products quality, simulate the real condition, idealized boundary conditions, et al. CFD software also has the high generalization advantage of grasping the question, at the same time, and providing the important theoretical basis for the optimization design of the automotive cooling system. Auto motors intercoolers adopt the air-air intercooled devices with the cross and immiscible flow fins and tubes, the structure is widely used in the energy, shipbuilding, vehicles, aviation, air conditioning, refrigeration and petrochemical fields. Yang and Web who are foreign scholars had completed a series of research work, reference [4] give the friction factor correlation which is measured for R-12 in multi-channel tubes with hydraulic diameters 1.56-2.64 mm .Another scholar[6] had studied on multi-channel tubes that hydraulic diameters were 2.72and5.45 mm , and put forward the friction factor correlation and compared with the $Blasuis$ equation. However, Domestic scholars also adopted numerical simulation method and different boundary conditions to study heat transfer, but most of them did not consider the fluid-solid coupled. In this paper, we study on the heat transfer performance and resistance performance for five different structural parameters that hydraulic diameters are 2.6-4.2 mm , and made a comparative study with single channel flat tube.

2. Computational physics model

In this paper，Fig.1a and 1b show the physics model of cooling tubes in intercoolers, its size is $62\times6\,mm$ and length is 652 mm ,thickness is 0.6 mm ,rib thickness is 0.4 mm , the hot fluid flow in the multi-channels. Each of flat tube is divided into eight rectangular channels with the same hydraulic diameter and two irregular sections, the heat will be transferred to outer wall surface through the conduction of the aluminum wall and taken away by convection with cooling air. These structural channels disturb the flow heat exchange boundary layers, strengthen the turbulence intensity of hot fluid and enhance the heat exchange. Obviously, this is a complex multi-dimensional heat transfer problem about solid-fluid coupled.

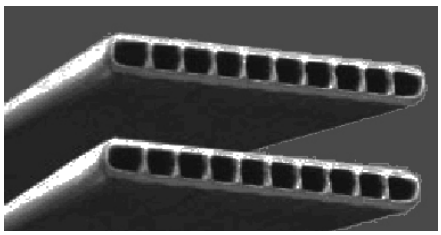


Figure1a .the Physical model of multi-channel tubes

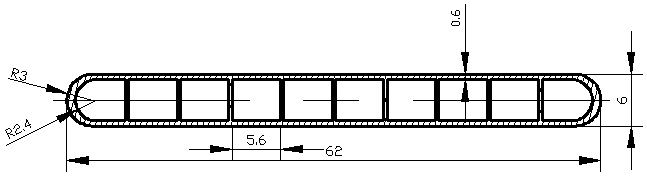


Figure1b. The photograph of cross-section for the multi-channels flat tube

The parameters of cross-section for each multi-channel flat tube are showed in Table 1.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.86

Table1.The parameters of cross-section for multi-channel flat tubes

NO	Channels number N	Long axle A/mm	Short axle B/mm	A/B	Cross section area S_j/mm^2	Heat transfer area S_{exc}/mm^2	Hydraulic diameter D_h/mm
1#	2	27.7	4.8	5.77	284.01792	88280	4.2
2#	4	14.4	4.8	3	278.25792	99236	3.65
3#	6	9.4	4.8	1.96	271.53792	109920	3.22
4#	8	7.1	4.8	1.48	269.61792	121920	2.88
5#	10	5.6	4.8	1.17	265.77792	133400	2.6
6#	Single channel flat tube						

2.1 Calculation equations and simplified assumptions

Based on the above structures, when establish the finite element models, we need to make some basic assumptions are as follows:

The hot fluid flow and heat transfer is in the steady state the continuous fluid medium is in the computational domain;, ignoring the thin flow

The Mach number is less than 0.3, incompressible, and the effect of fluid viscosity plays a significant role in fluid flow

Ignores the fouling resistance, the phase transition, reaction flow and the influence of gravity

The hot fluid be distributed uniformity to the micro-channels; considering the fluid-solid coupled of each micro-channel flow.the mass flow and temperature of hot fluid be distributed uniformity in the inlet, constant.

The differential form of computational equations are as follows.

The mass conservation equation is defined as :

$$\frac{\partial(\rho\mu_i)}{\partial x_i}=0 \tag{1}$$

The momentum conservation equation is defined as :

$$\frac{\partial(\rho\mu)}{\partial t}+div(\rho\mu\mu)=div(u.grad\mu)-\frac{\partial p}{\partial x}+S_\mu \tag{2}$$

$$\frac{\partial(\rho\upsilon)}{\partial t}+div(\rho\upsilon\mu)=div(u.grad\upsilon)-\frac{\partial p}{\partial y}+S_\upsilon \tag{3}$$

$$\frac{\partial(\rho\omega)}{\partial t}+div(\rho\omega\mu)=div(u.grad\omega)-\frac{\partial p}{\partial z}+S_\omega \tag{4}$$

The energy conservation equation is defined as :

$$\frac{\partial(\rho T)}{\partial t}+div(\rho\mu T)=div\left(\frac{k}{c_p}gradT\right)+S_T \tag{5}$$

The momentum equation in the solid computational domain does not exist, but the energy equation (5) still be applicable, in which the velocity of inlet is zero.

2.2 Boundary conditions

Inlet: mass flow

Outlet: pressure outlet

Symmetry: use Symmetrical boundary conditions

Outer wall: the second thermal boundary conditions

The interface between fluid and solid: coupled

Others: adiabatic boundary conditions in the reference [7,8]

For the heat transfer with fluid-solid coupled，we use the FVM[2] numerical method for solving. The computational domains of fluid and solid all use the idea of zoning and multi-block grid meshing, and the Hex/Wedge cells with the method of cooper. To ensure the accuracy of numerical calculation, the fluid domain which adjacent to solid domain is meshed by boundary layers. The finite element models are given in Fig.2a and 2b.

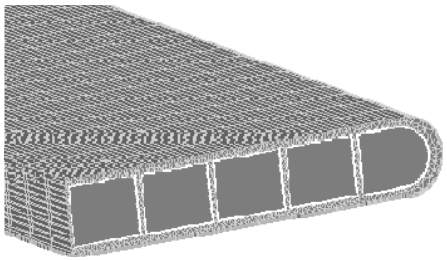


Figure 2a .Solid domain grid

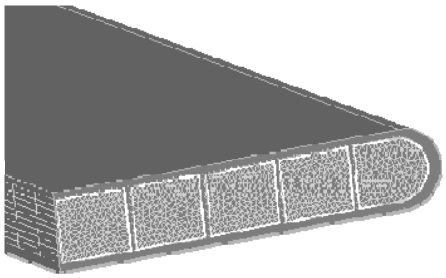


Figure 2b.Fluid domain grid

3. Results and discussions

For the heat transfer coefficient h in tubes is given by the equations:

$$h = \frac{Q}{S_{exc}\left(T_w - T_{ave}\right)} \tag{6}$$

where Q is the heat transfer rate,(w); S_{exc} is the heat transfer surface area in tubes,(m^2); T_w is the average wall temperature in tubes; T_{ave} is the fluid qualitative temperature, T_{ave} =（$T_i + T_o$）/2,(K); T_i is the average temperature of fluid in the inlet, T_o is the average temperature of fluid in the outlet.

3.1 Heat transfer correlation

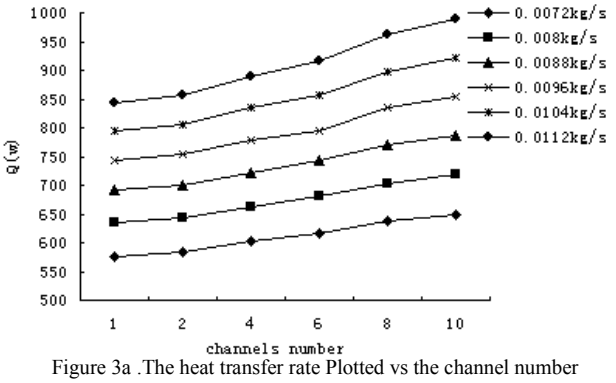


Figure 3a .The heat transfer rate Plotted vs the channel number

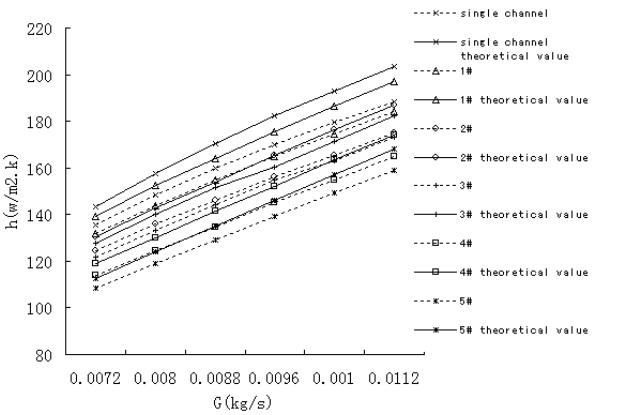


Figure 3b.The heat transfer coefficient plotted vs the channel number

Fig.3a shows the heat transfer rate plotted vs the channel number of flat tubes. for the same mass flow, as expected, the heat transfer of rectangular plain tubes are obviously better than the single channel tube, the heat transfer rate increases with increasing the channel number. This is because the channel number increases, the flow cross-sectional area of rectangular plain tubes are 1%.3.01%.5.35%.6.02%.7.36% less than the single channel tube respectively. The equivalent hydraulic diameter decreases 7.08%.19.2%. 28.8%.36.3%.42.5% respectively, which increases the average flow velocity and fluid turbulent intensity, the increase of flow velocity strengthens the scour of wall boundary layers, which improve the heat transfer performance and be conducive to increase the enthalpy difference between inlet and outlet. So it enhances the heat transfer rate. For the same structure parameters, the curve of heat transfer parallel to each other at different mass flow, the heat transfer rate increases with increasing the mass flow, but the growth rate slows down gradually, which indicate that the mass flow rate increases, it be capable of increasing the heat transfer coefficient and the total enthalpy difference between inlet and outlet, thereby, enhancing heat transfer

Fig.3b shows the heat transfer coefficient plotted vs the channel number of flat tubes. For the same structure parameters, the heat transfer coefficient increases with increasing the mass flow rate. for the same mass flow rate, the size sequence of heat transfer coefficient is the single channel tube＞1#＞2#＞3#＞4#＞5#.Form the structure parameters of flat tubes, the heat exchange area of each rectangular plain tube is 6.09%, 19.3%of 32.1%, 46.5%, 60.3 more than that of single channel tube respectively, the heat transfer rate of each rectangular plain tube is 1.6%, 4.7%, 7.1%, 11%, 13% more than that of single channel tube, respectively, with increasing the channel number of flat tubes, but the heat transfer coefficient of each rectangular plain tube is 2.65%, 8.04%, 10.2%, 16%, 20% less than that of single channel flat tube, respectively, so it is reasonable that the heat transfer coefficient decreases relatively. Therefore, when consider the fluid-solid coupled, the heat transfer area increases so quickly with increasing the channel number of flat tubes, it appears that the heat exchange area plays a significant role in affecting the heat transfer performance in each of rectangular plain tube. Also, for the same mass flow, By comparing the heat transfer coefficient with the theoretical results, the errors are in 6%, the numerical analysis data are provided a very good correlation

3.2 Friction factor correlation

The friction factor f of each tube using the *Blasuis* equation [4] as a reference

$$f_1 = 0.079\, R_e^{-0.25} \quad (2500 < R_e < 20000) \tag{7}$$

For the heat exchange of hot fluid flow in the multi-channel tubes, The total pressure drop includes the friction pressure drop, the accelerated pressure drop, the gravity pressure drop, the entrance pressure drop, the exit pressure drop. In the parallel flow intercooler, the hot fluid flow along horizontal direction, because no heat transfer occurred in the horizontal cross section, no acceleration and gravity terms are included in the total pressure drop, In foreign scholars Yang and Web [3,4], the ratio of friction pressure drop in total pressure drop had studied systematically, so here only consider the friction pressure drop and ignore the entrance pressure drop and the exit pressure drop. the friction factor is given by:

$$f = \frac{2\Delta p.D_h}{\rho.l.\mu^2} \tag{8}$$

where Δp is the pressure drop,(p_a); ρ is the fluid density, kg/m^3 ; μ is the fluid velocity, m/s ; l is the length of flat tubes, mm ; D_h is the hydraulic diameter of flat tubes, mm .

At the same mass flow rate, the friction coefficient in different parameters of flat tubes change with different channel number as shown in Fig.4a

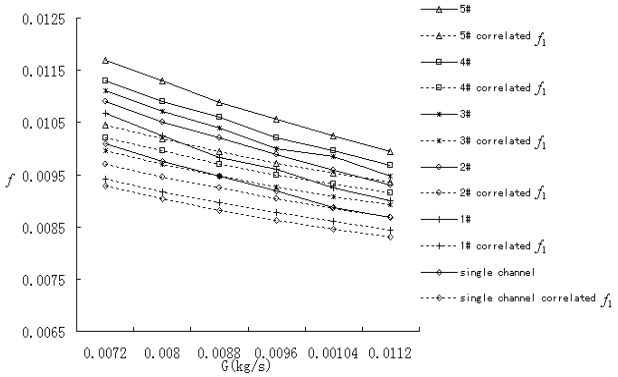


Figure4a. Friction factor plotted vs. the channel number

Fig.4a shows the friction factor decreases with increasing the mass flow rate at the same structure parameters; the value of descending order is 5#＞4#＞3#＞2#＞1#＞single channel tube. This is because at the same conditions in the inlet mass flow rate, the number of channels increases that lead to reduce the cross section area and hydraulic diameter, also, the high flow velocity and pressure drop impact the overall performance in the intercoolers. Therefore, the number of channels in flat tubes are not suitable for too much in the engineering.

By comparing the computational results of each tube with the *Blasuis* equation, the results show that the computational results are the average of 6%-13% higher than the correlation results when the Reynolds number is in 2500-20000 and the mach number is less than 0.3,based on the data range in the Fig.4a, the correlation is applicable. In Yang and Web [4] the research results were compared with the correlation equation, it shows that the measured results are the average of 14% higher than the correlation results, which shows the numerical simulation results are closer to Yang and Web measured results and further verify the correctness of computation model.

4. Conclusions

Established the coupled heat-transfer model about multi-channel tubes heat conduction and hot fluid flow heat-transfer for intercoolers and conducted the numerical simulation in the test conditions. By comparing the heat transfer coefficient h with the theoretical calculation value ,the friction coefficient f with the correlation f_1 results, respectively, the errors are uniformly in 6%and6%-13%, and f is more closer to the results of Yang and Web [4] measured

Obviously, the multi-channel tubes has good heat transfer performance than the single channel tube, the heat transfer enhancement effect increases with increasing the channel number of flat tubes, considering fluid-solid coupled, the heat exchange area plays a significant role in affecting the heat transfer performance in each of multi-channel tube. However, too much channels would increase the pressure drop and further affect the efficiency of the engine intake.

Select the structure parameters of cooling flat tubes for the parallel flow intercoolers, it should be considered the effect of flat tube parameters on heat exchange area, heat transfer coefficient and pressure drop to obtain the satisfactory heat transfer performance for flat tubes

References

[1] S. M. Yang and W. Q. Tao, Heat transfer (2006).

[2] J. Y. Tu and C.Q. Liu, Computational fluid dynamics (2009).

[3] Yang CY, Webb RL, Condensation of R-12 in small hydraulic diameter extruded Aluminum tubes with and without micro-fins 39(4) (1996), 791-800.

[4] Yang CY, Webb RL, Friction pressure drop of R-12 in small hydraulic diameter extruded aluminum tubes with and without micro-fins 39(4) (1996), 801-809.

[5] B.L.Xiao, X.L.Yu,S. Han, The study of effects of fin parameters on thermal hydraulic performance of a vehicular charged air cooler 44(11)(2010).

[6] M. W. Wambsganss, J. A. Jendrzejczyk, D. M. France and N. T. Obot, Two-phase flow patterns and frictional pressure gradients in a small rectangular channel : a comparison between two horizontal orientations, Argonne National Laboratory Report No. ANL-90/46 (1990).

[7] G.X. Li, N. Li, and X. C. Zhang, Experimental study on turbulent characteristics in air-air intercoolers 16(1998), 373- 374.

[8] G.X. Li, X. J. Sun and Y.G.Li,3-D Numerical analysis for the heat transfer process in air-air intercoolers 17(1999), 361-364.

[9] M. W. WambsgzLnss, J. A. Jendrzejczyk and D. M. France, Two-phase flow and pressure drop in flow passages of compact heat exchangers, SAE Technical Paper Series, no. 920550 (1992).

[10] Y. L. He，L.W. Tang and W. Q. Tao, Performance evaluation of an inter-cooler for turbo-charger system of internal combustion engine (5) (2005), 447—449.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Access Selection in Always Best Connected Networks

Ziqian Tu
School of Internet of Things Engineering
Jiangnan Uninversity
Wuxi, China
e-mail: keppeltu@gmail.com

***Abstract*—In the recent years, the wireless communication is developing in a very tremendous speed. Various kinds of radio access technologies have been used all over the world that brings us an opportunity of being always best connected. In this paper, we will focus on the access selection of the network and make an architecture to give users a simple, effective and convenient way to be best connected. In designing the optimizer of the architecture, we take QoS and prices of the services both into consideration to satisfy the needs of most ordinary users who are not proficient in the wireless access services. Throughout analysis, this architecture demonstrates that it will benefit users a lot in the near future.**

Keywords:radio access technologies, always best connected, architecture

I. INTRODUCTION

In the recent years, the wireless communication is developing in a very tremendous speed. The second-generation (2G) cellular systems have been used in a worldwide area. The Global System for Mobile Communications (GSM) can enable the user to stay connected almost all over the world, and with the help of General Packet Radio Service (GPRS), the user can connect to the Internet at the speed of 115Kbps. The third-generation (3G), taking Universal Mobile Telecommunications System (UMTS) and Code-Division Multiple access 2000 (CDMA2000) as representation, has greatly improved the transmission data rate so that the 3G users can experience better mobile Internet services. Based on an all-IP packet switched network, the fourth-generation (4G) will provide much higher rate for the users. Meanwhile, in parallel with the mobile access, the local area access technologies have also changed people's daily life. For example, Wireless local area networks (WLAN) can also provide users with a high data rate.

Although the user can have a great deal of choices to choose the services from multiple access technologies, the user can barely know which service is the most suitable for him/her when he/she decides to use some applications. Different wireless networks have different bandwidths, timeliness, bit error rate and costs. As for a user, he/she may use different applications when he/she has different demands. Some of the applications may call for large bandwidth, while others may need low bit error rate (BER). For example, we do not need a large bandwidth when we use voice communications, but we do need a large one when we want to watch on-line video shows fluently. According to the scenario, it is necessary to design an architecture which can allow the user to be connected in the best way when facing the dynamic and heterogeneous access networks to best utilize the available services.

According to the concept and architecture of "Always Best Connected (ABC)" [1, 2], access selection should take terminal-based selection, network-based selection and user intervention into consideration. In this paper, we mainly concentrate in the issue of access selection from available wireless network. Reference [3] proposes a new architecture capable of supporting ABC service. In [4], the author demonstrates a terminal management structure to handle the problem of access selection. Reference [5] proposed a network selection that involves analytic hierarchy process (AHP) and gray relational analysis (GRA). Reference [6] modeled the problem of multi-constraint dynamic access selection (MCDAS) as a variant of bin packing problem. Reference [7] maps access selection of ABC service to a class of combinatorial optimization problems that are non-deterministic polynomial-time hard (NP-hard).

All these approaches either demonstrate the concepts, or attempt to take too many factors into consideration to be easily realized, so an appropriate architecture should be set up. We want to give the user a simple, effective and convenient way to be best connected to the service. Simple: Users do not need to know the features or characteristics of the services. What the user should do is just to choose whether to use their profiles or to use the service provided by the optimizer. Effective: The architecture we designed automatically chooses the best service to the user according to the application on account of QoS and cost that not only satisfies the demand of the users but also save money for the user. Convenient: The user-friendly operating interface will ask only a few questions for the user to collect useful information. So it will not bother users too much. The architecture will automatically complete the optimization of service access.

The rest of the article is organized as follows. First, we will describe the model of the architecture. Second, the procedure of the architecture will be present. Finally, we will conclude the paper and describe our future work.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.21

II. ARCHITECTURE

The architecture of network access is showed in figure 1. The access frame consists of three parts, they are: ABC Server, Network Detector and Network Optimizer.

ABC Server: The server directly connects with the user equipment (UE). It authorizes the user to access to the network which is determined by the network optimizer and store the user's information such as user's profile, user's account and user's authority.

Network Detector: The detector collects the real-time data about the services in a certain area. It detects what services can be supplied to the user and analyzes the services such as bandwidth, timeliness, bit error rate and cost. It will send the services which can be supplied and the status of the services to the ABC Server and Network Optimizer (listed below) separately.

Network Optimizer: The optimizer will analyze the status of the services and the application used by the user to determine which services is the best for users. Then it will send the results to the Server.

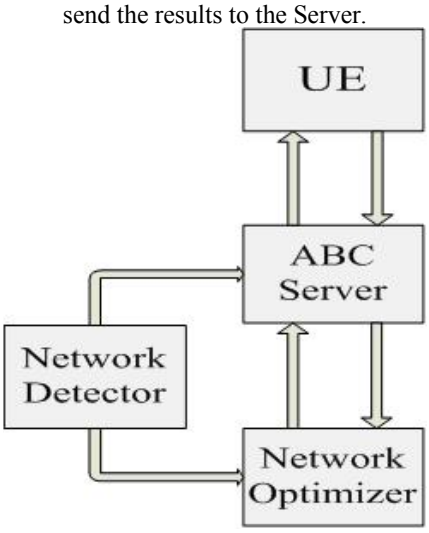


Figure 1. Architecture of Network Access

III. WORKING PROCEDURE

When the user decides to use some kind of application, the user will have two options:

Option One is to use the user's own profile stored in the server.

Option Two is to let the server supply the best suitable services for the user.

If the user choose option one, the UE will send a message to the server. The format of the message goes like this:

ID: 137-514-368
Application: Web Surfing
Type: Profile

When the server receives this message, it will automatically index the user's profile with user's ID. And the profile can be divided into 2 different types:

The first type we call it Correspondence Type (CT), because there is correspondence between Services and applications. The user configures the applications with services in one-to-one relationship. If the server cannot provide the user with the services as the user demands, the server will reply the user that the service is not available in this region and ask the user whether he or she is willing to connect to the service which has been determined by the optimizer according to the application the user wants to use.

The second type we call it Priority Type (PT). The reason is that the user lists his/her preferences about the services in a priority order. The server will try to access to the services as the user lists, from highest priority to lowest priority. If the server cannot provide the user with the services as the user demands, the server will reply the user that the service is not available in this region and ask the user whether he or she is willing to connect to the service which has been determined by the optimizer according to the application the user wants to use.

If the user chooses Option Two, the UE will send a message to the server in the format like this:

ID: 137-514-368
Application: Voice
Type: Optimization
services_supported:WLAN, GPRS, CDMA2000

When the server receives this message, it will forward the message to the optimizer. The first step optimizer will do is to make a comparison between the services which can be supported by the UE and the services which can be provided by the server to find out the same services. The Purpose of the first step to make the optimizer not analyze the services that cannot be supported by the UE is to save overhead and improve efficiency. Then the optimizer will analyze the application which the user is asked for and start processing.

We design a function to decide which the best service for the user's application is. By calculating the function, we can get a value. The service with the highest value will be the most suitable service which we refer to the best service in the area. Here, in this formula, we take the bandwidth, timeliness, bit error rate and cost into consideration. And we set different weights on different applications.

The function goes like this:

$$Value = W_B \cdot \ln\frac{B_a}{B_s} + W_T \cdot \ln\frac{T_s}{T_a} + W_E \cdot \ln\frac{E_s}{E_a} - W_C \cdot \ln\frac{C_a}{C_{\max}}$$

Where B_a , T_a , E_a and C_a represent the actual bandwidth, timeliness, bit error rate and cost of the service respectively, while B_s , T_s , E_s mean the lowest required bandwidth, timeliness, bit error rate of applications and $C_{\max}$ means the highest price of the services provided in the area. However, if the service is free, we set C_a as 0.01.

W_B , W_T, W_E and W_C are the weights of bandwidth, timeliness, cost and bit error rate. The weights vary

according to different applications and the sum of the weights is 1.

The optimizer will run the function to test the services. Different applications have different lowest required parameters (see table I). The service with highest value will be chosen.

We now use a case to describe the working procedure of our architecture. When a user is at home and he/she wants to give somebody a call: his/her UE will first connect with the server, and the server will reply the available services to the UE.(see Table II) Then the user decides whether he/she chooses his/her owe profile or the optimized service provided by the optimizer. The user chooses to use the optimized service. So when the user makes the call, the optimizer will set the W_B, W_T, W_E and W_C as 0.1, 0.35, 0.15 and 0.4 considering that users do not need a large bandwidth and they want to save money when they make a call. The optimizer will count the values of all the available network. The WLAN with the value of 3.07 which is the highest is selected. It will send the message goes like this:

ID: 137-514-368
Application: Voice
Type: Optimization
Service name: WLAN

to the server.

The server will supply the service which is determined by the Optimizer for the user. And the user will receive a message like this:

ID: 137-514-368
Application: Voice
Type: Optimization
Service name: WLAN
Available Bandwidth: 5Mbps
Timeliness: 150ms
BER: 10^{-5}
Cost: 0

Soon, when the user leaves home and takes a subway at the speed of 80km/h to his/her company, the WLAN and WiMAX802.16d is unavailable, so, if the user continues to talk, the optimizer will choose GPRS with the value of 1.33 as the best service for him/her. When the user finishes his voice phone, someone calls him with a video phone, and then the optimizer changes the weights. It sets W_B, W_T, W_E and W_C as 0.3, 0.2, 0.1 and 0.4. So the CDMA2000 is selected with the value higher than that of GPRS. Half an hour later, when the user arrives at his/her office, he/she has a large emergent file to download, and again, the optimizer changes the weights with W_B, W_T, W_E, W_C as 0.4, 0.1, 0.4 and 0.1. So the WiMAX802.16d is selected with the highest value.

IV. Conclusion and Future Work

An architecture capable of best utilizing the wireless network is proposed in this study. It provides users a simple, effective and convenient way to access the network with the consideration of users' actual demand under different situations. The architecture can not only save user's money with appropriate QoS but also can improve the utilization of heterogeneous wireless network.

Future research will focus on set of the parameters which correspond with different applications and the issue how to solve the problem when the optimizer indicates a great deal of UE to use the same service at the same time and causes the blocking.

TABLE I. The lowest required parameters for some applications

Application	Bandwidth	Timeliness	Bit error rate
Voice Phone	64kbps	200ms	10^{-3}
Video Phone	512kbps	250ms	10^{-5}
Big File Transmission	2Mbps	300ms	10^{-5}

TABLE II. Hypothetical standard parameters of some radio access technologies

Services	Bandwidth	Charge	Timeliness	Bit error rate	Mobility
WLAN	5Mbps	No charge	150ms	10^{-5}	Low
GPRS	115kbps	$0.1/min	200ms	10^{-4}	120km/h
CDMA2000	2Mbps	$0.3/min	250ms	10^{-4}	120km/h
WiMAX802.16d	20Mbps	$1/min	200ms	10^{-5}	Low

References

[1] Gustafsson, E.; Jonsson, A.; , "Always best connected," Wireless Communications, IEEE , vol.10, no.1, pp. 49- 55, Feb. 2003

[2] Fodor, G.; Eriksson, A.; Tuoriniemi, A.; , "Providing quality of service in always best connected networks," Communications Magazine, IEEE , vol.41, no.7, pp. 154- 163, July 2003

[3] Chen Yiping; Yang Yuhang; , "A new 4G architecture providing multimode terminals always best connected services," Wireless Communications, IEEE , vol.14, no.2, pp.36-41, April 2007

[4] Adamopoulo, E.; Demestichas, K.; Koutsorodi, A.; Theologou, M.; , "Intelligent Access Network Selection in Heterogeneous Networks - Simulation Results," Wireless Communication Systems, 2005. 2nd International Symposium on , vol., no., pp.279-283, 7-7 Sept. 2005

[5] Qingyang Song; Jamalipour, A.; , "Network selection in an integrated wireless LAN and UMTS environment using mathematical modeling and computing techniques," Wireless Communications, IEEE , vol.12, no.3, pp. 42- 48, June 2005

[6] Xing, B.; Nalini Venkatasubramanian; , "Multi-constraint dynamic access selection in always best connected networks," Mobile and Ubiquitous Systems: Networking and Services, 2005. MobiQuitous 2005. The Second Annual International Conference on , vol., no., pp. 56- 64, 17-21 July 2005

[7] Gazis, V.; Alonistioti, N.; Merakos, L.; , "Toward a generic "always best connected" capability in integrated WLAN/UMTS cellular mobile networks (and beyond)," Wireless Communications, IEEE , vol.12, no.3, pp. 20- 29, June 2005

THE REMOTE FAULT MAINTENANCE SUPPORT SYSTEM OF LOGISTICS CARRY VEHICLE BASED ON NETWORK *

Gongliang Jiang
School of Traffic and Transportation，Chongqing Jiaotong University
Chongqing, 400074，china
Email：JGL6388@sina.com

ABSTRACT

Road transport is one of the most important parts of logistics carry. In order to improve efficiency and quality of logistics carry, It is required that diagnose and maintain remote trouble carry vehicle rapidly. This paper researched the Remote fault Maintenance Support System of Logistics carry Vehicle based on network. Some techniques including embedded computer, wireless communication, network communication and database are used to build logistic carry vehicle remote fault diagnosis and maintenance system, in this system, carry vehicle's situation can be detected in time, and can be connected to the center of remote maintenance support by wireless communication. Trouble automobile can be remotely diagnosed and received maintenance technique assistance, realize remote diagnosis and maintenance to logistic carry vehicle. The efficiency and feasibility for reducing the time of faulty diagnosis and maintenance, sharing the information resource, improving the carry vehicle availability and credibility and reducing repaired cost are proved by application.

Keywords: network; Support System; Logistics Equipment, Carry Vehicle, Remote Maintenance, Failure Diagnosis

1. INTRODUCTION

Along with development rapidly of road transport establishment and automobile industrial in china, the carry ability by road has been enhanced constantly. Road transport has become one of the most important parts of modern logistics carry. It is most essential for logistics carry enterprise survival how to realize rapid carry, high efficiency, and low cost in logistics carry. in order to adapt the conveyance demand of great capacity, far distance, shortcut and high efficiency, the development direction of logistics carry vehicle is great type, complication , automaticity and more and more high technology. So the difficulty of fault diagnosis and maintenance is increased. because of the logistics carry vehicles often run on open country and different district, it has highly traffic intensity, long run time, great change traffic environment, therefore, the probability of fault occurred is more high relatively. When the Logistics carry vehicle faults occurred on operating road or different district, and the diagnosis and maintenance of the fault is not made timely, it would be happened that cargo carried could be send to destination in good time. Therefore, it will bring losing and infection badly to the logistics carries enterprise.

* *The research is supported by opening foundation of Chongqing key laboratory of communications engineering (N0.2008cqjyoo3).

The mode existing maintenance could not satisfy demand for meet an emergency diagnosis and maintenance of logistics carry vehicle on way rapidly. Therefore, The Remote Fault Maintenance Support System of Logistics Carry Vehicle based on network is required to establish. This system is built by high integrated electron technique in carry vehicle, modern network, communicate, and embed computer and database technique. The status of Logistics Carry Vehicle is supervised real time by diagnose system along with vehicle, in case the fault signal appeared，the fault information is collected and disposed in time, and send to Remote Fault Maintenance Support center. The remote diagnosis and maintenance for fault carry vehicle is made by the computer communication network. Through the remote maintenance decision, Remote fault carry vehicle can break the limit of traditional maintenance in resource, method and place; realize the limitless share of maintenance resource and method in time and space. So the efficiency and quality of fault vehicle diagnosis and maintenance in carry are improved availably.

2. THE PRINCIPLE OF CARRY VEHICLE REMOTE FAULT DIAGNOSIS

The logistics carry vehicle may produce various information on the operating, when the function of carry vehicle become badly, the corresponding abnormity information will appear going with fault. the estate signal of vehicle would be change if the carry vehicle occurs abnormity and fault, the mechanical signal concerned with vehicle state would produced such as abnormity vibrancy, noise，temperature raise, oil consume increasing ,and the chemic signal would change at the course of machine become badly as wear and tear particulate, oil and liquid , chemic component of gas(Xianhua,2005).therefore, how to capture the estate signal speediness and exactly is the essential method for making fault survey and diagnosis to logistics carry vehicle.

There is an on-line detection system in the remote fault maintenance support system for logistic carry vehicle, this system detect fault character signal using embedded computer system with detection instrument, and capture sensitive carry vehicle's situation signal and fault information, then send to the remote maintenance service center by wireless communication(Ruili et al,2009). The information is disposed, and analyzed in the remote maintenance service center. The carry vehicle's situation can be diagnosed in time, and can be evaluated and identified .Trouble vehicle can be remotely received maintenance technique assistance using technology resource in web server.

The process of remote fault diagnose is shown in Fig 1.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.94

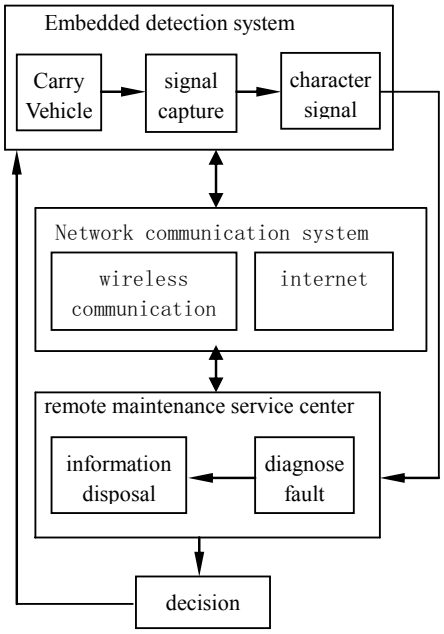


Fig 1 the process of remote fault diagnose

3 THE CORE FUNCTIONS OF REMOTE FAULT MAINTENANCE SUPPORT SYSTEM FOR LOGISTIC CARRY VEHICLE

The system of remote fault maintenance support system for logistic carry vehicle takes GPS/GIS/GSM, INTERNET, as well as information network technology as support.

The system is established through effective fusion of advanced information technology, data communication technology, automatic control technology, remote detection, vehicle theory, artificial intelligence, as well as the information processing technology, and is applied in the fault diagnosis and maintenance support for logistic carry vehicle. It has the characteristic of information, integration and intelligent. It can display function in wide range, and is a real-time, accurate, highly effective intelligence diagnosis and maintenance management system. The core functions of remote fault maintenance support system are shown as follows.

(1) The real time detection for the carry vehicle situation

The fault detection of carry vehicle is that measure estate signal concerned with the carry vehicle's situation. The estate signal is only carrier of vehicle's fault information, and only according as diagnosis. So that is very importance for capturing estate signal in time, and accurately. The estate signal is captured by sensor, or else detection instrument. The real time detection on-line system is established through effective integration of fault diagnosis technology, computer technology, and build in carry vehicle. it can collection data of vehicle situation, and capture character information from the estate signal concerned with the carry vehicle.

(2) The remote communications and position of logistic carry vehicle

In order to obtain estate signal and fault information, and confirm position of carry vehicle, as well as the carry vehicle would communicate real time with maintenance service center, this remote fault maintenance support system use GSM module as remote wireless communication equipment. It can realize real time detection to carry vehicle and information transmission. At the same time, this system uses GPS/GIS as vehicle position.

(3) The remote fault diagnosis and maintenance support

The remote fault maintenance support system uses GSM module for receiving information from carry vehicle, analyze the received information, estimate the vehicle situation, diagnose fault if the vehicle situation is abnormity, and bring forward maintenance advice.

The work principle of Remote fault Maintenance support system based on network technology is shown in Fig 2.

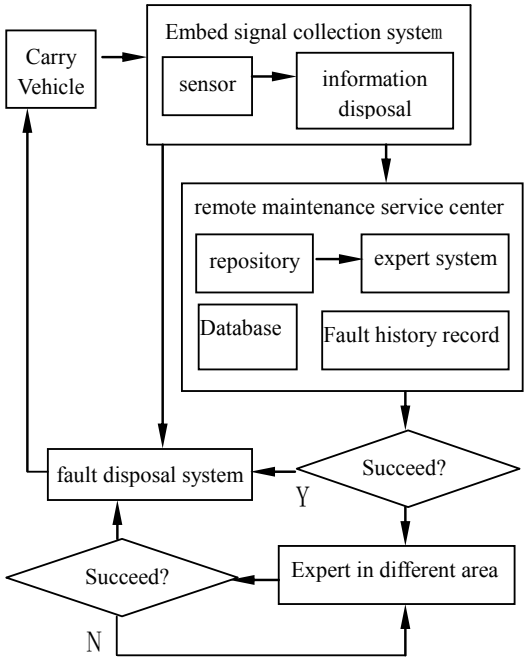


Fig 2 the work principle of Remote fault diagnose and Maintenance

4 THE FRAMEWORK OF THE REMOTE FAULT MAINTENANCE SUPPORT SYSTEM OF LOGISTICS CARRY VEHICLE BASED ON NETWORK

The framework of The Remote Fault Maintenance Support System of Logistics Carry Vehicle based on network is shown in Figure 3, which includes four main parts.

(1) The embed real time detection and diagnosis system: include embed real time detection module, vehicle situation analyze and fault diagnosis, wireless communication module etc.

(2) The remote maintenance service support system: include fault diagnosis expert system consisting of repository, consequences machine, and database; inquire about resource information; fault history record etc.

(3) The network support environment: Software and hardware support environment consist of GPS/GIS/GSM, INTERNET technology, web server.

(4) The network service resource: expert and maintain net in different area.

5 THE CHARACTERISTIC OF THE REMOTE FAULT MAINTENANCE SUPPORT SYSTEM

(1) This system can detect remote logistic carry vehicle, control vehicle's situation real time, and provide technology ensure for vehicle operating reliably.
(2) This system is propitious to implement the cooperation serve of logistic carry, improve accuracy and reliability of fault diagnosis, and guarantee logistic carry in time.
(3) This system has good expansibility, is favorable for data accumulate and sharing resource.

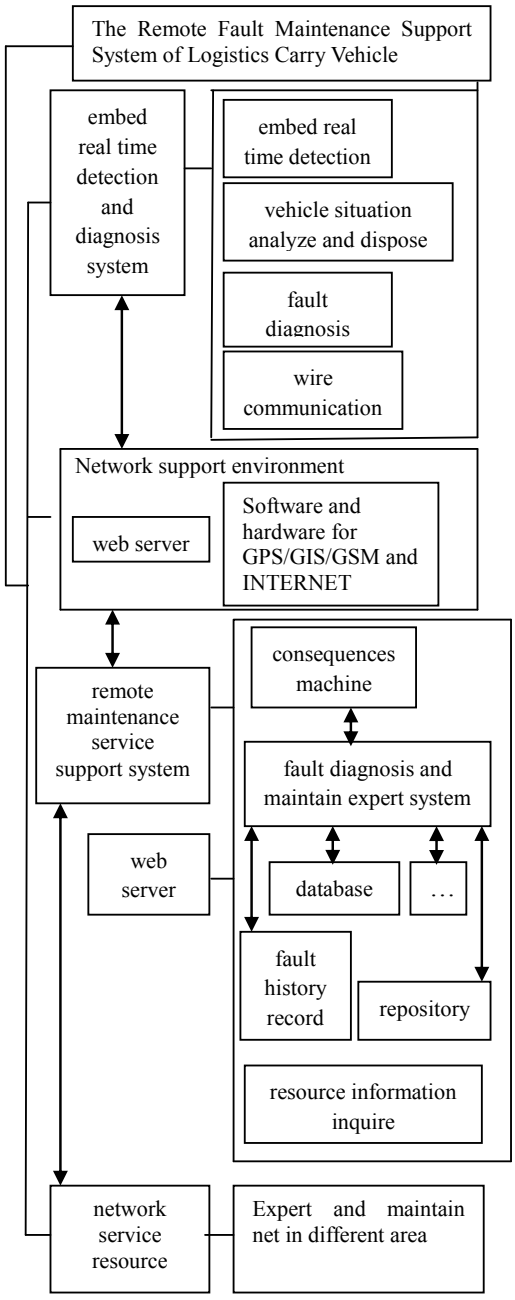


Fig 3 the framework of The Remote Fault Maintenance Support System of Logistics Carry Vehicle

6 CONCLUSIONS

In order to realize rapid carry, high efficiency, and low cost in logistics carry. Some techniques including embedded computer, wireless communication, network communication and database are used to build remote faulty maintenance support system of Logistics carry Vehicle, in this system, logistic carry vehicle's situation can be detected in time, and can be connected to the center of remote diagnosis and maintenance by wireless communication. Trouble vehicle can be remotely diagnosed and received maintenance technique assistance. The efficiency and feasibility for reducing the time of faulty diagnosis and maintenance, sharing the information resource, improving the carry vehicle availability and credibility and reducing repaired cost are proved by application.

REFERENCE

[1]Xianhua,zhao.(2005). "Sketch of remote fault diagnosis and maintenance system for constructionmachinery."architecture mechanization，3.34-36.
[2]Ruili,zeng,and yunkui,xiao(2009)."Research on automobile remote faulty diagnosis and maintenance system." electron measure technique, 7.129-131.
[3]Shufeng Wang,Junyou Zhang, Qun Yu. The Test of Vehicle Operation Stability applied ADAMS. Journal of China Agriculture University. 2001（6）：81-84.

A Way to Detect Computer Trojan Based on DLL Preemptive Injection

Guo Yucheng, Wu Peng, Lin Juwei, Guo Qingping
Department of Computer Science
Wuhan University of Technology
Wuhan, Hubei
ycheng.g@gmail.com; ejoywx@163.com; L01_ujs@163.com; qpguo@whut.edu.cn

***Abstract*: Computer Trojan has been threatening the information security of computer systems since its birth. While many companies offer many solutions to deal with computer Trojan, users can also buy a wide range of anti-virus and anti-Trojan software on the market so far, the information security software, facing to Trojan horse's attack, has been in the embarrassing position where they had to defend it passively. In this paper, in the Windows X86 platforms, based on the DLL Preemptive Injection, we introduce boldly a method against computer Trojans, and implement a downloader Trojan detection model.**

Keywords: Computer Trojan; Information Security; Windows X86; DLL Preemptive Injection

I. INTRODUCTION

Computer Trojan has been threatening the security of computer systems since its birth. While many companies offer many solutions to deal with computer Trojan, users can also buy a wide range of anti-virus and anti-Trojan software on the market, so far, the information security software, facing to Trojan horse's attack, has been in the embarrassing position where they had to defend passively: On the one hand, they can be efficiently detect and remove the known Trojans; on the other hand, the new kind of unknown Trojan, they are powerless.

In recent years, in the Windows operating system, the emergence of Kaspersky, ESET, McAfee, etc. anti-virus software and other outstanding at the same time, there have been 'Panda', 'AV Terminator', 'robot', etc. unconventional Trojans. In fact, these techniques used by Trojans are not wiser than those anti-Trojan software, only the sword easy road, just used what a number of information security software vendors do not want to analyze and study. Event some Trojans directly copied these companies anti-virus, anti-Trojan technology.

In the new kind of known Trojans, some parasitized their executable code into the modules of the normal programs, some divided the executable code itself into multiple relatively safe modules, and some use Direct-Disk-Access, Shellcode, Rootkit and other technologies to protect themselves and bypass the information security software so that anti-virus and anti-Trojan software are able to do nothing.

Face to new kind of computer Trojans, the help which the information security software offers computer users is very limited. Therefore, when we are to design a solution against Trojan, we need to analyze and study the characteristics of Trojans, so that active defense - DLL Preemptive Injection can satisfy this need very well.

II. ABOUT DLL PREEMPTIVE INJECTION

Before talking about the technology of DLL Preemptive Injection, DLL Injection must be mentioned for DLL Preemptive Injection is from DLL Injection.

In computer programming, DLL injection is a technique used to run code within the address space of another process by forcing it to load a dynamic-link library. DLL injection is often used by third-party developers to influence the behavior of a program in a way its authors did not anticipate or intend. For example, the injected code could trap system function calls, or read the contents of password textboxes, which cannot be done the usual way [1].

DLL Preemptive Injection is an enhanced version of DLL Injection. It is that, when the operating system is loading the DLL, brutally Cutting off the normal operating procedures, priority loading the specially designed DLL which achieves some kind of functions, and then restore the previous normal operating procedures - This is similar to the Operating system interruption mechanism: Forcibly intercept the normal execution flow of the processor, make the processor jump to the point where there is the high priority sequence of instructions until these sequence to be executed, then return to the original operational procedures.

In the Windows platform, when the operating system creates a process, the operating system must first load the modules such as DLL and ocx that the process needed, and some other data into the process space, then that process is considered as a system object can be executed. When the operating system loads a DLL, the DLL must also be mapped into the target process space. If a DLL is not loaded into the process space, it is not really a valid executable module. When a process obtains the required modules and data, it must call the appropriate API to create threads and start the threads, then it is regarded as a living process [2]. Thus, DLL Preemptive Injection is achieved in a variety of ways, such as Hooking NtCreateThread/NtResumeThread and Hooking NtMapViewOfSection and so on [3]. As shown in Figure 1 and Figure 2, they are the schematic diagrams of DLL Preemptive Injection based on hooking NtCreateThread and hooking NtMapViewOfSection.

As noted above, from the perspective of the operating system, DLL Preemptive Injection is not a secure and stable technique because it has to change the operating system's kernel, but it is a powerful means of the information security

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.18

technology for it can achieve the effect of fighting fire with fire.

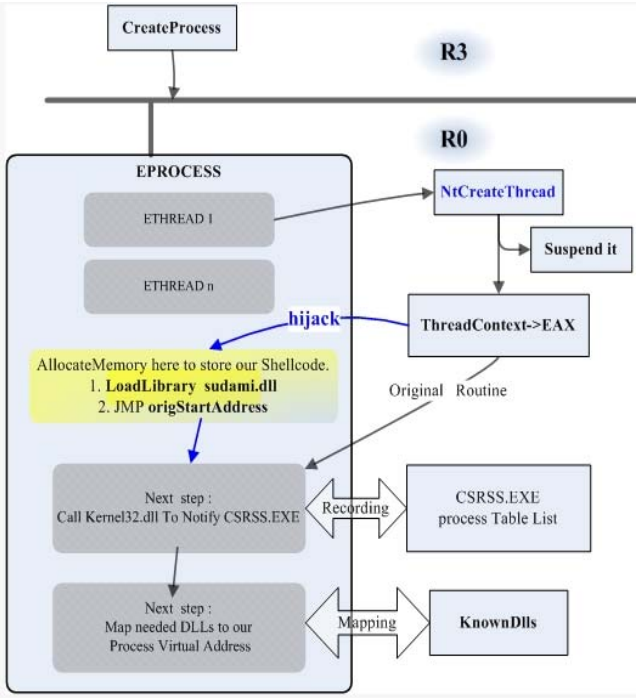


Figure 1. DLL Preemptive Injection based on Hooking NtCreateThread

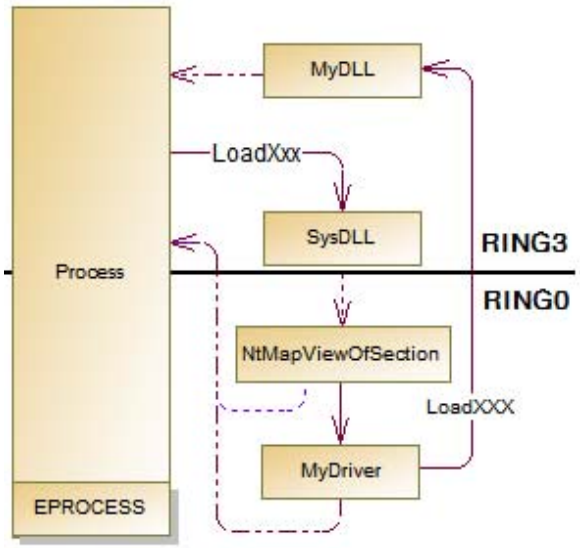


Figure 2. DLL Preemptive Injection based on Hooking NtMapViewOfSection

III. DLL PREEMPTIVE INJECTION 'S APPLICATION

This section is divided into three parts, namely, the definition of behavioral characteristics of the new kinds of Trojan, the new Trojan defense model, and the design and implementation of the downloader Trojan detection engine.

A. Trojan's Behavioral Characteristics

The author has been deeply studying the new kind of computer Trojan and surprisedly finds that their behavior has the following characteristics [4].

1) Trojan must depend on the APIs that the operating system provides, and these API are in the operating system's KnownDLLs.

2) The Trojan runs firstly, its memory image code is relatively safe. Some Trojans contain malicious instruction modules, which have to be loaded into Trojan's process space.

3) Trojan's malicious modules sometimes are in the resource section of the executable program; or the Trojan program, because of specific needs, downloads malicious modules with specific function from the remote host - that is the downloader Trojan.

4) The Trojan runs, through a number of tricks to avoid the detection of information security software, then load those malicious modules.

5) Trojan's various modules are loaded completely, the Trojan formally implements its established business.

6) Trojan does not have the mechanism to detect its own security so that it is relatively easy to kill itself.

B. The Trojan Defense Model

According to the preceding analysis, we can design a model against Trojan as Figure 3.

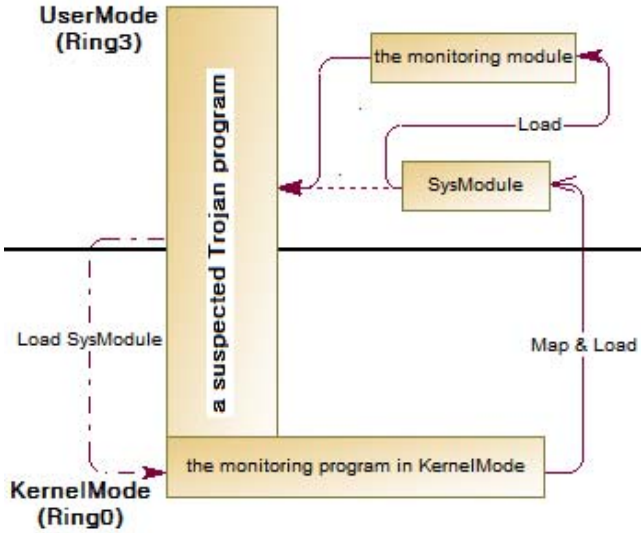


Figure 3. A Trojan Defense Model

During a suspected Trojan successfully runs, the monitoring program in the kernel of operating system, will intercept the access to some of the APIs in necessary modules. At the moment of the suspected Trojan program to call some specific APIs, the monitoring program in Kernel-mode preemptively injects a monitoring module working in User-mode into the suspected Trojan's process space, then let the injected module run.

The monitoring module in User-mode process works in the suspected Trojan to collect the concerned data so as to know where the current process is safe or not - if it is not

safe, in other words, it is likely to be an unknown Trojan, and then the monitoring module triggers "suicide" instructions to have the current process to suicide.

In this way, the aim is achieved to defend the information security of the computer system against Trojan.

C. The Downloader Trojan Detection Engine

The downloader Trojan detection engine consists of two sub-modules, the DLL Preemptive Injection engine in Kernel-Mode and the downloader Trojan detection module in User-Mode, as shown in Figure 4. The former's work is to intercept the operating system loading UrlMon.dll, then inject the latter which typically is a DLL into the process space. The latter analyzes the behavior of the host process to call APIs.

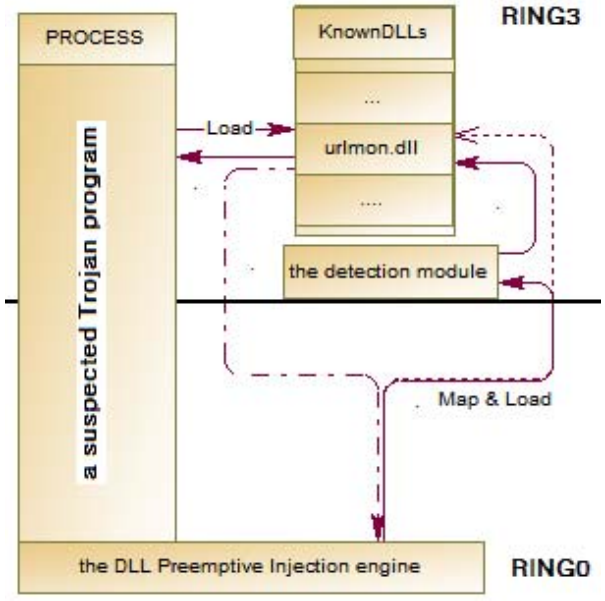


Figure 4. The Downloader Trojan Detection Engine

1) The DLL Preemptive Injection engine in Kernel mode is relatively complex to be designed and implemented, and contains three sub-modules, whose functions are to intercept the target DLLs' loading, construct the shellcode to load the monitoring module which works in User-mode, inject the shellcode into the target process space and enable it run.

As shown in the following code, it is a standard shellcode template, which can load the DLL module in User-mode after the appropriate amendments and adjustments [5].

```
unsigned char data [54] = {
    0x60, 0x6A, 0x65, 0x68, 0x6F, 0x6E, 0x45, 0x79, 0x68,
0x55, 0x72, 0x6C, 0x4D, 0x8B, 0xC4, 0x50,
    0xB8, 0xC6, 0x1D, 0x80, 0x7C, 0xFF, 0xD0, 0x58,
0x58, 0x58, 0x61, 0x90, 0x90, 0x90, 0x90, 0x90,
    0x90, 0x90, 0x90, 0x90, 0x90, 0x90, 0x90, 0xB8, 0x44,
0x43, 0x42, 0x41, 0x83, 0xC0, 0x05, 0xFF,
    0xE0, 0xCC, 0xCC, 0x90, 0x90, 0x90};
```

It is easy to implement intercepting the target DLLs' loading, so here is the detail description and implementation to inject the shellcode into the target process space and enable it run.

a) The System calls NtMapViewOfSection to map and load modules, intercept it and make it jump into the proxy function of NtMapViewOfSection , later return to the original NtMapViewOfSection routine to continue.

b) In the proxy function, based on SectionObject which is one of NtMapViewOfSection's parameters, analyze whether the current loaded module is the module needed to be monitored - urlmon.dll, if not, let it go simply; Otherwise, call ZwAllocateVirtualMemory / ZwWriteVirtualMemory and other system kernel API routine to inject the shellcode into the process space which is correspond with one of NtMapViewOfSection's parameters, ProcessHandle.

c) Modify the instructions in the entry point of the memory image of urlmon.dll to achieve firstly executing shellcode and then returning to execute the instructions in urlmon.dll. Obviously, the memory image of urlmon.dll is not easy to be directly rewrite, it is necessary to call ZwProtectVirtualMemory routine to change the memory properties in the very position, so that where keep in the writable state. Unless to do so, or the goal is never completed.

2) The downloader Trojan detection module in User mode is not complicated to be designed and implemented, and there are a variety of ways achieve that, so this article only briefly describes it. Its function is to hook a number of API series that the Trojan likes to call, such as URLDownloadToFile / URLDownloadToCacheFile, ShellExecuteEx / WinExec / LoadLibrary and OpenProcess / WriteProcessMemory / CreateRemoteThread etc., by analyzing the behavior and features of the host process to call the API, in general, This module can know whether the current process is a Computer Trojan or not. If so, the hosting process wills suicide so as in a timely manner to maintain the information security of the current system.

IV. EXPERIMENTS

In this section, the author will use a known method of bypassing Kaspersky anti-virus software, as the following source code [5], give a test to the Downloader Trojan Detection Engine described in the previous.

```
#include <Windows.h>
#include <urlmon.h>
#include <Shellapi.h>
int APIENTRY WinMain( IN HINSTANCE hInstance,
IN HINSTANCE hPrevInstance, IN LPSTR lpCmdLine, IN
int nShowCmd )
{//IDE is MSVC6.
        HANDLE
hEvent=CreateEvent(NULL,NULL,NULL,"ByPass");
        if(0==GetLastError())
        {
                STARTUPINFO             si;
                PROCESS_INFOMATION      pi;
```

```
            CHAR
        szPath[MAX_PATH];
            ZeroMemory(&si,sizeof(si));
            ZeroMemory(&pi,sizeof(pi));
            si.cb=sizeof(si);
        GetModuleFileName(NULL,szPath,MAX_PATH);
        CreateProcess(szPath,NULL,NULL,NULL,FALSE
,0,NULL,NULL,&si,&pi);
            for(int i=0;i<1000000000;i++){_asm nop;}
            CloseHandle(hEvent);
            return 0;
        }
        CHAR szFileName[MAX_PATH] ={0};
        URLDownloadToCacheFile(NULL,"file://c:\\wind
ows\\notepad.exe",szFileName,MAX_PATH,0,NULL);
        ShellExecute(0,"open",szFileName,NULL,NULL,S
W_SHOW);
        return 0;
    }
```

In the author's PC, the anti-virus software is ESET's EAV4.2 which is one of the Global TOP 10 [7]. The downloader Trojan is from the source code given earlier, successfully escapes from EAV's monitor but is caught by the downloader Trojan Detection Engine.

Of course, this result is not powerfully convincing, but at least some extent, proves the anti-Trojan method is feasible and effective.

V. FURTURE TO THE RESEARCH

Determining whether a program is a Trojan or not just according to the behavior and features of its calling APIs, it is unwise, the ruling mechanism based on the program's instruction stream is indispensable, in other words, the anti-Trojan model this article describes is lack of an engine to analyze the program's instruction stream. In the near future, the author plans to develop an instruction flow analysis engine on the Libclamav.

VI. CONCLUSION

Today DLL Injection is in the abuse, it is undeniable, as one of the evil hacker techniques, it is not intrinsically different from DLL Preemptive Injection..

The key point of DLL Preemptive Injection is preemptive, if it is integrated with the Trojan detection techniques, it is able to eliminate the threat in the bud. Hybridizing the anti-Trojan and DLL Preemptive Injection, a powerful anti-Trojan capability will be obtained. The monitoring module Injected into the Trojan process space, triggers the instructions to suicide instructions, the Trojan process will fall apart immediately and die.

ACKNOWLEDGMENT

The author would like to give a big thank to his tutor, Prof. Guo, he could not successfully finish the paper without his tutor's support. Thanks to Li Juwei, in his help, I can succeed in completing the downloader Trojan detection engine based on DLL Preemptive Injection. Finally, the author likes to thank the Chinese Academy of Sciences, the technology that this paper relates to and its verification work are never implemented without the equipments from the ICT. What is more, This work is supported by the Innovation research fund for graduate students free to explore innovative projects of Wuhan University of Technology2010-ZY-JS-006).

REFERENCES

[1] *Wikipedia, "DLL injection".* *http://en.wikipedia.org/wiki/DLL_injection*

[2] Pan Aiming, "Principle and Implementation of Windows kernel". Publishing House of Electronics Industry,2010.

[3] Sudami, "N kinds of ideas and the implementation of DLL Injection from the kernel". http://hi.baidu.com/sudami

[4] Peter Szor, "The Art of Computer Virus Research and Defense". Machinery Industry Press, 2007.

[5] Kris Kaspersky, "Shellcoder's Programming Uncovered". A-List Publishing, 2007.

[6] cnbeta, "VB100 test results released in October". http://www.cnbeta.com/articles/124115.htm

[7] dangdang, "the Method of Bypassing Kaspersky Heuristic Virus Scanning" http://secinn.appspot.com/pstzine/read?issue=4&articleid=9

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Study on Multi-Grade Intrusion Detection Model Based on Data Mining Technology

Halqam Ablat
school of Mathematics and Information Technology
Xinjiang Education Institute,
Urumqi, Xinjiang, 830043, China
e-mail:sxhalqam@xjei.cn；halqam@tom.com

***Abstract*—Focusing on the deficiencies of conventional intrusion detection model, WenkeLee, Salvatore J. Stolfo et al. propose the intrusion detection system based on data mining technology. It solves the problem that self-adaptability of the system is poor, and the conditions of misreport or omission are also further improved. However, as far as mass data are concerned, more and more resources need to be consumed in the intrusion detection system based on data mining technology, and the detection speed gets slower and slower. In the article, the multi-grade intrusion detection model based on data mining technology is proposed, and the objective to improve the detection speed is reached.**

Key Words: ***Intrusion detection model; IDES model; data mining technology; multi-grade intrusion detection model***

Ⅰ. INTRODUCTION

A complete intrusion detection system needs to have the features of being accurate, timely, complete, expansible, fault-tolerate and etc. After data mining technology is introduced into the intrusion detection system, the intrusion detection system can find out the mode of users' abnormal behaviors effectively. With strengthened detection ability, it solves the problem that self-adaptability of the system is poor, and the conditions of misreport or omission are also further improved. However, with increased network information flow, its detection efficiency becomes lower and lower, the resources required to be consumed in the intrusion detection system gets more and more, and the detection speed gets slower and slower. The multi-grade intrusion detection model based on data mining technology can solve the problem.

Ⅱ. INTRUSION DETECTION MODEL BASED ON DATA MINING TECHNOLOGY

A. Intrusion detection model

The earliest intrusion detection model was proposed by Denning in 1986, and named IDES (The Intrusion Detection Expert System) model. Since IDES model could not detect new attack methods, Teresa Lunt et al. improved the intrusion detection model proposed by Denning in 1988, and developed an intrusion detection expert system based on the improved model. The system includes a detector of abnormal conditions and an expert system, which are used for detection of abnormal conditions and misuse respectively.

B. Intrusion detection model based on data mining technology

Most products based on IDES model adopt the misuse detection technology. The deficiencies of such intrusion detection method are mainly shown as follows: Poor self-adaptability, high misreport and omission rate, detection efficiency to be further improved, problem about operating performance. At present, in order to solve the foresaid problems, experts have proposed the method in which data mining technology is introduced into the intrusion detection system.

Data mining itself is a universal knowledge discovery technology for the purpose of extracting data information in which we are interested from mass data. The application of data mining technology to the

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.85

intrusion detection system aims to carry out intellectualized processing of mass safe audit data and extract useful information to find out intrusion detection. At present, there are mainly two directions: The first is to find out the rules and modes of intrusion and combine it with the detection methods matching to the modes; the second is to be used for detection of abnormal conditions, find out users' normal behaviors, and establish the database of users' normal behaviors.

Data mining technology is used in the intrusion detection system. Through association analysis, classification analysis and etc. in data mining, the analysis of historical data is made. It can extract users' behavioral features, summarize the rules of intrusion behaviors, and set up relatively complete rule database to carry out intrusion detection.

WenkeLee and Salvatore J. Stolfo apply data mining technology to the research field of intrusion detection for the first time. Its basic idea is to extract hidden intrusion information (It is shown in the form of concept, regulation, rule, mode and etc.) from audit data or data flow, and use these information to detect abnormal intrusion and known intrusion. The introduction of data mining technology into the intrusion detection system can also solve the problem about manual compilation of feature library, so that machine can modify the feature library in a self-adaptive way. Meanwhile, it can also improve efficiency and find out intrusion behaviors as soon as possible.

The application of data mining technology to the intrusion detection system can bring the advantage of data mining technology in handling large-quantity data into full play, find out knowledge or rules in data unknown to people, and improve efficiency and accuracy of detection accordingly.

III. MULTI-GRADE INTRUSION DETECTION MODEL

Compared to conventional intrusion detection system, the introduction of data mining technology into the intrusion detection system can effectively find out the mode of users' abnormal behaviors. With strengthened intrusion detection ability, it solves the problem that self-adaptability of the system is poor, and the conditions of misreport or omission are also further improved. However, with increased network information flow, the system resources required to be consumed in the intrusion detection system gets more and more, the detection efficiency of t he intrusion detection system becomes lower and lower, and the detection speed also gets slower and slower. In view of this, we make improvement on the intrusion detection model based on data mining technology, and propose a kind of multi-grade intrusion detection model based on data mining technology.

See Fig.1 for the multi-grade intrusion detection system model.

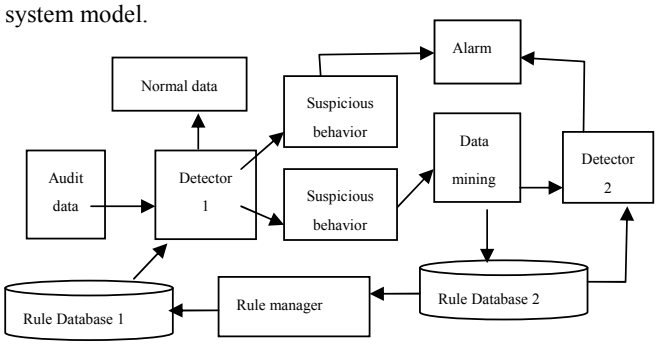


Fig 1 Multi-grade intrusion detection model

The multi-grade intrusion detection model is based on distribution cooperation technology. The whole system divides the intrusion detection procedure into two grades: Primary intrusion detection and advanced intrusion detection. The primary intrusion detection module is an intrusion detection system based on IDES model, while the advanced intrusion detection module is an intrusion detection system based on data mining technology. The primary intrusion detection module and the advanced intrusion detection module constitute the distribution system. The misuse detection technology is used in both modules for primary intrusion detection and advanced intrusion detection.

Data package from network gets rid of invalid data through preprocessing module first, sort out original data in the format which can be identified in the data mining algorithm, and form audit data accordingly. Then it is sent into the primary intrusion detection module for primary intrusion detection. Detector 1 makes it have mode matching with the rules in Rule Database 1, so that audit data can be divided into three types including data about normal behaviors, data about intrusion behaviors and data about suspicious behaviors. Among them, intrusion behaviors generally refer to various known common

attack behaviors (such as known attack behaviors utilizing system holes, attack modes released online and etc.); normal behaviors refer to users' behaviors in strict accordance with all the features of normal user mode generated by the system; suspicious behaviors refer to such kind of behaviors which are deviated from normal behaviors and are not unknown attack behaviors. They may be either normal behaviors or intrusion behaviors. If it is normal behavior, no processing needs to be made; if it is intrusion behavior, alarm should be made. Suspicious data in which intrusion behaviors may exist should be marked. By doing so, the procedure of primary intrusion detection is completed. Some common intrusion rules can be stored in Rule Database 1 in advance at the beginning. Later, some rules in Rule Database 2 can be added into Rule Database 1 through rule manager. The rule manager plays a coordinative role between the primary intrusion detection module and advanced intrusion detection module.

In the primary intrusion detection module, Detector 1 needs to filter normal data, which can reduce the data required to be processed by the advanced intrusion detection module. The processing method to filter normal data and abnormal data is different. For normal data, since users' normal behaviors are quite different, it is very hard to use limited rules to describe all the normal behaviors of users. Therefore, we can take the method of anomaly comparison to solve it. The idea of anomaly comparison is similar to the similarity comparison. In Detector 1, we take the idea to filter normal data. After a data package arrives, the probability P (%) that such kind of data appear is calculated according to known data, and then the original anomaly is calculated according to the probability P (%):

$$A(x) = -\log_2(P(x))$$

For normal data, a large number of continuous normal data of the same kind may exist. In order to prevent the value of anomaly from being highly concentrated and unbalanced, we adopt the method of relative anomaly in the system, that is, the original anomaly is divided by the highest original anomaly, and a value between 0 and 1 is obtained:

$$0 < Ar(x) = A(x) / HighestA(x) \leq 1$$

If relative anomaly does not exceed certain threshold value (According to previous experience, we set the threshold value as 0.8 in the system),the data are normal data and should be filtered. If relative anomaly is more than the threshold value, the data may be suspicious data or intrusion data; for the aggregate of suspicious data and intrusion data, we will use the misuse detection method matching to the rule to detect known data about intrusion behaviors. By doing so, intrusion data can be judged accurately, and the remaining data are suspicious data which should be processed by the advanced intrusion detection module.

In the system, we take the misuse detection technology, for the data about intrusion behaviors only account for a very small part among total data for any network application system. Thus there is only little space covered by the intrusion rule database. Such small-scale intrusion rule database is easy for maintenance and update. The main task of primary intrusion detection module is to filter intrusion behaviors and normal behaviors. No further analysis and confirmation will be made to data which are not matching. Instead, they will be processed by the advanced intrusion detection module. The advanced intrusion detection module carries out data mining processing to suspicious data and find out new intrusion rules. The newly found intrusion rules will be added to Rule Database 2, which ensures that the system can update the rule database according to the change of data online. The frequency of retraining of classifier should be also noted. According to the intrusion detection system based on data mining technology we have realized, we can set the time interval as 35 seconds.

No matter whether the misuse detection technology or the anomaly detection technology, they are based on such a hypothesis: Normal behaviors and intrusion behaviors can be distinguished according to the data provided by the review data sources. However, with the increase of the quantity of rules in the rule database, to inspect all the characteristic attributes once not only increases the data processing, but also consumes time, which will reduce the efficiency of intrusion detection. To take the multi-grade intrusion detection method and make operation between the primary intrusion detection and the advanced intrusion detection can improve the detection

speed to a great extent. Meanwhile, the intrusion rule database is divided into two, so that the detection scope is reduced half, which will be more beneficial to improve the efficiency of intrusion detection, and will not reduce the accuracy of detection at the same time.

At the beginning, the rules in Rule Database 1 and Rule Database 2 are obtained through training of data classified manually at the development stage, and these intrusion rules can only be applied to the existing network. After the intrusion detection system is put into operation, with update and development of network application, new intrusion rules will be continuously added in Rule Database 2 with the support of data mining module. Detector 1 will classify network data according to the intrusion rules in Rule Database 1, then data mining module will carry out data mining processing to suspicious data, extract new intrusion rules, and update Rule Database 2 continuously. Meanwhile, according to the conditions to use the rules in two rule databases, the rule manager makes dynamic adjustment to the rules in two rule databases, put common intrusion rules and rules corresponding to intrusion into Rule Database 1, and put those rare and advanced intrusion rules into Rule Database 2.

Thus, the main task of the primary intrusion detection module is to detect some common intrusion behaviors and filter normal behaviors. By doing so, it can reduce the data required to be processed by data mining module to a great extent. The main task of the advanced intrusion detection module is to find out new intrusion behaviors and detect some uncommon intrusion behaviors at the same time. With specific division of two modules, cooperative processing is made with the rule manager, so that the intrusion detection efficiency of the whole system is improved.

Ⅳ. Experimental Analysis Of Multi-Grade Intrusion Detection System

In the following, a detailed introduction will be given to the experimental analysis procedure of the system that we have realized and improved.

A. Experimental analysis of the intrusion detection system

We take the data mining software Weka to carry out experimental analysis of the intrusion detection system. Explorer is the basic and the most common function of Weka. In our experimental analysis, only Explorer function is used.

In order that the experimental analysis results are true and can reflect actual conditions, the network data in the experiment come from Lawrence Berkeley National Laboratory. The following attributes are included: timestamp,duration,protocol,bytes_sent_orginator,bytes_sent_responder,local_host,remote_host,state,flags. They represent the marks of starting time of connection, duration, service type, number of sent bytes, number of received bytes, source host, destination host, end condition of connection, internal connection or external connection. Among them, “state” refers to the state when connection is ended. Various different abnormal behaviors can be judged through it. The contents shown by “state” are stated in Table Ⅰ.

TABLE I. TABLE OF THE STATE FOR TCP CONNECTION END

State	Description
SF	Normal completion of TCP dialogue
REJ	SYN is sent from one side, RST is sent from the other side as response
SO	SYN is sent from one side, no response is given from the other side
S1	After connection is established (SYN exchange is completed), there is no further data exchange
S2	After connection is established, the sender closes the connection.
S3	After connection is established, the receiver closes the connection.
S4	SYNack is received, but corresponding SYN is not received.
RSTOSn	Connection is reset by the sender when connection is in the state n.
RSTRSn	Connection is reset by the receiver when connection is in the state n.
SS SYN	The closed connection is received.
SH	Connection in the state 0 is closed before SYNack is received.
SHR	Connection in the state 4 is closed before initial SYN is received.

OOS1	SYNack and SYN are not matching.
OOS2	SYN is retransmitted with different serial No. at the beginning.

There are 782281 network connection records in total in the experimental data set [32][33][34]. Due to the restriction of experimental conditions, we select 65535 connection records as training set, and select extra 65535 connection records as test set. Two attributes of "timestamp" and "duration" here are used to judge the size of sliding windows when candidate features are constituted. "flags" is the mark to judge whether the connection is internal connection or external connection. The remaining six features are used for the classification of classifiers. In order to simplify experimental analysis, we do not use the constituted five candidate features for the classification of classifiers.

The intrusion detection system realized in the article mainly gives consideration to improving the detection speed of the system. Meanwhile, there is very little relevance among the attributes in network connection records, which meets the condition about independence of category. Therefore, we select Naïve Bayesian Algorithm to constitute classifier. The realized intrusion detection system is carried out focusing on Naïve Bayesian Algorithm at the second stage for preprocessing, that is, the discretization of continuous attributes is carried out.

The test results of the intrusion detection system based on data mining technology are stated as follows:

Number of correct classification records 60727 92.6635%

Number of wrong classification records 4808 7.3365%

Time for the establishment of model 0.06s

See Table Ⅱ for the specific classification results with the attribute "state" as the target classification attribute

TABLE II. TABLE OF THE TEST RESULTS OF THE INTRUSION DETECTION SYSTEM

TP Rate	FP Rate	Precision	Recall	Class
0.966	0.259	0.957	0.966	SF
0.017	0	1	0.017	RSTRS3
0.557	0.002	0.894	0.557	REJ
0.845	0.016	0.81	0.845	S0
0.637	0.014	0.587	0.637	RSTOS2
0	0	0	0	S1
0	0	0	0	RSTRS1
0.466	0	1	0.466	RSTOS1
0	0	0	0	RSTOS4
0	0	0	0	OOS1
0	0	0	0	S2
0.872	0.006	0.391	0.872	RSTOS3
0	0	0	0	SS
0	0	0	0	SHR
0.527	0	0.65	0.527	RSTRS2
0	0	0	0	S3
0	0	0	0	RSTOS0
0	0	0	0	RSTRS4
0	0	0	0	SH
0	0	0	0	OOS2

In which: TP Rate refers to the proportion of the number of samples for correct classification of target samples.

FP Rate refers to the proportion of the number of wrong samples in the classification of non-target samples into target samples

Precision refers to the detection rate of abnormal behaviors

Recall refers to the identification rate of anomaly made by the system

Class refers to the attribute of classification targets

From the test results, we can see that the proportion of the number of correct classification records reaches 92.6635%. Such proportion is relatively high. It also shows that it is feasible that we take Naïve Bayesian Algorithm to constitute classifier. Besides, the time for the establishment of models is 0.06s, which is relatively short. In the specific classification result with "state" as the attribute of classification targets, the proportion of the number of samples in correct classification of SF category reaches 96.6%, the detection rate of abnormal behaviors reaches 95.7%, the identification rate of anomaly made by the system reaches 96.6%. All the proportions are relatively high, while such proportions in some categories are relatively low. It is mainly caused by the difference of the number of records. We take 65535 record data in total, while there are 56022 data of SF category. It also shows the advantages of Naïve Bayesian

Algorithm in handling mass data.

B. *Experimental analysis of multi-grade intrusion detection system*

When data required to be processed are increased, we have improved the realized system and proposed the multi-grade intrusion detection model. When the multi-grade intrusion detection technology is used in the intrusion detection system, the primary intrusion detection module mainly filters normal data and common intrusion data, just marks suspicious data, calculates corresponding probability according to the frequency that data appear in data concentration, calculates the original anomaly of every piece of datum according to the probability value, and then calculates relative anomaly of every piece of datum. The threshold value is set as 0.8, normal data are filtered, and some common intrusion data are removed (Some data in the experimental data are known intrusion data). Thus, there are 9508 data left among 65535data, which makes the amount of data processing reduced to a great extent. The 9508 data will be taken as the training set to establish classification model, and 9508 data will be selected from the test set to be a new test set for test. The test result is stated as follows:

Number of correct classification records 8466 89.0408%

Number of wrong classification records 1042 10.9592%

Time for the establishment of model 0.0001s

See TableIII for the specific classification results with the attribute "state" as the classification target attribute

TABLE III. TABLE OF TEST RESULTS OF THE MULTI-GRADE INTRUSION DETECTION SYSTEM

TP Rate	FP Rate	Precision	Recall	Class
0	0	0	0	SHR
0	0	0	0	SS
0.842	0.002	0.41	0.842	OOS1
0	0	0	0	OOS2
0.861	0.021	0.891	0.861	REJ
0.069	0	0.5	0.069	RSTOS0
0.568	0.006	0.766	0.568	RSTOS1
0.894	0.03	0.885	0.894	RSTOS2
0.914	0.005	0.847	0.914	RSTOS3
0.041	0	0.667	0.041	RSTOS4
0.148	0.001	0.409	0.148	RSTRS1
0.784	0.006	0.5	0.784	RSTRS2
0.724	0.005	0.797	0.724	RSTRS3
0	0	0	0	RSTRS4
0.955	0.087	0.919	0.955	S0
0.892	0	0.983	0.892	S1
0.737	0.001	0.737	0.737	S2
0	0	0	0	S3

Since the data involved in data mining processing are reduced to a great extent, the time for the establishment of classification model is very short, some time can be saved to have a large number of data processed in actual system. The time for the processing of data mining is an important parameter of the intrusion detection system based on data mining technology, thus the detection speed of the intrusion detection system based on data mining technology can be improved. Meanwhile, we can see that the proportion of the number of correct classification records in source data is 89.0408%, which is reduced to some extent. It is mainly attributed to the decrease of data. The Naïve Bayesian Algorithm is more advantageous in the processing of mass data, while there is a great number of data required to be processed in actual system. It is also the reason why we select the Naïve Bayesian classification algorithm to constitute classifiers. From the foresaid comparison and analysis, we can see that the intrusion detection technology proposed in the article reaches the aim to improve detection speed.

V. CONCLUDING REMARKS

In the article, a brief introduction is given to the intrusion detection model proposed by Denning first, and then the intrusion detection model based on data mining technology which is proposed by WenkeLee is introduced. Focusing on the problem that the detection speed is reduced when mass data are processed in the intrusion detection system based on the mode, the multi-grade intrusion detection model is proposed in the article, and the realized system is improved based on the model, so that the multi-grade intrusion detection system based on data mining technology is realized.

REFERENCES

[1] Tang Zhengjun, Li Jianhua, The Intrusion Detection Technology, Beijing: Tsinghua University Press, 2004．6～7

[2] Yang Yongchuan, Li Dongjing, Information Safety, Beijing: Tsinghua University Press, Chinese People's Public Security Press, 2007.96～97

[3] Lu Yong, Cao Yang et al., The Framework of the Intrusion Detection System Based on Data Mining [J]. Journal of Wuhan University (Science Edition), 2002(1).62～63

[4] Jiawei Han, Micheline Kamber, Concept and Technology of Data Minning [M], Translated by Fan Ming et al. Beijing: Mechanical Industry Press, 2001.3～5

[5] Chen Zhibo, Data Warehouse and Data Mining, Beijing: Tsinghua University Press, 2009.152～153

[6] Sheng Siyuan, Zhan Shouyi, Shi Yaobin The Intrusion Detection System Based on Data Mining [J]. Computer Engineering, 2003(1)．156～157

[7] Jiang Yichuan, Tian Shengfeng, The Application of Data Mining Technology in the Intrusion Detection System Computer Engineering, 2001(4).21～25

[8] Empirically-Derived Analytic Models of Wide-Area TCP Connections, V. Paxson, *IEEE/ACM Transactions on Networking*, 2(4), pp. 316-336, August 1994;

[9] Growth Trends in Wide-Area TCP Connections, V. Paxson, IEEE Network, 8(4), pp. 8-17, July 1994;

[10] Wide-Area Traffic: The Failure of Poisson Modeling, V. Paxson and S. Floyd, *IEEE/ACM Transactions on Networking*, 3(3), pp. 226-244, June 1995.

The Application in the Intrusion Detection System, Computer Engineering, 2001(4).21～25

A Novel Information Hiding Algorithm Based on Page Object of PDF Document

Huang Simin[1] Sun Xingming[2,1] Fu Zhangjie[1]
[1]School of Information Science and Engineering ,Hunan University
Changsha,China
[2]Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology
Nanjing,China
e-mail:sunnudt@163.com, xiaoyu520820@163.com

***Abstract*—After analyzing the PDF file and PS file, a novel information-hiding algorithm based on the Page Object of PDF file is proposed. First, a piece of PS code will be finished in PS file to create a new Page Object of PDF, and then the secret information is camouflaged to form the attribute value of the new object. The embedding and extracting of information in PDF document have been implemented.Theoretical analysis and experimental results show that this algorithm does not affect the output of the readers, the editors and the printers, and capacity and robustness are superior to other similar methods.**

Keywords-Information-hiding;PageObject; PDF file; PS file; secret information

I. INTRODUCTION

A. Background

The rapid development of multimedia technology and the popularity of the Internet have brought great convenience to people's life. People can exchange information and share network resource through the Internet,and it comes the serious problem of Information Security .How to protect the security of the information has become more and more important.Nowadays, structured-text are wildly used by people for exchange of information,such as MS office word,HTML,Adobe PDF,etc.which play important roles in network information delivers in our daily life .PDF document has become one of the important ways which are used for information deliver because of openness, portability, safety, efficiency and cross-platform.Therefore, information hiding based on PDF document has become a very hot area in the research of information security,which has a wide range of application prospects.

B. Situation of Research

The general model of hiding data in other data can be described as follows. The embedded data is the message that one wishes to send secretly. It is usually hidden in an innocuous message referred to as a cover-text,or cover-image or cover-audio as appropriate, producing the stego-text or other stego-object[1]. As a structured-text, PDF document is also a cover-text which is used to embed secret information in.Unlike an image or audio, a structured-text has little redundancy information for secret communication[2]. A Steganography system named PDFStego based on PDF is presented by Shangping Zhong etc,which embed the data information in the the right margin ,if it is not ragged[3]. After Arnold transformation , a binary image is embedded in PDF document by Chunyan Gu etc. Which embed the data of the image in character spacing[4]. There is a algorithm which is proposed by Youji Liu etc. who designed a system for information-hiding in PDF document,in this algorithm,a redundancy object will be construct in PDF file to embed secret information[5].In [6][7][8][9][10],there are some steganographic methods are proposed almost by varying the line or word or character spacing or by varying certain character features slightly.These steganographic methods have some weaknesses, for example: A. The embedding capacity is small; B. The robustness is weak; C. The security is low, etc.

II. THE WORKING MECHANISM AND PRINCIPLE OF PDF AND POSTSCRIPT

In the PDF documents, PDF regards every text character as the image model, while each character corresponds to a specific graphics. The character graphics are categorized by typeface, which contains a particular set of character codes and graphics, and defines the corresponding relationship between character codes and the graphics. When the program runs, each typeface will generate a dictionary with a name. The program firstly finds the dictionary which presents that character, then gets the character code and finally obtains the character graphics according to it.

Postscript which is a programming language, abbreviated to PS, is suitable for printing images and text (paper, film or CRT materials all will do). PostScript is a kind of language describing the pages, whose main purpose is to provide a convenient way to describe the images for a relatively independent device[11]. Such device independence means not to take a specific device characteristic as a reference to describe the image, so there is no need to make any modifications to apply to any PostScript printers.

The contents of PDF documents and the format of the page can be obtained by the corresponding PS files, which are the programs written in the language of Postscript. PDF and PS can interchange to each other through Adobe Acrobat Professional and Acrobat Distiller. To simplify the process of page description and achieve the random orientation of documents, PDF employs a strictly-defined structure, while PS can provides a reasonably readable programming language structure. Therefore, PS files are

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.9

more likely to achieve the encoding and decoding of the watermark and insert the watermark to PDF documents by modifying the output format of PS documents.

III. FEASIBILITY ANALYSIS AND PROCESS OF ALGORITHM

A. Analysis of Features

PDF represents text and graphics using the Adobe 2D vector-graph positioning model which is the same model as the one used by PostScript language[11]. Like a Postscript language program, a PDF page description draws a page by placing text and graphics on selected areas,we called these areas "Page Object".

PDF defined these Page Object's positions in terms of pairs of coordinates on the Cartesian plane by some operators and a coordinate pair is a pair of real numbers x and y.These operators are "Path Construction Operators"[11]. For example "mo", it is a operator which is used to begin a new subpath by moving the current point to coordinates(x, y), omitting any connecting line segment."li" is also a kind of "Path Construction Operators", it is used to append a straight line segment from the current point to the point(x, y). The new current point is (x, y). There are some other operators such as "C","V","R",etc.,which are used to operator a cubic Bezier curve.

The "mo" moves a Page object from a position to another position,so the operand of "mo" can be taken as the coordinates of the Page Object. An example of the use of "mo" is shown in Fig.1.There are some characters in these two Page Objects, and PDF defined the operator "xsh" to control the position of Page Object.For example,if it is 6.12345 mn from left side of "B" to left side of "s",Postscript language will describe it like [6.12345 0].Of course, this is only a hypothesis that there is no character behind "s". A page-description example of a part of a PDF text is shown in Fig.2.

B. Feasibility of the Algorithm

Write a PS code,and the content of the object is "this is a test".

The code is as follows:

8.22345 9.23345 mo

(this is a test)

[8.23276 8.23276 8.23276 8.23276 8.23276 8.23276 8.23276 8.23276 8.23276 8.23276 8.23276 8.23276 8.23276 8.23276 0] xsh

Add the code to PS file, the new PDF file is shown in Fig. 3.

If the "this is a test" is replaced with "spaces",and modify the coordinates to 0.00xxx 0.00xxx ,the page object will be moved to the position which is next to the origin, and then the secret information is camouflaged to form the attribute value of the space . This method will not only meet the transparency,but also robust when tampering attacks,and the capacity is unlimited in theory.

C. Specific Algorithm

1) Pretreatment of secret information

First,a new object need to be constructed, and then the secret information is camouflaged to form the attribute value of the space which is in the Page Object. It is worth mentioning that only two decimal places for Embedding position can be chosen,which it's the rule of coordinate precision in PDF. And this rule will not be discussed in this paper.The determine of the number of spaces will be discussed in following section.

a) Number of spaces for one line

Assume that the number of the data information in one line is N.If the N is a even number, the number of space which need to construct is (N/2+1+1), If N is a odd number, the number of space which need to construct is {(N+1)/2+1+1}. we use A to represents the number of spaces.

- If N is a even number

$$A = N / 2 + 2 \quad (1)$$

- If N is a odd number

$$T = \{(N + 1)/ 2 + 2\} \quad (2)$$

b) Number of spaces for all lines

Assume that there are n lines data information, so,there should be $A_1,A_2,A_3.....A_N$.In this algorithm,we also need a attribute value of the space to record the number of lines,therefore the number of the spaces is T.

$$T = A_1 + A_2 + A_3 + A_4 + ... + A_N + 1 \quad (3)$$

2) Embedding and extracting

a) Embedding Algorithm

- Input: cover-text,data information
- Output: stego-text
- Step1: Scan the cover-text and create a page object. Let the marked information be "0.00xxx 0.00xxx mo"
- Step2: calculate the "T","N" ,"A".
- Step3: Reference to the rules of 3.2.1 to construct the spaces, the 1st,2nd and the 3rd position were used to record the "T","N_1" and "A_1" respectively.
- Step4: Divide the data which is in the 1st line into groups by every two data, if N_1 is a odd number, the last group will be added a "0" in it.
- Step5: Embedding the encrypted information in the two decimal places of the attribute value, the other decimal and integer are random.
- Step6: Imitate the Step3,Step4,and Step5, the data information in other lines will be embedded into the page object.
- Step7: Close the cover-file, and generate a stego-file.

The flowchart of embedding is shown in Fig.4.

b) Extracting Algorithm

- Input: stego-text

- Output: cover-text
- Step1: Scan the stego-text ,find the marked information "0.00xxx 0.00xxx mo".If there is no marked information,return "Extraction failed".
- Step2: IF there is a marked information, continue to find the attribute value of spaces.
- Step3: Read the 1st three attribute value, extract the secret information in 1st line,and then extract all secret information one line by one line.
- Step4: Output the secret information and close the stego-text.
- The flowchart of extracting is as is shown as Fig.5.

IV. FEASIBILITY ANALYSIS AND PROCESS OF ALGORITHM

Selected 200 PDF documents randomly for test, the following conclusions can be obtained.

A. Robustness Test

The experimental result shows that it's robust under the general manipulations of Adobe Acrobat Professional such as addition of text content, postil, stamp, signature, background and delete text content of the PDF document.

TABLE I. STATISTICS OF ROBUSTNESS TEST

Type of attack	*Intensity of attack*	*Subhead*
Addition of postil	Arbitrary but limited	100%
Addition of stamp	Arbitrary but limited	100%
Addition of signature	Arbitrary but limited	100%
Addition of background	Arbitrary but limited	100%
Delete text content	Nearly 100%	>95%
Add text content	Unlimited	100%

B. Capacity Evaluation

This is a similar method of adding redundant objects,although the new page object will be displayed,in fact,it is "invisible ".So,once the required number of spaces by the rules is calculated, the capacity will be infinite in theory.But as a means of information hiding, the size change of the text should be considered, therefore,nearly 10% of capacity is appropriate,which is a very satisfactory capacity.

C. Transparency Evaluation

Part of cover-text is shown in Fig.6,and Part of stego-text is shown in Fig.7. There is a new Page Object in the PDF which is shown in Figure 7.First ,the "space" character is invisible in PDF document, second, the order of magnitude of the Coordinates is micron, therefore, this algorithm is transparent. Experimentation shows that, it is hardly to discover the new page object even if zoom 64 times to this stego-text.

D. Compare with other Algorithms

After the analysis of other algorithms,the simulation were implemented on these 200 PDF documents and the statistical results are shown in Tabel 2.

TABLE II. COMPARE WITH OTHER ALGORITHMS

	The algorithm of this paper	*Other similar methods[4,5,6,7]*	*Algorithm based on structure of PDF document[8]*	*Algorithm based on Syntax and semantics*
Maximum capacity	Infinite in theory	capacity<0.667% of text size	Infinite in theory	Formatdocument : capacity<0.5% Plane document : capacity<2.5%
Robustness	better	worse	fragile	worse
transparency	No Optically Changed	No Optically Changed	No Optically Changed	May affect the semantics

V. CONCLUSIONS AND FUTURE WORK

A novel steganographic technique for hiding data information in PDF text have been presented, the design of the system was finished on VC 6.0.Theoretic analysis and the experimentation show that the system is secure, experimental results show that this algorithm does not affect the output of the readers, the editors and the printers, and the capacity and robustness of this algorithm are superior to other similar methods,and it is transparent.

Although the secret information which used in this algorithm are numbers,but itcan be replaced "spaces" with Chinese characters and English characters in theory. The requirements will drive researchers to study the color of the characters to ensure the transparency of the object, therefore,how to embed the various character in PDF document is a challenge in our future work.

VI. ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China (No. 60736016, No. 60973128, No. 60973113, No. 61073191 and No. 61070196), and in part by National Basic Research Program of China (No. 2009CB326202, and No. 2010CB334706), PAPD.

VII. REFERENCES

[1] Fabien.A.P.Petitcolas,Ross.J.Anderson,Makus.G.Kuhn,"Information-Hiding--A Survey," Special Issue on Protection of Multimedia Content, vol. 87, pp. 1062-1078, July 1999.

[2] Stefan Katzenbeisser and Fabien A. P. Petitcolas, "Information Hiding Techniques for Steganography and Digital Watermarking," Artech House Publishers, December 1999.

[3] Shangping Zhong, Xueqi Chen, Tierui Chen, "Data hiding in a kind of PDF texts for secret communication," International Journal of Network Security, vol. 4, pp. 17-26,2007.

[4] Chunyan Gu,Yang Yang, "A Text Digital Watermarking Algorithm for PDF Document Based on Scrambling Technique," Journal of Foshan University, vol. 3, pp. 43-47,2009.

[5] Youji Liu,Xingming Sun,Gang Luo, "A Novel Information Hiding Algorithm Based on Structure of PDF Document.,"Computer Engineering, vol. 32, pp. 230-232, 2006.

[6] Mohan S K, Hau K F,"Watermarking of Electronic Text Documents," Electronic Commerce Research, vol. 2, pp. 169-187, 2002.

[7] Qiang Wang, Xingtong Liu, "A New Watermarking Al gorithm of PDF Document Based on Correct Coding," Computing Technology and Automation,vol. 28, pp. 137-141,September 2009.

[8] Qiuyu Zhang,Dongmei Yu,Wei Guan, "Digital watermarking algorithm based on Chinese PDF documents,"Computer Engineering and Design,vol. 28, 5983-5987, December 2007.

[9] Keyu Liao,Binfa Li,Zenghui Ma,"Digital Watermarking Algorithm based on PDF Document," Modern Computer,vol. 5, pp. 4-7, May 2005.

[10] Xu Liang, Zhiyong Yuan, Ming Huang, "Text digital watermarking algorithm based on line spacing code," Information Technology,vol. 3, pp. 17-21, 2008.

[11] PDF Reference fifth edition ADOBE Portable Document Format 1.7, December 2007.

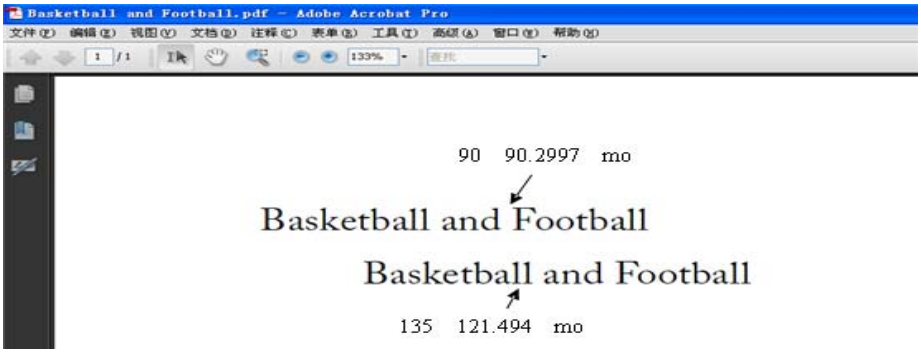


Figure 1. The use of "mo" in PDF document

```
basketball and football.ps*
/r /s /t /u /v /w /x /y
/z 133{/.notdef}rp]
EAMTBS+ACaslonPro-Regular nf
EAMTBS+ACaslonPro-Regular*1 [18 0 0 -18 0 0 ]msf
90 90.2997 mo
(Basketball and Football )
[11.8819 7.68784 6.40985 8.92984 7.66971 5.8878 9.07376 7.68782 4.73575 4.60974 4.55936 7.68784
9.57776 9.37975 4.55936 10.0279 8.96582 8.96582 5.8878 9.07376 7.68784 4.73575 4.60974 0
]xsh
135 121.494 mo
(Basketball and Football )
[11.8819 7.68784 6.40985 8.92984 7.66971 5.8878 9.07376 7.68782 4.73575 4.60974 4.55936 7.68784
9.57776 9.37975 4.55936 10.0279 8.96584 8.96582 5.88779 9.07376 7.68784 4.73575 4.60974 0
```

Figure 2. A page-description example of a part of a PDF text

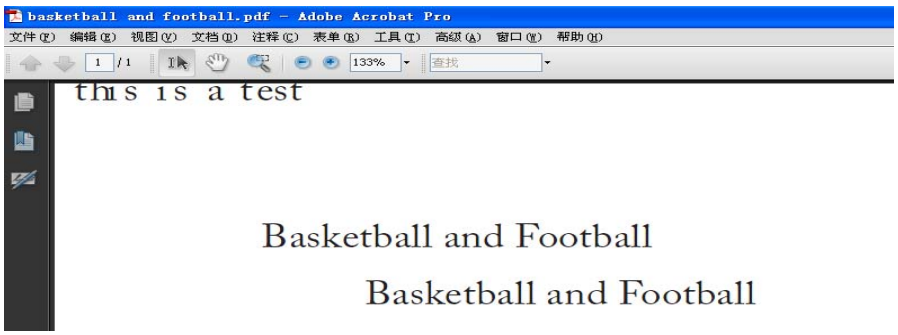


Figure 3. The new PDF file

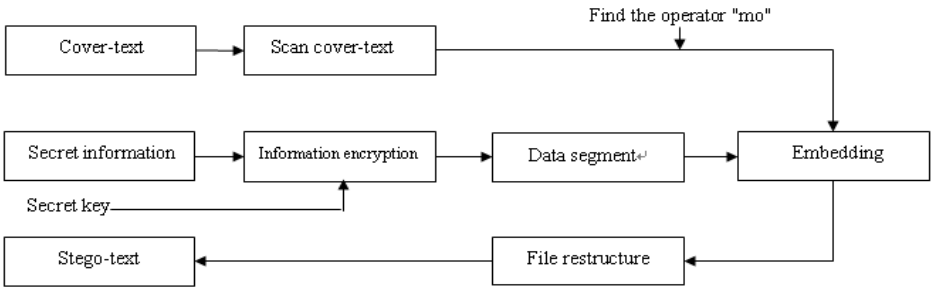


Figure 4. The flowchart of embedding

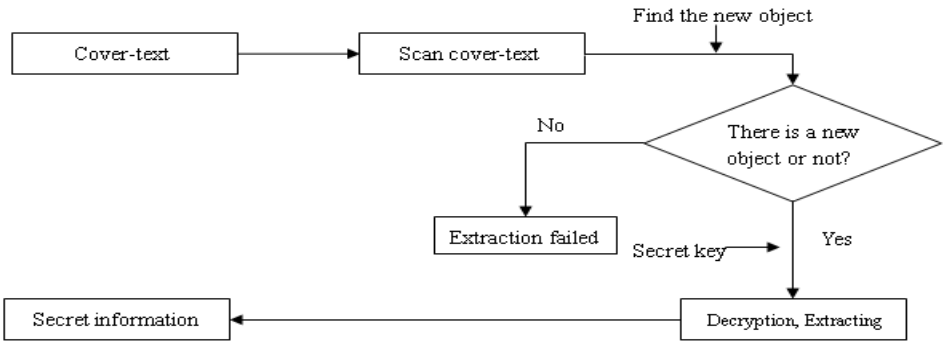
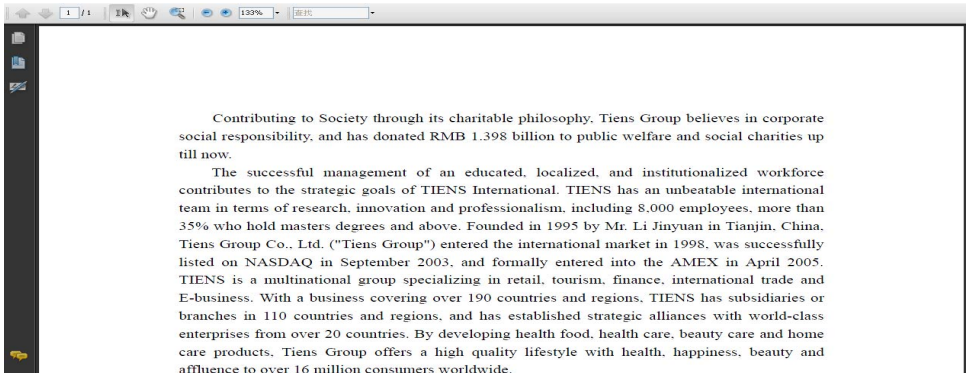


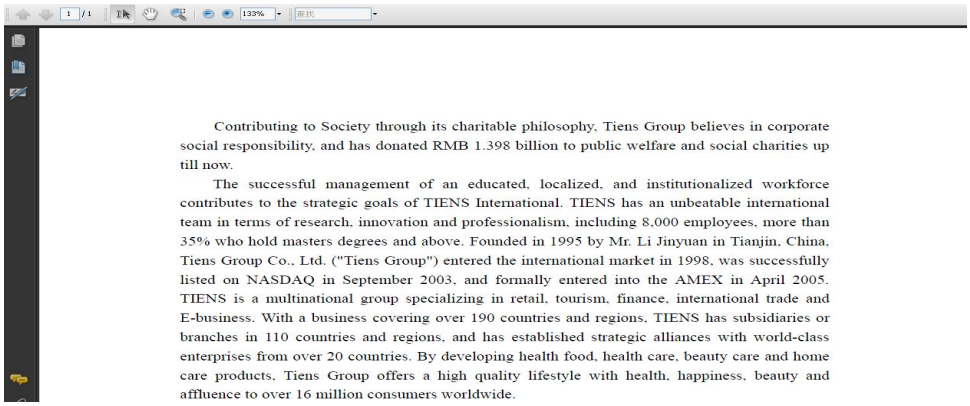
Figure 5. The flowchart of extraction



Contributing to Society through its charitable philosophy, Tiens Group believes in corporate social responsibility, and has donated RMB 1.398 billion to public welfare and social charities up till now.

The successful management of an educated, localized, and institutionalized workforce contributes to the strategic goals of TIENS International. TIENS has an unbeatable international team in terms of research, innovation and professionalism, including 8,000 employees, more than 35% who hold masters degrees and above. Founded in 1995 by Mr. Li Jinyuan in Tianjin, China, Tiens Group Co., Ltd. ("Tiens Group") entered the international market in 1998, was successfully listed on NASDAQ in September 2003, and formally entered into the AMEX in April 2005. TIENS is a multinational group specializing in retail, tourism, finance, international trade and E-business. With a business covering over 190 countries and regions, TIENS has subsidiaries or branches in 110 countries and regions, and has established strategic alliances with world-class enterprises from over 20 countries. By developing health food, health care, beauty care and home care products, Tiens Group offers a high quality lifestyle with health, happiness, beauty and affluence to over 16 million consumers worldwide.

Figure 6. A part of cover-text



Contributing to Society through its charitable philosophy, Tiens Group believes in corporate social responsibility, and has donated RMB 1.398 billion to public welfare and social charities up till now.

The successful management of an educated, localized, and institutionalized workforce contributes to the strategic goals of TIENS International. TIENS has an unbeatable international team in terms of research, innovation and professionalism, including 8,000 employees, more than 35% who hold masters degrees and above. Founded in 1995 by Mr. Li Jinyuan in Tianjin, China, Tiens Group Co., Ltd. ("Tiens Group") entered the international market in 1998, was successfully listed on NASDAQ in September 2003, and formally entered into the AMEX in April 2005. TIENS is a multinational group specializing in retail, tourism, finance, international trade and E-business. With a business covering over 190 countries and regions, TIENS has subsidiaries or branches in 110 countries and regions, and has established strategic alliances with world-class enterprises from over 20 countries. By developing health food, health care, beauty care and home care products, Tiens Group offers a high quality lifestyle with health, happiness, beauty and affluence to over 16 million consumers worldwide.

Figure 7. A part of stego-text

The Transformation of Exponential Operators In Quantum Mechanics

Zhang Ao
Science Education Department
Beijing Institute of Graphic Communication
Beijing, China
e-mail: ahf200805@sina.com

Pu Jia-Ling
Beijing Institute of Graphic Communication
Beijing, China
e-mail: pujialing@bigc.edu.cn

***Abstract*—It is very important to transform exponential operator to analytic equation in quantum mechanics. The general functions can be expressed as linear combinations of the exponential operators. It is regarded as the method of linear superposition. Three-dimensional Lie algebra is very useful for the transformation of exponential operator. By the way, one can obtain the diffusion constant which is very important for molecular crystal.**

Keywords-component; differential equation; Lie algebra; commutation relation

I. INTRODUCTION

Exponential operator expansions and solutions to differential equations occur usually in quantum mechanics [1]. It is assumed that the general functions can be expressed as linear combinations of exponential operators. It has been instructively discussed by Aizu [2] that the parameter differentiation of quantum-mechanical linear operators. Although we know how to differentiate an ordinary exponential function, it is not widely realized that a general closed formula exists for the derivative of an exponential operator with respect to a parameter [3]. The non-Abelian two-dimensional algebra is the lowest-dimensional Lie algebra for physical interest [4]. Three-dimensional Lie algebra is very important in quantum mechanics [5]. By three-dimensional Lie algebra, one can obtain an equation of motion for the density matrix which is solved to give the diffusion constant after some transformations [6]. In this paper, the transformation of exponential operator is discussed in three-dimensional Lie algebra.

II. THE THREE-DIMENSIONAL LIE ALGEBRA IN THE QUANTUM PHYSICS

In quantum mechanics, we usually use sets of operator X_1 , X_2 , $\cdots$, X_n . The $X_i's$ are defined as elements of Lie algebra. The lowest-dimensional Lie algebra for the physicist is the non-Abelian two-dimensional algebra.

The three-dimensional Lie algebra is very important for the quantum physics. The operators P,Q,I are defined as element of the three-dimensional Lie algebra. The operators satisfy the commutation relations

$$[P,Q]=PQ-QP=cI \tag{1}$$

$$[P,I]=PI-IP=0 \tag{2}$$

$$[Q,I]=QI-IQ=0 \tag{3}$$

Where P , Q and I represent the coordinate, the momentum and the unit operator, respectively.

We define the operator W

$$W=P^2+Q^2 \tag{4}$$

For the three-dimensional Lie algebra, the commutation relations have the form

$$[W,P]=WP-PW=-2cQ \tag{5}$$

$$[W,Q]=WQ-QW=-2cP \tag{6}$$

$$[W,I]=WI-IW=0 \tag{7}$$

The operators are defined as

$$Z=\alpha P+\beta Q+\gamma W \tag{8}$$

$$G=dP+rQ+gW \tag{9}$$

$$G(x)=d(x)P+r(x)Q+gW+u(x)I \tag{10}$$

The $G(x)$ satisfies the differential equation

$$\frac{dG(x)}{dx}=[Z,G(x)] \tag{11}$$

From Eqs.(10) and (11), we can get the form

$$\begin{aligned} &d'(x)P+r'(x)Q+u'(x)I \\ &=[\alpha P+\beta Q+\gamma W,d(x)P+r(x)Q+gW+u(x)I] \\ &=\alpha r(x)[P,Q]+\alpha g[P,W]+\beta d(x)[Q,P] \\ &\quad+\beta g[Q,W]+\gamma d(x)[W,P]+\gamma r(x)[W,Q] \end{aligned} \tag{12}$$

From Eqs.(12), we can find the differential equations

$$d'(x)=2c\gamma r(x)-2c\beta g \tag{13}$$

$$r'(x)=2c\alpha g-2c\gamma d(x) \tag{14}$$

$$u'(x)=c\alpha r(x)-c\beta d(x) \tag{15}$$

From Eqs.(13), we can get the quadratic differential equation

$$d''(x)=2c\gamma r'(x)\ =2c\gamma[2c\alpha g-2c\gamma d(x)] \tag{16}$$

Then we can get the form of the differential equation

$$d''(x)+\left(2c\gamma\right)^2 d(x)=\left(2c\right)^2\gamma\alpha g \tag{17}$$

Subject to the initial condition

$$d(0)=d \tag{18}$$

The solution of Eqs.(17) and (18) is easily found to be given by

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.100

$$d(x)=(d-\frac{\alpha g}{\gamma})\cos(2c\gamma x)+C_2\sin(2c\gamma x)+\frac{\alpha g}{\gamma} \tag{19}$$

From Eqs.(14), we can get the quadratic differential equation

$$r''(x)=-2c\gamma d'(x)=-2c\gamma[2c\gamma r(x)-2c\beta g] \tag{20}$$

Then we can get the form of the differential equation

$$r''(x)+\left(2c\gamma\right)^2 r(x)=\left(2c\right)^2\gamma\beta g \tag{21}$$

Subject to the initial condition

$$r(0)=r \tag{22}$$

The solution of Eqs.(21) and (22) is easily found to be given by

$$r(x)=(r-\frac{\beta g}{\gamma})\cos(2c\gamma x)+D_2\sin(2c\gamma x)+\frac{\beta g}{\gamma} \tag{23}$$

From Eqs.(13), (19) and (23), we can get the follow equation

$$\begin{aligned}&(d-\frac{\alpha g}{\gamma})(-2c\gamma)\sin(2c\gamma x)+\mathrm{C}_2(2c\gamma)\cos(2c\gamma x)\\&=(2c\gamma)(r-\frac{\beta g}{\gamma})\cos(2c\gamma x)+D_2(2c\gamma)\sin(2c\gamma x)\end{aligned} \tag{24}$$

From Eqs.(24)， it follows that

$$\mathrm{C}_2=r-\frac{\beta g}{\gamma},\ D_2=-(d-\frac{\alpha g}{\gamma}) \tag{25}$$

Then we can get the forms of $d(x)$ and $r(x)$

$$\begin{aligned}d(x)&=(d-\frac{\alpha g}{\gamma})\cos(2c\gamma x)\\&\quad+(r-\frac{\beta g}{\gamma})\sin(2c\gamma x)+\frac{\alpha g}{\gamma}\end{aligned} \tag{26}$$

$$\begin{aligned}r(x)&=(r-\frac{\beta g}{\gamma})\cos(2c\gamma x)\\&\quad-(d-\frac{\alpha g}{\gamma})\sin(2c\gamma x)+\frac{\beta g}{\gamma}\end{aligned} \tag{27}$$

Substituting Eqs.(26) and (27) into Eqs.(15), one obtains

$$\begin{aligned}u'(x)&=c\alpha r(x)-c\beta d(x)\\&=c(\alpha r-\beta d)\cos(2c\gamma x)\\&\quad-c(\alpha d+\beta r-\frac{\alpha^2+\beta^2}{\gamma}g)\sin(2c\gamma x)\end{aligned} \tag{28}$$

The solution of the above differential equation has the form

$$\begin{aligned}u(x)&=\frac{1}{2\gamma}(\alpha r-\beta d)\sin(2c\gamma x)\\&\quad+\frac{1}{2\gamma}(\alpha d+\beta r-\frac{\alpha^2+\beta^2}{\gamma}g)\cos(2c\gamma x)+D\end{aligned} \tag{29}$$

By the initial condition

$$u(0)=0 \tag{30}$$

We can get the form of the constant

$$D=-\frac{1}{2\gamma}(\alpha d+\beta r-\frac{\alpha^2+\beta^2}{\gamma}g) \tag{31}$$

Substituting Eqs.(31) into (29), one obtains

$$\begin{aligned}u(x)&=\frac{1}{2\gamma}(\alpha r-\beta d)\sin(2c\gamma x)\\&\quad+\frac{1}{2\gamma}(\alpha d+\beta r-\frac{\alpha^2+\beta^2}{\gamma}g)[\cos(2c\gamma x)-1]\end{aligned} \tag{32}$$

From Eqs.(26), (27) and (32), one obtains

$$e^{Z}F(P,Q)e^{-Z}=F(J,K) \tag{33}$$

Where

$$\begin{aligned}J&=P\cos(2c\gamma x)-Q\sin(2c\gamma x)\\&\quad+\frac{1}{2\gamma}\{\alpha[\cos(2c\gamma x)-1]-\beta\sin(2c\gamma x)\}\end{aligned} \tag{34}$$

$$\begin{aligned}K&=P\sin(2c\gamma x)+Q\cos(2c\gamma x)\\&\quad+\frac{1}{2\gamma}\{-\alpha\sin(2c\gamma x)+\beta[\cos(2c\gamma x)-1]\}\end{aligned} \tag{35}$$

When $x=1$, we can get the form

$$\begin{aligned}J&=P\cos(2c\gamma)-Q\sin(2c\gamma)\\&\quad+\frac{1}{2\gamma}\{\alpha[\cos(2c\gamma)-1]-\beta\sin(2c\gamma)\}\end{aligned} \tag{36}$$

$$\begin{aligned}K&=P\sin(2c\gamma)+Q\cos(2c\gamma)\\&\quad+\frac{1}{2\gamma}\{-\alpha\sin(2c\gamma)+\beta[\cos(2c\gamma)-1]\}\end{aligned} \tag{37}$$

One can obtain the motion equation for the density matrix by the transformation of Eqs.(36) and (37)

III. DISCUSSION

For a model exciton-phonon Hamiltonian containing both linear and quadratic coupling terms, we can obtain the equation of motion for the density matrix by some transformations of exponential operator. We assume that the general functions can be expressed as linear combinations of exponential operators. By this paper, we can obtain the solution of the transformation of exponential operator by three-dimensional Lie algebra way in quantum mechanics, and obtain the diffusion constant. It is very important for the molecular crystal.

.

ACKNOWLEDGMENT

This project was supported by financially supported

By PHR20090515, BIGC(Ea201016).

REFERENCES

[1] R.M.Wilcox, J.Math.Rev.8,962(1967)
[2] K.Aizu, J.Math.Rev.4,762(1963).
[3] R.W.Munn, R.Silbey, J.Chem.Phys.68,2493(1978).
[4] R.W.Munn, J.Chem.Phys.58,3230(1973).
[5] R.W.Munn, J.Chem.Phys.58,3230(1973).
[6] Yang Zhao, David W.Brown and Katja Lidenberg, J.Chem.Phys.100, 2335(1994)

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Generalized Laguerre Interpolation and Its Application to Differential Equation

Zhang Xiao-yong
Department of Mathematics, Shanghai Maritime University, Pudong Avenue, 1550, Shanghai, 200135, China.
zhxiyo@gmail.com

Sui Jiang Hua
Mechanical engineering college, Dalian Ocean University, China

***Abstract*—In this paper,we develop the generalized Laguerre pseudospectral method for solving initial problems of second order ordinary differential equation. We also propose a multi-step version of the generalized Laguerre collocation method, which are very efficient for long-time numerical simulations. The numerical solutions possess the spectral accuracy. The global convergence of proposed algorithms are proved. Numerical results demonstrate the effectiveness of proposed methods and coincide well with theoretical analysis.**

***Index Terms*—Generalized Laguerre interpolation, Collocation method, Second order ordinary differential equation.**

I. Introduction

Many practical problems arising in science and engineering require solving initial problems of second order ODEs. For examples, after spacial discretization, a large class of nonlinear wave equations, such as Klein-Gorden and sine-Gorden equations, are reduced to certain systems of second order ODEs. We may reformulate such problems to certain systems of first order ODEs and then solve them numerically. Whereas, for saving work, it seems reasonable to solve them directly.

In many literatures, one focused on finite difference schemes for the equation $\frac{d^2u}{d\rho^2} = f(u,\rho)$, which are usually divided into two classes. In the first class of finite difference schemes, the coefficients depend on known periods or oscillation frequencies of solutions, including exponential-fitted and trigonometrically-fitted methods and some linear multi-step.In the second class of finite difference schemes, the coefficients are constants, such as Runge-Kutta-Nystrom method, linear multi-step method, hybrid method,collocation method and so on.

Relatively, there have been less existing results on numerical methods for the general equation $\frac{d^2u}{d\rho^2} = f(u, \frac{du}{d\rho}, \rho)$, for which one oftentimes used Runge-Kutta-Nystrom method and linear multi-step method. Some authors also developed other methods, for instance, the collocations method using spline and Chebyshev approximations.

As we know, spectral method employs global orthogonal polynomials as trial functions. It provides exceedingly accurate numerical results. However,so far,there is no work concerning the collocation methods keeping the spectral accuracy, for initial value problems of second order ODEs, since it is not easy to design proper algorithms and analyze their numerical errors precisely.

This paper is for a new collocation method for solving initial value problems of the equation $\frac{d^2u}{d\rho^2} = f(u, \frac{du}{d\rho}, \rho)$. In the next section, we derive the new generalized Laguerre interpolation approximation results. Then, we use the new results to approximate the exact solution directly and analyze the numerical error in Section 3. This algorithm is easy to be implemented and possesses the spectral accuracy. Roughly speaking, for fixed mode N, the smoother the exact solutions ,the more accurate the numerical result. Moreover, even if the solution has certain weak singularity, the numerical result still keeps the high accuracy. In opposite, the commonly used finite difference schemes do not have such merits. In Section 4, we propose a multi-step version of collocation method by using the generalized Laguerre interpolation. By this approach with moderate mode N, we could evaluate numerical solution step by step as time increases. This feature simplifies actual calculations and saves much work. Moreover, it not only shares the same advantages as the first method, but also possesses the global convergence. Therefore,it is more appropriate foe long-time calculation. In particular, for any smooth solution and fixed τ (mesh size in time), the numerical error decay exponentially as N increases, while for any fixed mode N, the numerical error decay as $\tau \longrightarrow 0$, often faster than those of many usual finite difference schemes. It is noted that this new algorithm also works well even for some solutions growing rapidly as $\tau \longrightarrow \infty$. In Section 5, we present some numerical results which demonstrate the spectral accuracy of proposed schemes and coincide well with the theoretical analysis.The final section is for some concluding remarks.

II. Orthogonal Approximation by Using Generalized Laguerre Polynomials

In this section, we investigate some results about the generalized Laguerre approximation. Let $\Lambda = \{\rho \mid 0 < \rho < \infty\}$. We define

$$L^2_\chi(\Lambda) = \{v \mid v \text{ is measurable on } \Lambda \text{and } \|v\|_{\chi,\Lambda} < \infty\},$$

with the following inner product and norm

$$(u,v)_{\chi,\Lambda} = \int_\Lambda u(\rho)v(\rho)\chi(\rho)d\rho, \qquad ||v||_{\chi,\Lambda} = (v,v)^{\frac{1}{2}}_{\chi,\Lambda}.$$

For any integer $m \geq 0$, we define the space

$$H^m_\chi(\Lambda) = \{v \mid \frac{d^k v}{d\rho^k} \in L^2_\chi(\Lambda),\ 0 \leq k \leq m\},$$

equipped with the following inner product, semi-norm and norm

$$(u,v)_{m,\chi,\Lambda} = \sum_{0\leq k\leq m} (\frac{d^k u}{d\rho^k}, \frac{d^k v}{d\rho^k})_{\chi,\Lambda},$$

$$|v|_{m,\chi,\Lambda} = \|\frac{d^m u}{d\rho^m}\|_{\chi,\Lambda}, \qquad \|v\|_{m,\chi,\Lambda} = (v,v)^{1/2}_{m,\chi,\Lambda}.$$

For any $r > 0$, the space $H^r_\chi(\Lambda)$ and its norm $\|v\|_{r,\chi,\Lambda}$ are defined by space interpolation as in [2]. In particular, ${}_0H^1_\chi(\Lambda) = \{v \in H^1_\chi(\Lambda) \mid v(0) = 0\}$.

Let $\omega_{\alpha,\beta}(\rho) = \rho^\alpha e^{-\beta\rho}, \alpha > -1,\ \beta > 0$. The corresponding generalized Laguerre polynomials of degree l are defined by

$$\mathcal{L}^{(\alpha,\beta)}_l(\rho) = \frac{1}{l!}\rho^{-\alpha}e^{\beta\rho}\frac{d^l}{d\rho^l}(\rho^{l+\alpha}e^{-\beta\rho}), \quad l = 0,1,2,\cdots$$

They fulfill the following recurrence relations

$$\frac{d\mathcal{L}^{(\alpha,\beta)}_l(\rho)}{d\rho} = -\beta\sum_{k=0}^{l-1}\mathcal{L}^{(\alpha,\beta)}_k(\rho), \tag{2.1}$$

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.48

$$\frac{d\mathcal{L}_l^{(\alpha,\beta)}(\rho)}{d\rho} = -\beta\mathcal{L}_{l-1}^{(\alpha+1,\beta)}(\rho). \tag{2.2}$$

Therefore, it is straightforward to derive the following property (cf. [1])

$$\mathcal{L}_l^{(\alpha,\beta)}(0) = \frac{\Gamma(l+\alpha+1)}{\Gamma(\alpha+1)\Gamma(l+1)}, \qquad l \geq 0. \tag{2.3}$$

The set of $\mathcal{L}_l^{(\alpha,\beta)}(\rho)$ is the complete $L^2_{\omega_{\alpha,\beta}}(\Lambda)$-orthogonal system, namely,

$$(\mathcal{L}_l^{(\alpha,\beta)},\ \mathcal{L}_m^{(\alpha,\beta)})_{\omega_{\alpha,\beta}} = \begin{cases} \gamma_l^{(\alpha,\beta)}, & l = m, \\ 0, & l \neq m. \end{cases}$$

$$\text{and}(\frac{d\mathcal{L}_l^{(\alpha,\beta)}(\rho)}{d\rho},\ \frac{d\mathcal{L}_m^{(\alpha,\beta)}(\rho)}{d\rho})_{\omega_{\alpha+1,\beta}} = \begin{cases} \beta^2\gamma_{l-1}^{(\alpha+1,\beta)}, & l = m, \\ 0, & l \neq m. \end{cases}$$

where

$$\gamma_l^{(\alpha,\beta)} = \frac{\Gamma(l+\alpha+1)}{\beta^{\alpha+1}\Gamma(l+1)}. \tag{2.4}$$

Now, let N be any positive integer and $\mathcal{P}_N(\Lambda)$ be the set of all algebraic polynomials of degree at most N. Furthermore, ${}_0\mathcal{P}_N(\Lambda) = \{v \in \mathcal{P}_N(\Lambda) \mid v(0) = 0\}$.

In forthcoming discussions, we shall use the following Lemma

Lemma 2.1. For any $\phi \in \mathcal{P}_N(\Lambda)$, integer $r \geq 0$

$$\|\phi\|^2_{r,\omega_{\alpha,\beta},\Lambda} \leq c(\beta N)^{2r}\|\phi\|^2_{\omega_{\alpha,\beta},\Lambda}$$

The proof see (cf. [7]).

Now, let $\rho_{G,N-m-1,j}^{(\alpha+m,\beta)}, 0 \leq j \leq N-m-1$, denote the zeros of $\dfrac{d^m\mathcal{L}_N^{(\alpha,\beta)}(\rho)}{d\rho^m}$. By (2.2), we get that

$$\frac{d^m\mathcal{L}_N^{(\alpha,\beta)}(\rho)}{d\rho^m} = (-\beta)^m\mathcal{L}_{N-m}^{(\alpha+m,\beta)}(\rho) = c\prod_{j=0}^{N-m-1}(\rho-\rho_{G,N-m-1,j}^{(\alpha+m,\beta)}). \tag{2.5}$$

By virtue of these formulas, we obtain the following results:

Theorem 2.1. Let N is a fixed positive number. Denoting by $\rho_{G,N-m-1,j}^{(\alpha+m,\beta)}, 0 \leq j \leq N-m-1$, the zeros of $\dfrac{d^m\mathcal{L}_N^{(\alpha,\beta)}(\rho)}{d\rho^m}$. There exists a unique set of $N-m$ real numbers ω_j^m such that the following equality holds for any polynomial ϕ belongs to $\mathcal{P}_{2N-2m-1}(\Lambda)$.

$$\int_0^{+\infty}\phi(\rho)\omega_{\alpha+m,\beta}d\rho = \sum_{j=0}^{N-m-1}\phi(\rho_{G,N-m-1,j}^{(\alpha+m,\beta)})\omega_j^m \quad \forall\phi \in \mathcal{P}_{2N-2m-} \quad {}_{\omega_\beta} \tag{2.6}.$$

Theorem 2.2. Let $\rho_{G,N-m-1,j}^{(\alpha+m,\beta)}, 0 \leq j \leq N-m-1$, are the zeros of $\dfrac{d^m}{d\rho^m}\mathcal{L}_N^{(\alpha,\beta)}(\rho)$. There exists

(1) a unique set of $N-m$ real numbers $\varpi_j^m, 0 \leq j \leq N-m-1$,

(2) a unique set of m real numbers $\sigma_0^{m,k}, 0 \leq k \leq m-1$

such that the following equality holds for any polynomial ϕ belongs to $\mathcal{P}_{2N-m-1}(\Lambda)$.

$$\int_0^{+\infty}\phi(\rho)\omega_{\alpha,\beta}d\rho = \sum_{j=0}^{N-m-1}\phi(\rho_{G,N-m-1,j}^{(\alpha+m,\beta)})\varpi_j^m + \sum_{k=0}^{m-1}\frac{d^k\phi(0)}{d\rho^k}\sigma_0^{m,k}. \tag{2.7}$$

Taking $m = 2$, and $\rho_{G,N-3,j}^{(\alpha+2,\beta)}, \quad 0 \leq j \leq N-3$ are zeros of $\mathcal{L}_{N-2}^{(\alpha+2,\beta)}(\rho)$ in Theorem (2.2), for any ϕ in $\mathcal{P}_{2N-3}(\Lambda)$ the following equality holds

$$\int_0^{+\infty}\phi(\rho)\omega_{\alpha,\beta}d\rho = \sum_{j=0}^{N-3}\phi(\rho_{G,N-3,j}^{(\alpha+2,\beta)})\varpi_j^2 + \phi(0)\sigma_0^{2,0} + \frac{d\phi(0)}{d\rho}\sigma_0^{2,1}. \tag{2.8}$$

Next,we define the discrete inner product

$$[\phi,\psi]_{N,2} = \sum_{j=0}^{N-3}\phi(\rho_{G,N-3,j}^{(\alpha+2,\beta)})\psi(\rho_{G,N-3,j}^{(\alpha+2,\beta)})\varpi_j^2 + (\phi\psi)(0)\sigma_0^{2,0} + \frac{d(\phi\psi)(0)}{d\rho}\sigma_0^{2,1}. \tag{2.9}$$

By (2.8),for any $\phi\psi \in \mathcal{P}_{2N-3}(\Lambda)$

$$[\phi,\psi]_{N,2} = \int_0^{+\infty}\phi(\rho)\psi(\rho)\omega_{\alpha,\beta}d\rho. \tag{2.10}$$

About the discrete inner product, we have that

Theorem 2.3. If ϕ belongs to $\mathcal{P}_{N-1}(\Lambda)$, then

$$c\|\phi\|^2_{\omega_{\alpha,\beta}(\Lambda)} \leq [\phi,\phi]_{N,2} \leq \|\phi\|^2_{\omega_{\alpha,\beta}(\Lambda)}. \tag{2.11}$$

Now let $\omega_\beta(\rho) = \omega_{0,\beta}(\rho) = e^{-\beta\rho},\ \beta > 0$, we define

$$\mathcal{L}_l^{(\beta)}(\rho) = \mathcal{L}_l^{(0,\beta)}(\rho), \quad l = 0,1,2,\cdots$$

For technical reason, we need another Sobolve space associated with a positive integer α,

$$H^r_{\omega_\beta}(\Lambda,\alpha) = \{v \in H^r_{\omega_\beta}(\Lambda) | \rho^{\frac{\alpha}{2}}v \in H^r_{\omega_\beta}(\Lambda)\}.$$

This space is provided with the associated natural norm $\|v\|_{r,\omega_\beta,\alpha} = \|v(1+\rho)^{\frac{\alpha}{2}}\|_{r,\omega_\beta}$

We define the projection operator $P_{N,\beta} : L^2_{\omega_\beta}(\Lambda) \longrightarrow \mathcal{P}_N(\Lambda)$ as

$$\int_0^{+\infty}(v - P_{N,\beta}v)\phi\omega_\beta d\rho = 0, \qquad \forall\phi \in \mathcal{P}_N(\Lambda).$$

The following result characterize the property of $P_{N,\beta}$.

Theorem 2.4. Let $r \geq 0$ be a positive integer.α is the largest integer for which $\alpha < r+1$.Then for any any function v in $H^r_{\omega_\beta}(\Lambda,\alpha)$, the following estimate holds

$$\|v - P_{N,\beta}v\|_{\omega_\beta} \leq cN^{-\frac{r}{2}}\|v\|_{r,\omega_\beta,\alpha}. \tag{2.12}$$

Lemma 2.2. For any $v \in H^r_{\omega_\beta}(\Lambda,\alpha)$, and integer $0 \leq 2\mu \leq r$,

$$\|P_{N,\beta}\partial_\rho v - \partial_\rho P_{N,\beta}v\|^2_{\mu,\omega_\beta} \leq c(\beta N)^{2\mu-r}\|v\|^2_{r,\omega_\beta,\alpha}.$$

Using this Lemma, we obtain the following Theorem

Theorem 2.5. For any $v \in H^r_{\omega_\beta}(\Lambda,\alpha)$, and $0 \leq 2\mu \leq r$,

$$\|P_{N,\beta}v - v\|^2_{\mu,\omega_\beta} \leq c(\beta N)^{2\mu-r}\|v\|^2_{r,\omega_\beta,\alpha}.$$

where α is the largest integer such that $\alpha < r+1$.

Next, we define the interpolation operator $\mathcal{I}_{N-1,\beta}$. For any ϕ in $H^1_1(\Lambda)$, $\mathcal{I}_{N-1,\beta}\phi \in \mathcal{P}_{N-1}(\Lambda)$ and

$$\begin{cases} \mathcal{I}_{N-1,\beta}\phi(0) = \phi(0), \quad \dfrac{d}{d\rho}\mathcal{I}_{N-1,\beta}\phi(0) = \dfrac{d}{d\rho}\phi(0), \\ \mathcal{I}_{N-1,\beta}\phi(\rho_{G,N-3,j}^{(2,\beta)}) = \phi(\rho_{G,N-3,j}^{(2,\beta)}), \quad 0 \leq j \leq N-3. \end{cases}$$

where $\rho_{G,N-3,j}^{(2,\beta)}, \quad 0 \leq j \leq N-3$, are the zeros of $\mathcal{L}_{N-2}^{(2,\beta)}(\rho)$.

To obtain the interpolation error, we study the asymptotic behaviors of generalized Laguerre-Gauss interpolation nodes and weights. Let $\rho_{G,N,j}^{(\alpha,\beta)}$ and $\rho_{R,N,j}^{(\alpha,\beta)}, \quad 0 \leq j \leq N$, be the zeros of $\mathcal{L}_{N+1}^{(\alpha,\beta)}(\rho)$ and $\rho\frac{d}{d\rho}\mathcal{L}_{N+1}^{(\alpha,\beta)}(\rho)$, respectively. They are arranged in ascending order. Denote $\omega_{Z,N,j}^{(\alpha,\beta)}, \quad 0 \leq j \leq N, \quad Z = G,R$, the corresponding Christoffel numbers such that

$$\int_0^{+\infty}\phi\omega_{\alpha,\beta}d\rho = \sum_{j=0}^{N}\phi(\rho_{Z,N,j}^{(\alpha,\beta)})\omega_{Z,N,j}^{(\alpha,\beta)}, \qquad \forall\phi \in \mathcal{P}_{2N+\lambda_Z}(\Lambda).$$

where $\lambda_Z = 1$ and 0 for $Z = G$ and R,respectively. We now deduce the expressions of the weights. Indeed,

$$\omega_{G,N,j}^{(\alpha,\beta)} = \frac{1}{\frac{d}{d\rho}\mathcal{L}_{N+1}^{(\alpha,\beta)}(\rho_{G,N,j}^{(\alpha,\beta)})}\int_0^{+\infty}\frac{\mathcal{L}_{N+1}^{(\alpha,\beta)}(\rho)}{\rho - \rho_{G,N,j}^{(\alpha,\beta)}}\omega_{\alpha,\beta}d\rho, \ \ 0 \leq j \leq N. \tag{2.13}$$

Similarly, for the Gauss-Radau weights, we have

$$\omega_{R,N,j}^{(\alpha,\beta)}=\frac{1}{\frac{d}{d\rho}(\rho\frac{d}{d\rho}\mathcal{L}_{N+1}^{(\alpha,\beta)}(\rho))|_{\rho=\rho_{R,N,j}^{(\alpha,\beta)}}}\int_0^{+\infty}\frac{\rho\frac{d}{d\rho}\mathcal{L}_{N+1}^{(\alpha,\beta)}(\rho)}{\rho-\rho_{R,N,j}^{(\alpha,\beta)}}\omega_{\alpha,\beta}d\rho, \tag{2.14}$$

By (2.2), we have that

$$\rho_{R,N,j}^{(\alpha,\beta)}=\rho_{G,N-1,j-1}^{(\alpha+1,\beta)},\quad 1\le j\le N. \tag{2.15}$$

By virtue of (2.2),(2.14) and (2.15)

$$\omega_{R,N,j}^{(\alpha,\beta)}=(\rho_{R,N,j}^{(\alpha,\beta)})^{-1}\omega_{G,N-1,j-1}^{(\alpha+1,\beta)},\quad 1\le j\le N. \tag{2.16}$$

Using formula (15.3.15) of [3],we can verify that for a certain fixed number $\eta>0$,

$$\omega_{G,N,j}^{(\alpha,\beta)}\cong\frac{\pi}{\sqrt{\beta N}}e^{-\beta\rho_{G,N,j}^{(\alpha,\beta)}}(\rho_{G,N,j}^{(\alpha,\beta)})^{\alpha+\frac{1}{2}},\quad 0\le\rho_{G,N,j}^{(\alpha,\beta)}\le\frac{\eta}{\beta}. \tag{2.17}$$

Set $\rho_{G,N,-1}^{(\alpha,\beta)}=0$, By the formula (2.4),(2.5) and (2.7) of [6]

$$\omega_{G,N,j}^{(\alpha,\beta)}\sim\omega_{\alpha,\beta}(\rho_{G,N,j}^{(\alpha,\beta)})(\rho_{G,N,j}^{(\alpha,\beta)}-\rho_{G,N,j-1}^{(\alpha,\beta)}),\quad 0\le j\le N. \tag{2.18}$$

By virtue of (2.16), we obtain from (2.17) and (2.18) that

$$\begin{aligned}\omega_{R,N,j}^{(\alpha,\beta)}\quad&\cong\frac{\pi}{\sqrt{\beta(N-1)}}e^{-\beta\rho_{R,N,j}^{(\alpha,\beta)}}(\rho_{R,N,j}^{(\alpha,\beta)})^{\alpha+\frac{1}{2}},\\ &\text{if}\ \ 0\le\rho_{R,N,j}^{(\alpha,\beta)}\le\tfrac{\eta}{\beta},\quad 1\le j\le N.\end{aligned} \tag{2.19}$$

and

$$\omega_{R,N,j}^{(\alpha,\beta)}\sim\omega_{\alpha,\beta}(\rho_{R,N,j}^{(\alpha,\beta)})(\rho_{R,N,j}^{(\alpha,\beta)}-\rho_{R,N,j-1}^{(\alpha,\beta)}),\quad 1\le j\le N. \tag{2.20}$$

(2.14) together with formula (3.6.2) of [5], yields

$$\omega_{R,N,0}^{(\alpha,\beta)}=\frac{(\alpha+1)\Gamma^2(\alpha+1)\Gamma(N+1)}{\beta^{\alpha+1}\Gamma(N+\alpha+2)}. \tag{2.21}$$

Using Theorem 8.9.2 of [3], we can verify that for a certain fixed number $\eta>0$,

$$2\beta^{\frac{1}{2}}(\rho_{G,N,j}^{(k,\beta)})^{\frac{1}{2}}=\frac{1}{\sqrt{N+1}}(j\pi+\mathcal{O}(1)),\ \ 0<\rho_{G,N,j}^{(k,\beta)}\le\frac{\eta}{\beta},\ \ k=0,1,2. \tag{2.22}$$

Using Theorem 6.31.3 of [3], for large j, we have that

$$\begin{aligned}&\tfrac{c_1j^2}{\beta(N+\frac{k}{2}+\frac{3}{2})}<\rho_{G,N,j}^{(k,\beta)}<\tfrac{c_2j^2}{\beta(N+\frac{k}{2}+\frac{3}{2})},\\ &c_1\cong\tfrac{\pi^2}{4},\quad c_2\cong 4.\quad k=0,1,2.\end{aligned} \tag{2.23}$$

From Theorem 6.31.2 of [3], the largest node satisfies

$$\begin{aligned}&\rho_{G,N,N}^{(k,\beta)}<\beta^{-1}(2(N+1)+k+1+((2(N+1)+k+1)^2+\tfrac{1}{4}-k^2)^{\frac{1}{2}})\\ &\cong 4\beta^{-1}(N+1).\quad k=0,1,2.\end{aligned} \tag{2.24}$$

Next,we give out formula of ϖ_i^2 with $\alpha=0$ and $\sigma_0^{2,k}$ with $\alpha=0$. Taking $\alpha=0,\phi=\rho^2\frac{\mathcal{L}_{N-2}^{(2,\beta)}(\rho)}{\rho-\rho_{G,N-3,i}^{(2,\beta)}}$ in (2.8),then

$$\int_0^{+\infty}\rho^2\frac{\mathcal{L}_{N-2}^{(2,\beta)}(\rho)}{\rho-\rho_{G,N-3,i}^{(2,\beta)}}\omega_\beta d\rho=(\rho_{G,N-3,i}^{(2,\beta)})^2\frac{d\mathcal{L}_{N-2}^{(2,\beta)}(\rho_{G,N-3,i}^{(2,\beta)})}{d\rho}\varpi_i^2.$$

So, by (2.14), (2.15)

$$\varpi_i^2=\frac{\omega_{R,N-2,i+1}^{(1,\beta)}}{\rho_{R,N-2,i+1}^{(1,\beta)}}\qquad 0\le i\le N-3. \tag{2.25}$$

Taking $\alpha=0,\phi=\rho\mathcal{L}_{N-2}^{(2,\beta)}(\rho)$ in (2.8)

$$\int_0^{+\infty}\rho\mathcal{L}_{N-2}^{(2,\beta)}(\rho)\omega_\beta d\rho=\frac{d(\rho\mathcal{L}_{N-2}^{(2,\beta)})}{d\rho}(0)\sigma_0^{2,1}.$$

Using (2.14),(2.21) leads to

$$\sigma_0^{2,1}=\omega_{R,N-2,0}^{(1,\beta)}\cong\frac{1}{(\beta N)^2}. \tag{2.26}$$

Taking $\alpha=0,\phi=\mathcal{L}_{N-2}^{(2,\beta)}(\rho)$ in (2.8),we obtain from (2.3),(2.25) that

$$\sigma_0^{2,0}\quad=\frac{\mathcal{L}_{N-1}^{(1,\beta)}(0)-1}{\beta\mathcal{L}_{N-2}^{(2,\beta)}(0)}+\beta\frac{\mathcal{L}_{N-3}^{(3,\beta)}(0)}{\mathcal{L}_{N-2}^{(2,\beta)}(0)}\sigma_0^{2,1}\cong\frac{1}{\beta N}. \tag{2.27}$$

Using these results leads to

Theorem 2.6. If ϕ belongs to $H_{\omega_\beta}^2(\Lambda)\cap H_{\omega_{\frac{\beta}{2}}}^1(\Lambda)$, then

$$\|\mathcal{I}_{N-1,\beta}\phi\|_{\omega_\beta,\Lambda}^2\le c(N^{-1}\|\phi\|_{H_{\omega_\beta}^2(\Lambda)}^2+\ln N\|\phi\|_{H_{\omega_{\frac{\beta}{2}}}^1(\Lambda)}^2). \tag{2.28}$$

Theorem 2.7. If ϕ belongs to $H_{\omega_{\frac{\beta}{2}}}^s(\Lambda,\alpha)$, for any integer $s>7$,

$$\|\phi-\quad\mathcal{I}_{N-1,\beta}\phi\|_{\omega_\beta,\Lambda}\le cN^{\frac{3-s}{2}}\|\phi\|_{s,\omega_{\frac{\beta}{2}},\alpha}. \tag{2.29}$$

$$\|\phi-\quad\mathcal{I}_{N-1,\beta}\phi\|_{2,\omega_\beta,\Lambda}\le cN^{\frac{7-s}{2}}\|\phi\|_{s,\omega_{\frac{\beta}{2}},\alpha}. \tag{2.30}$$

III. Generalized Laguerre Pseudospectral Method

First, we note that the set of $\mathcal{L}_l^{(\beta)}(\rho)$ is the complete $L_{\omega_\beta}^2(\Lambda)$-orthogonal system, namely,

$$(\mathcal{L}_l^{(\beta)}(\rho),\ \mathcal{L}_m^{(\beta)}(\rho))_{\omega_\beta,\Lambda}=\begin{cases}\frac{1}{\beta}, & l=m,\\ 0, & l\ne m.\end{cases}$$

Using (2.3), we derive that

$$\mathcal{L}_l^{(\beta)}(0)=1,\qquad l\ge 0. \tag{3.1}$$

Obviously, by (3.1),$\mathcal{I}_{N-1,\beta}v(\rho)$ can be expanded as

$$\mathcal{I}_{N-1,\beta}v(\rho)=\sum_{l=0}^{N-1}\hat{v}_l\mathcal{L}_l^{(\beta)}(\rho)\ \text{and}\ \mathcal{I}_{N-1,\beta}v(0)=\sum_{l=0}^{N-1}\hat{v}_l=v_0.$$

By virtue of (2.1), we deduce that

$$\frac{d}{d\rho}\mathcal{I}_{N-1,\beta}v(\rho)\quad=-\beta\sum_{l=1}^{N-1}\hat{v}_l(\sum_{m=0}^{l-1}\mathcal{L}_m^{(\beta)}),\quad-\beta\sum_{l=1}^{N-1}l\hat{v}_l=v_0^{'}.$$

By the above two formulas ,we have that

$$\hat{v}_{N-1}=(2-N)v_0-\frac{1}{\beta}v_0^{'}+\sum_{l=0}^{N-3}(N-2-l)\hat{v}_l. \tag{3.2}$$

$$\hat{v}_{N-2}=(N-1)v_0+\frac{1}{\beta}v_0^{'}-\sum_{l=0}^{N-3}(N-1-l)\hat{v}_l. \tag{3.3}$$

where

$$\begin{aligned}\hat{v}_l\quad&=\beta\int_0^{+\infty}\mathcal{I}_{N-1,\beta}v(\rho)\mathcal{L}_l^{(\beta)}(\rho)\omega_\beta d\rho=\beta[\mathcal{I}_{N-1,\beta}v,\mathcal{L}_l^{(\beta)}]_{N,2}\\ &=\sum_{j=0}^{N-3}a_j^lv(\rho_{G,N-3,j}^{(2,\beta)})+a_0^lv_0+\widehat{a}_0^lv_0^{'},\quad 0\le l\le N-3.\end{aligned} \tag{3.4}$$

Clearly,

$$a_0^l=\beta(\mathcal{L}_l^{(\beta)}(0)\sigma_0^{2,0}+\frac{d}{d\rho}\mathcal{L}_l^{(\beta)}(0)\sigma_0^{2,1}),\quad\widehat{a}_0^l=\beta\mathcal{L}_l^{(\beta)}(0)\sigma_0^{2,1}$$

$$a_j^l=\beta\mathcal{L}_l^{(\beta)}(\rho_{G,N-3,j}^{(2,\beta)})\varpi_j^2,\quad 0\le j\le N-3.$$

By virtue of (3.2),(3.3) and (3.4)

$$\hat{v}_{N-1} = \sum_{j=0}^{N-3} a_j^{N-1} v(\rho_{G,N-3,j}^{(2,\beta)}) + a_0^{N-1} v(0) + \widehat{a}_0^{N-1} v^{'}(0). \tag{3.5}$$

$$\hat{v}_{N-2} = \sum_{j=0}^{N-3} a_j^{N-2} v(\rho_{G,N-3,j}^{(2,\beta)}) + a_0^{N-2} v(0) + \widehat{a}_0^{N-2} v^{'}(0). \tag{3.6}$$

where

$$a_0^{N-1} = (2-N) + \sum_{l=0}^{N-3} (N-2-l) a_0^l,$$

$$\widehat{a}_0^{N-1} = -\frac{1}{\beta} + \sum_{l=0}^{N-3} (N-2-l)\widehat{a}_0^l, \quad a_j^{N-1} = \sum_{l=0}^{N-3} (N-2-l) a_j^l,$$

$$a_0^{N-2} = (N-1) - \sum_{l=0}^{N-3} (N-1-l) a_0^l,$$

$$\widehat{a}_0^{N-2} = \frac{1}{\beta} - \sum_{l=0}^{N-3} (N-1-l)\widehat{a}_0^l, \quad a_j^{N-2} = -\sum_{l=0}^{N-3} (N-1-l) a_j^l.$$

We consider the following problem

$$\begin{cases} \dfrac{d^2 v}{d\rho^2} = f(v(\rho), v^{'}(\rho), \rho), & \rho \geq 0 \\ v(0) = v_0, \dfrac{dv(0)}{d\rho} = v_0^{'}. \end{cases} \tag{3.7}$$

Let

$$G_{\beta,1} = \frac{d^2}{d\rho^2}\mathcal{I}_{N-1,\beta} v(\rho) - \mathcal{I}_{N-1,\beta}\frac{d^2 v}{d\rho^2}. \tag{3.8}$$

$$\frac{d^2}{d\rho^2}\mathcal{I}_{N-1,\beta} \quad v(\rho_{G,N-3,j}^{(2,\beta)}) = \mathcal{I}_{N-1,\beta}\frac{d^2 v}{d\rho^2}(\rho_{G,N-3,j}^{(2,\beta)}) + G_{\beta,1}(\rho_{G,N-3,j}^{(2,\beta)}), \quad 0 \leq j \leq N-3. \tag{3.9}$$

Next, we derive an expression for the left side of (3.9). Let $\hat{v}_l$ is the coefficient of $\mathcal{I}_{N-1,\beta} v$ in terms of $\mathcal{L}_l^{(\beta)}(\rho)$. By virtue of (2.1),

$$\frac{d^2}{d\rho^2}\mathcal{I}_{N-1,\beta} v(\rho) = \sum_{l=2}^{N-1} \hat{v}_l (\beta^2 \sum_{k=0}^{l-2} (l-k-1)\mathcal{L}_k^{(\beta)}). \tag{3.10}$$

Set $c_l(\rho) = \beta^2 \sum_{k=0}^{l-2} (l-k-1)\mathcal{L}_k^{(\beta)}, \quad 2 \leq l \leq N-1$. Then

$$\begin{aligned} \frac{d^2}{d\rho^2}\mathcal{I}_{N-1,\beta} v(\rho) &= \sum_{j=0}^{N-3} (\sum_{l=2}^{N-1} a_j^l c_l(\rho)) v(\rho_{G,N-3,j}^{(2,\beta)}) \\ &+ (\sum_{l=2}^{N-1} a_0^l c_l(\rho)) v_0 + (\sum_{l=2}^{N-1} \widehat{a}_0^l c_l(\rho)) v_0^{'} \\ &= \sum_{j=0}^{N-3} b_j(\rho) v(\rho_{G,N-3,j}^{(2,\beta)}) + b_0(\rho) v_0 + \widehat{b}_0(\rho) v_0^{'}. \end{aligned} \tag{3.11}$$

Taking $\rho = \rho_{G,N-3,i}^{(2,\beta)}$ in (3.11),we deduce that

$$\begin{aligned} \frac{d^2}{d\rho^2}\mathcal{I}_{N-1,\beta} v(\rho_{G,N-3,i}^{(2,\beta)}) &= \sum_{j=0}^{N-3} b_j(\rho_{G,N-3,i}^{(2,\beta)}) v(\rho_{G,N-3,j}^{(2,\beta)}) \\ &+ b_0(\rho_{G,N-3,i}^{(2,\beta)}) v_0 + \widehat{b}_0(\rho_{G,N-3,i}^{(2,\beta)}) v_0^{'}, \quad 0 \leq i \leq N-3. \end{aligned} \tag{3.12}$$

Let

$$a_{i,j} = b_j(\rho_{G,N-3,i}^{(2,\beta)}), \quad 0 \leq i \leq N-3, \quad 0 \leq j \leq N-3,$$

$$A^{(N-2)} = (a_{i,j}),\ d^{(N-2)} = (b_0(\rho_{G,N-3,0}^{(2,\beta)}), \cdots, b_0(\rho_{G,N-3,N-3}^{(2,\beta)}))^{\top},$$

$$g^{(N-2)} = (\widehat{b}_0(\rho_{G,N-3,0}^{(2,\beta)}), \cdots, \widehat{b}_0(\rho_{G,N-3,N-3}^{(2,\beta)}))^{\top},$$

$$G^{(N-2)} = (G_{\beta,1}(\rho_{G,N-3,0}^{(2,\beta)}), \cdots, G_{\beta,1}(\rho_{G,N-3,N-3}^{(2,\beta)}))^{\top},$$

$$\begin{aligned} F^{(N-2)} &= (f(v(\rho_{G,N-3,0}^{(2,\beta)}), v^{'}(\rho_{G,N-3,0}^{(2,\beta)}), \rho_{G,N-3,0}^{(2,\beta)}), \cdots, \\ &\quad f(v(\rho_{G,N-3,N-3}^{(2,\beta)}), v^{'}(\rho_{G,N-3,N-3}^{(2,\beta)}), \rho_{G,N-3,N-3}^{(2,\beta)}))^{\top}, \end{aligned}$$

$$V^{(N-2)} = (v(\rho_{G,N-3,0}^{(2,\beta)}), \cdots, v(\rho_{G,N-3,N-3}^{(2,\beta)}))^{\top}.$$

Obviously, by the definition of $\mathcal{I}_{N-1,\beta}$

$$\mathcal{I}_{N-1,\beta}\frac{d^2}{d\rho^2} v(\rho_{G,N-3,i}^{(2,\beta)}) = f(v(\rho_{G,N-3,i}^{(2,\beta)}), v^{'}(\rho_{G,N-3,i}^{(2,\beta)}), \rho_{G,N-3,i}^{(2,\beta)}), \quad 0 \leq i \leq N-3.$$

Then, (3.9) reads

$$A^{(N-2)} V^{(N-2)} = F^{(N-2)} + G^{(N-2)} - v_0 d^{(N-2)} - v_0^{'} g^{(N-2)} \tag{3.13}$$

Next,we construct the numerical scheme. To do this,we approximate $v(\rho)$ by $v_{N-1}(\rho)$. where $v_{N-1}(\rho) \in \mathcal{P}_{N-1}(\Lambda)$ and $v_{N-1}(0) = v_0, \frac{dv_{N-1}(0)}{d\rho} = v_0^{'}, v_{N-1}(\rho_{G,N-3,i}^{(2,\beta)}) = v(\rho_{G,N-3,i}^{(2,\beta)}), 0 \leq i \leq N-3$.

Clearly, $\mathcal{I}_{N-1,\beta} v_{N-1}(\rho) = v_{N-1}(\rho)$.

Let $\vec{v}_{N-2} = (v_{N-1}(\rho_{G,N-3,0}^{(2,\beta)}), \cdots, v_{N-1}(\rho_{G,N-3,N-3}^{(2,\beta)}))^{\top}$,

$$\ddot{f}_i = f(v_{N-1}(\rho_{G,N-3,i}^{(2,\beta)}), v_{N-1}^{'}(\rho_{G,N-3,i}^{(2,\beta)}), \rho_{G,N-3,i}^{(2,\beta)}),$$

$$\vec{F}_{N-2}(\vec{v}_{N-2}) = (\ddot{f}_0, \cdots, \ddot{f}_{N-3})^{\top}.$$

We derive the following spectral scheme for (3.7)

$$A^{(N-2)} \vec{v}_{N-2} = \vec{F}_{N-2}(\vec{v}_{N-2}) - v_0 b^{(N-2)} - v_0^{'} g^{(N-2)}. \tag{3.14}$$

Obviously, the system (3.14)is equivalent to

$$\begin{cases} \dfrac{d^2 v_{N-1}(\rho_{G,N-3,i}^{(2,\beta)})}{d\rho^2} = f(v_{N-1}(\rho_{G,N-3,i}^{(2,\beta)}), v_{N-1}^{'}(\rho_{G,N-3,i}^{(2,\beta)}), \rho_{G,N-3,i}^{(2,\beta)}), \\ \qquad\qquad 0 \leq i \leq N-3, \\ v_{N-1}(0) = v_0, \dfrac{dv_{N-1}(0)}{d\rho} = v_0^{'}. \end{cases} \tag{3.15}$$

Next, we analyze the numerical error of (3.14). To do this, let $E_{N-1} = v_{N-1} - \mathcal{I}_{N-1,\beta} v$. Next,we assume that there exists a real number γ such that

$$|f(z_1, \tilde{z}_1, \rho) - f(z_2, \tilde{z}_2, \rho)| \leq \gamma(|z_1 - z_2|) + |\tilde{z}_1 - \tilde{z}_2|) \tag{3.16}$$

Under this condition, we obtain the following Theorem

Theorem 3.1. If v belongs to $H_{\omega_{\frac{\beta}{2}}}^r(\Lambda, \alpha)$, taking β such that $\beta^2 > \delta$,by Theorem 2.5 and Theorem 2.7,

$$\begin{aligned} \|E_{N-1}\|_{\omega_\beta} &\leq c(N^{\frac{3-s}{2}} \|\frac{d^2 v}{d\rho^2}\|_{s,\omega_{\frac{\beta}{2}},\alpha} + N^{\frac{7-s}{2}} \|v\|_{s,\omega_{\frac{\beta}{2}},\alpha} \\ &\quad + N^{\frac{6-s}{2}} \|v\|_{s,\omega_{\frac{\beta}{2}},\alpha} + N^{\frac{6-s}{2}} \|\frac{dv}{d\rho}\|_{s,\omega_{\frac{\beta}{2}},\alpha}). \end{aligned}$$

Remark 3.1. For the validity of convergence of usual integration processes, we impose certain condition on the constant γ in (3.16). This limits its applications seriously. However, for any γ, we could use the scheme (3.14) with suitable parameter $\beta^2 > \delta$ to solve (3.7) efficiently. Therefore, our new method is available for a large class of dynamical systems.

Remark 3.2. Suppose that $f(z,t)$ fulfills the following Lipschitz condition

$$|f(z_1, \rho) - f(z_2, \rho)| \leq L|z_1 - z_2|, \quad L > 0.$$

Then we have the same error estimate as (3.14).

Remark 3.3. The scheme (3.14)is an implicit scheme. If $f(z,t)$ is nonlinear function for z, we need a nonlinear iteration to solve this system. Suppose that $f(z,t)$ fulfills the following Lipschitz condition

$$|f(z_1, \bar{z}_1, \rho) - f(z_2, \bar{z}_2, \rho)| \leq L(|z_1 - z_2| + |\bar{z}_1 - \bar{z}_2|), \quad L > 0.$$

Then iteration process is convergent.

IV. **Multiple-step generalized Laguerre-Gauss collocation method**

In the previous section,we introduced integration process for second ordinary differential quations.Theoretically,its numerical errors with bigger mode N is smaller.But in actual computation,it is not convenient to use very big mode.On the other hand, the distance between the adjacent interpolation nodes $\rho_{G,N-3,j}^{(2,\rho)}$ and $\rho_{G,N-3,j-1}^{(2,\rho)}$ increases fast as N and j increase,especially for the nodes which are located far from the origin point $\rho = 0$.This feature is one of advantge of the Laguerre interpolation,since we can use moderate mode N to evaluate the unknown function at large ρ.But it is also its shortcoming.In fact,if the exact solution oscillates or changes very rapidly between two large adjacent interpolation nodes,then we may lose information about the structure of exact solution between those nodes. To remedy this deficiency,we may refine the numerical result. For example, let N_0 is a moderate positive integer, and the set of nodes $\{\rho_{G,N_0-3,j}^{(2,\rho)}\}_{j=0}^{N_0}$. we use (3.14) with interpolation nodes $\{\rho_{G,N_0-3,j}^{(2,\rho)}\}_{j=0}^{N_0}$ to obtain the original numerical solution $U_{0,N_0}(\rho), 0 \leq \rho < +\infty$.Then we take $\rho_{G,N_1-3}^{(2,\rho)} = \rho_{G,N_0-3,N_0-3}^{(2,\rho)}$ and consider the problem

$$\begin{cases} \dfrac{d^2U_1(\rho)}{d\rho^2} = f(U_1(\rho), U_1^{'}(\rho), \rho), & \rho > \rho_{G,N_0-3,N_0-3}^{(2,\rho)} \\ U_1(\rho_{G,N_1-3}^{(2,\rho)}) = U_{0,N_0}(\rho_{G,N_0-3,N_0-3}^{(2,\rho)}), \\ \dfrac{dU_1(\rho_{G,N_1-3}^{(2,\rho)})}{d\rho} = U_{0,N_0}^{'}(\rho_{G,N_0-3,N_0-3}^{(2,\rho)}) \end{cases} \tag{4.1}$$

By a shifting argument and using with the parameter N_1 interpolation nodes $\{\rho_{G,N_1-3,j}^{(2,\rho)}\}_{j=0}^{N_1}$, we get the refined numerical solution $U_{1,N_1}(\rho), \rho_{G,N_1-3,0}^{(2,\rho)} \leq \rho < +\infty$,especially the values of$U_{1,N_1}(\rho)$ at the interpolation points $\rho_{G,N_1-3,j}^{(2,\rho)}, 0 \leq j \leq N_1$.Repeatint the above procedure, we obtain the refined numerical solution $U_{m,N_m}(\rho)$ for $\rho_{G,N_m-3,0}^{(2,\rho)} \leq \rho < +\infty$.This algorithm saves work and provides more accurate numerical results.In actual computation, we may take $\rho_{G,N_m-3}^{(2,\rho)} = \rho_{G,N_{m-1}-3,N_{m-1}-k_{m-1}}^{(2,\rho)}, k_{m-1} = 3, 4, 5$.

REFERENCES

[1] G.Szegö, *Orthogonal Polynomials*, AMS,Providence,RI,1959.

[2] J. Bergh and J. Löfström, *Interpolation Spaces, An Introduction*, Spinger-Verlag, Berlin, 1976.

[3] M.M. Chawla, M.A. Al-Zanaidi, S.S. Al-Ghonaim, Single-implicit stabilized extended one-step method for second order initial value problems with oscillating solutions, Math. Comp. Model. 29 (1999) 63-72.

[4] W. Gautschi, Numerical integration of ordinary differential equations based on trigonometric polynomials, Numer. Math. 3 (1961) 381-397.

[5] H.M. El-Hawary, S.M. Mahmoud, On some 4-point spline collocation method for solving second-order initial value problems, Appl. Numer. Math. 38 (2001) 223-236.

[6] E. Hairer, G. Wanner, A theory for Nyström methods, Numer. Math. 25 (1976) 383-400.

[7] Zhang Xiao-yong and Ning Tong-ke,Mixed spherical harmonic-generalized Laguerre spectral approximation and its application to Navier-Stokes equations set in exterior domain,ICCIS2010

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

An Efficient Approach for Construction Business Process Based on SOA

Jianfeng Sha
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications Beijing, China
shage001314@sohu.com

Budan Wu
State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications Beijing, China
wubudan@bupt.edu.cn

***Abstract*—Service-oriented business process construction is a technology that Web service components are directly mapped to the visualization graphical components. Business people are allowed to construct main logic structure of the business process through dragging and dropping the graphical components, then platform automatically generates BPEL elements and attributes by parsing WSDL .The BPEL elements and attributes which platform could not automatically generate will be specified by business people with help of user-friendly visual interface. In addition, it's necessary to display available variables for business people in order to copy data from one variable to another. However, when the scale of business process is very large, it's unfeasible to show all available variables for business people. To solve this problem, in this paper, we propose and prove two theorems used to filter out some variables, and the remaining variables are sorted in accordance with similarity among the variables so that make business people are able to quickly navigate the variables needed. In this way, platform hides some details of BPEL language for business people. Therefore, as long as business people are familiar with the business knowledge, they could use platform to construct business process that meets the needs of business. Thus speeding up the development efficiency of business process and reducing cycle of business process reconstruction.**

Keywords-Service-oriented; web service; business process; similarity

I. INTRODUCTION

The growing trend in enterprises is to assemble business systems from a set of appropriate web services and no longer be written from scratch. Service Oriented Architecture (SOA) has gained a lot of attention, so that more and more enterprises have started to encapsulate their software components as service components and created their business applications by composing service components. Business people could choose many different web services association language to reassemble web service components, different web services association languages not only have different requirements for technical and development capabilities of business people, but also directly affect the efficiency and cycle of development. The most popular language of web services association languages is BPEL language [1]. The business people could write BPEL codes to assemble web services by hand , but this method requires more technical and development capabilities of business people—business people must not only be familiar with business knowledge of domain-specify , but also must be fully know BPEL syntax details. Therefore, this method has low development efficiency and long development cycle.

Now there have been many visual development environment based on BPEL such as "BPELProject","ActiveBPEL","Oracle JDeveloper". Etc. Business people could make use of these platforms to construct main logic structure of the business process through dragging and dropping graphical components which are BPEL elements correspond to, besides, these platforms provide many user-friendly visual interfaces to set various values of BPEL elements attributes. Then these platforms could automatically generate BPEL codes, so speeding up development efficiency, reducing development cycle. However, these visual development environments provide a large number of graphical components and properties graphical views to map the elements and attributes of BPEL. These platforms not only don't hide the details of BPEL language for business people, but also business people have to select a lot of properties views and dialogs from the platforms to set a large number of attributes of BPEL elements. Thus it's limiting to further improve development efficiency.

In order to further improve development efficiency, this paper describes the techniques and methods based on service-oriented business process construction. our platform directly makes a mapping between web service components in asserts of domain-specify and the visualization graphical components, and could automatically generate BPEL elements and setting corresponding values of attributes by parsing WSDL [2], besides, providing a small amount of user-friendly interfaces to allow business people to set the values of BPEL elements attributes which the platform could not automatically generate. Thus not only further hiding some details of BPEL syntax, but also avoiding business people to set a large number of values of BPEL elements attributes. So business people can construct business applications, facing business and needs of domain-specify directly. Thus further improving development efficiency and reducing the development cycle.

The remainder of this paper is organized as follows. In Section 2, we discuss the method that generates BPEL elements used to invoke a web services. In Section 3, we propose and prove two theorems used to filter out the variables besides; we define similarity among variables and discuss the method of calculating similarity. Finally, we make conclusions in Section 4.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.22

II. INVOKING WEB SERVICE

The core to construct business process based on BPEL is invoking web services. The most basic BPEL element used to invoke a web service is "<invoke>". The main part of the "<invoke>" definition is shown in Figure 1.

```
<invoke partnerLink="NCName"
portType="QName"?
operation="NCName"
inputVariable="BPELVariableName"?
outputVariable="BPELVariableName"?
Standard-attributes>
    ……
</invoke>
```

Figure 1. <invoke> definition

From the above definition, to invoke a web service, the "partnerLink", "portType", "operation", "inputVariable" and "outputVariable" which are standard-attributes of "<invoke>" need to be set. Most of these attributes have the corresponding BPEL elements definition in BPEL or WSDL; so main function of these attributes is to refer to the definition of the corresponding BPEL or WSDL elements. The "<invoke>" and the elements that are referred by standard-attributes of "<invoke>" are the most basic and important in BPEL.

Communication among web services is implemented primarily through the messages exchange [4], where the BPEL element "<assign>" to implement message exchange among web services. Besides, the WSDL is needed to be imported with the BPEL element "<import>".

Therefore, in order to generate "<invoke>" automatically, firstly, must generate the elements as follows: "<import>", "<partnerLink>", "<variable>" and "<assign>". The next article will describe the methods generating these elements by parsing WSDL automatically.

A. Generating <import>

When to invoke a web service, needing to specify the location of WSDL which service-invoked correspond to. In the BPEL, the element "<import>" primarily is used to specify the location and namespace of WSDL, its definition is shown in Figure 2.

```
<import namespace="anyURI"?
location="anyURI"?
 importType="anyURI" />*
```

Figure 2. <import> definition

The Figure 2 shows that needing to set the attributes of the "<import>"as follows:"namespace", "location" and "importType".

1) *Setting "namespace"*
   The value of attribute "namespace" is equal to the value of attribute "tartgetnamespace" of WSDL [2]. So its value can be getting by parsing WSDL.
2) *Setting "location"*
   The value of attribute "location" specifies the physical address of WSDL, which can be getting from Database;
3) *Setting "importType"*
   Many types of documents can be imported into the current BPEL document such as XMLShema documents, WSDL documents and other BPEL documents etc. Therefore, the value of attribute "importType" is used to distinguish these documents. For WSDL document, its value also can be getting by parsing WSDL [2].

So the BPEL element "import" which is used to import WSDL can be automatically generated by parsing WSDL.

B. Selecting <portType>

The element "<portType>" is defined in the WSDL. The definition of the "<portType>" in WSDL is shown in Figure 3.

```
<wsdl:portType .... > *
    <wsdl:operation name="nmtoken"
     parameterOrder="nmtokens">
      <wsdl:input name="nmtoken"?
      message="qname"/>
      <wsdl:output name="nmtoken"?
      message="qname"/>
      <wsdl:fault name="nmtoken"
    </wsdl:operation>
 </wsdl:portType >
```

Figure 3. <portType> definition

In a WSDL, there can be allowed to define one or more than one "<portType>" .Therefore, SGP(Service Generating Platform) provides a dialog to show all "<portTypes>" defined in the WSDL so that business people can select one of these elements <"portType">. The "<portType>" selected recorded as "choosePortType".

C. Generating < variable>

If the "choosePortType" has "input" or "output" elements, as shown Figure 3. Then must generate corresponding BPEL elements "<variable>". The "<variable>" is used to define a variable; its definition is shown in Figure 4.

```
<variable name="BPELVariableName"
 messageType="QName"?
  type="QName"?
  element="QName"?>
 </variable>
```

Figure 4. <variable> definition

1) *Setting "name"*

The value of attribute name is able to any valid character string. However, it's convenient to be refered and readable. For the input-variable and output-variable of the web service, In our SGP, "name":=variable is inputVariable?"BPEL"+choosePortType.operation.input.name:"BPEL"+choosePortType.operation.output.name;

2) *Setting "messageType"*
"messageType" , for the input-variable and output-variable of the web service , "messageType":=variable inputVariable?choosePortType.operation.input.message :
choosePortType.operation.output.message;

D. Generating < partnerLink>

The BPEL element "<partnerLink>" is used to define a partner, its definition is shown in Figure 5.

```
<partnerLink name="NCName"
partnerLinkType="QName"
 myRole="NCName"?
partnerRole="NCName"?
initializePartnerRole="yes|no"? />
```

Figure 5. <partnerLink> definition

1) *Setting "name"*
The value of the "name" is able to any valid character string;
2) *Setting "partnerLinkType"*
The value of the "partnerLinkType" can be getting with "choosePortType", because every "<portType>" in WSDL has a corresponding "partnerLinkType",its definition is shown in Figure 6.

```
<partnerLinkType name="NCName">
    <role name="NCName"
    portType="QName" />
     <role name="NCName"
    portType="QName" />?
</partnerLinkType>
```

Figure 6. <partnerLinkType> definition

So the value of the "partnerLinkType" is equal to the value of attribute "name" of the WSDL element "<partnerLinkType>" whose "portType "is equal to "choosePortType";
3) *Setting "myRole" and "partnerRole"*
The values of the "myRole" and "partnerRole" are getting by parsing WSDL,
"myRole":=wsdl.partnerLinkType.role[1],
"partnerRole":=wsdl.partnerType.role[0];

E. Generating <invoke>

After completion of the above steps, now can generate the "<invoke>" shown as Figure 1.

1) *Setting "partnerLink"*
The value of the "partnerLink" of the "<invoke>" is equal to the value of attribute "name" of the "<partnerLink>" as D part of Section 2;
2) *Setting "portType"*
invoke.portType:=choosePortType.name;
3) *Setting "operation"*
operation,invoke.operation:=choosePortType.operation.name;
4) *Setting"inputVariable" , "outputVariable"*
The values of the "inputVariable" and "outputVariable" of the "<invoke>" is equal to the value of attributes "names" of BPEL element variable as C part of section 2;

Therefore, the BPEL elements-related used to invoke a web service could be automatically generated by parsing corresponding WSDL. No longer needing a large number of graphical interfaces to get user's input to generate those BPEL elements, thus speeding up the development efficiency of business process. In addition to above those steps, needing to generate a BPEL element "<assign>" to initialize the input-variable of the service-invoked.

III. SETTING INPUT-VARIABLE OF SERVICE-INVOKED

The input-variable of service-invoked must be assigned before service is invoked. In the BPEL, the element "<assign>" can be used to set input-variable of service-invoked through copying data from one variable to another. Many visual development environments provide the dialogs to show all available variables so that business people could select some variables to set input-variable of service-invoked. This approach is feasible for the small-scale business process, but when the scale of business process is very large, perhaps a business process contains hundreds of variables, showing all variables to be provided for business people to choose is unfeasible.

In fact, many of these variables could not be used to set the input-variable of service-invoked. In this paper, two theorems will be Proposed and proved to filter out those variables which could not be used to set the input-variable of service-invoked. In addition, in order to allow business people to quickly navigate to the variables used to set the input-variable of service-invoked. This paper proposes a concept of similarity among variables, and specifying the method of calculating. Candidate-Variables (Variables which are the remaining variables after filtering.) are sorted in accordance with similarity between Candidate-Variables and input-variable of the service-invoked.

A. The Theorems Of Variables-Filter

Definition 1: *Target-Variable*, the input-variable of service-invoked;
Candidate-Variable, a variable which may be used to set the Target-Variable;
CVS , represents a set of all *Candidate-Variable*s.
Definition 2: $Path-SS$, represents a set of all Services-Invoked on a path which is from Entry Service (such as receive) to current Service-Invoked in the business

process. $Path-SSS$, represents a set of all Path-SS. suppose the $Path-SSS$ set has n elements.

$Path-SS-VS-k$, represents a set of all input-variables and output-variables of elements in $Path-SS-k$, where $Path-SS-k \in Path-SSS(k=1,2,..n)$.

Theorem 1: $CVS \subseteq \bigcup_{k=1}^{n} Path-SS-VS-k$.

Proof. If $v \notin \bigcup_{k=1}^{n} Path-SS-VS-k$, it means variable v is not input-variable or output-variable of services-invoked on any paths which are from Entry Service (such as receive) to current Service-Invoked in the business process, hence when business process is running in the Execution Engine, no matter which path will be executed. For Target-Variable, the value of the variable v is undefined, so any $v \notin \bigcup_{k=1}^{n} Path-SS-VS-k$,could not be used to set Target-Variable. It means any variable cv , if $cv \in CVS$,then $cv \in \bigcup_{k=1}^{n} Path-SS-VS-k$,hence $CVS \subseteq \bigcup_{k=1}^{n} Path-SS-VS-k$.

Theorem 2 : $CVS \subseteq \bigcap_{k=1}^{n} Path-SS-VS-k$.

Proof. If $v \in \bigcup_{k=1}^{n} Path-SS-VS-k$ and $v \notin \bigcap_{k=1}^{n} Path-SS-VS-k$, then at least exists one set $Path-SS-VS-t(t=1,2,3,...n)$ $v \notin Path-SS-VS-t$.

Therefore, if business process is executed along the paths which v doesn't lie on, in this case, For Target-Variable, the value of the variable v is undefined, hence v could not be used to set Target-Variable. It means any variable cv, if $cv \in CVS$, then $cv \in \bigcap_{k=1}^{n} Path-SS-VS-k$,hence $CVS \subseteq \bigcap_{k=1}^{n} Path-SS-VS-k$.

For the Target-Variable, Theorem 1 filters out the variables which don't lie on the paths that are from Entry Service to the current service-invoked. Theorem 2 further filters out the variables which only lie on some parts of paths that are from Entry Service to the current service-invoked. In addition, it's easy to see that

$$CVS \subseteq \bigcup_{k=1}^{n} Path-SS-VS-k \subseteq \bigcap_{k=1}^{n} Path-SS-VS-k$$

.So the numbers of the CVS which are used to set the Target-Variable are reduced. Thus it's unnecessary to display all available variables on the dialog so that business people could quickly find the variables that used to set the Target-Variable. In order to more quickly navigate the variables that used to set the Target-Variable, the CVS could be sorted.

B. CVS Sorted

The variables are composed from the basic variables such as xsd:string,xsd:int, xsd:float, xsd:boolean and so on. So the BPEL variable can be considered as a vector.

Definition 3 (Variable Vector) : Suppose that v is a BPEL variable, bvs_v is a set of all basic variables of v , if $bvs\_v = \{bv_1, bv_2, bv_3, ...bv_n\}$ and $v = \{i_1.bv_1, i_2.bv_2, i_3.bv_3, ...i_n.bv_n\}$, $bv_k \in bvs\_v$, i_k represents the number of bv_k $(k=1,2,...n)$,if i_k is equal zero means that variable v dons' t contain basic variable bv_k .Then definition variable vector $\vec{v} = (i_1, i_2, i_3, ...i_n)$.

Definition 4 (Variables Similarity):

if

$bvs\_v1 = \{bv1_1, bv1_2, bv1_3, ...bv1_t\}$

$v1 = \{v1i_1.bv1_1, v1i_2.bv1_2, v1i_3.bv1_3, ...v1i_t.bv1_t\}$

$bvs\_v2 = \{bv2_1, bv2_2, bv2_3, ...bv2_m\}$

$v2 = \{v2j_1.bv2_1, v2j_2.bv2_2, v2j_3.bv2_3, ..v2j_m.bv2_m\}$

and

$bvs\_v = bvs\_v1 \cup bvs\_v2 = \{bv_1, bv_2, bv_3, ...bv_n\}$

then $\vec{v1} = (i_1, i_2, i_3 ...i_n)$,if $bv_k = bv1_h \in bvs\_v1$ then $i_k = v1i_h (1 \le k \le n, 1 \le h \le t)$,if $bv_k \notin bvs\_v1$,then $i_k = 0$. $\vec{v2} = (j_1, j_2, j_3 ...j_n)$,if $bv_k = bv2_h \in bvs\_v2$ then $j_k = v2j_h (1 \le k \le n, 1 \le h \le m)$ if $bv_k \notin bvs\_v2$, then $j_k = 0$.

The Similarity of $v1$ ***and*** $v2$:

$$vs(v1,v2) = \frac{\sum_{k=1}^{n} i_k j_k}{\sqrt{\sum_{k=1}^{n} i_k^2} \sqrt{\sum_{k=1}^{n} j_k^2}} = \frac{|i_1 j_1 + i_2 j_2 + i_3 j_3 + ...i_n j_n|}{\sqrt{i_1^2 + i_2^2 + i_3^2 + ...i_n^2} \sqrt{j_1^2 + j_2^2 + j_3^2 + ...j_n^2}}.$$

If there are five candidate-variables as fellows, $cv1 = \{2.\text{int}, 1.boolean, 3.string\}$

$cv2 = \{1.\text{int}, 3.float, 2.string\}$

$cv3 = \{3.\text{int}, 2.string\}$, $cv4 = \{2.date\}$

and $cv5 = \{1.\text{int}, 2.float, 1.string\}$.

Target-Variable is $tv = \{1.\text{int}, 2.float, 1.string\}$ The results of ***similarity*** between Target-Variable and Candidate-Variables are shown in Table 1.

TABLE I. EXAMPLE OF SIMILARITY

$\overrightarrow{cv1}(2,0,3,1)$	$\overrightarrow{tv}(1,2,1,0)$	$vs(cv1, tv) = 0.55$
$\overrightarrow{cv2}(1,3,2)$	$\overrightarrow{tv}(1,2,1)$	$vs(cv2, tv) = 0.98$
$\overrightarrow{cv3}(3,0,2)$	$\overrightarrow{tv}(1,2,1)$	$vs(cv3, tv) = 0.57$
$\overrightarrow{cv4}(0,0,0,1)$	$\overrightarrow{tv}(1,2,1,0)$	$vs(cv4, tv) = 0$
$\overrightarrow{cv5}(1,2,1)$	$\overrightarrow{tv}(1,2,1)$	$vs(cv5, tv) = 1$

Table 1 shows the more similar the Candidate-Variable and the Target-Variable, the greater ***similarity***. Therefore, the Candidate-Variables should be displayed in according to descending ***similarity*** so that the business people can quickly navigate to the variables which are used to set the Target-Variable. So the order of displaying for Candidate-Variables above should be as follows: $cv5 \rightarrow cv2 \rightarrow cv3 \rightarrow cv4$.

IV. CONCLUSION

In this paper, we present an efficient method to construct business process based on SOA, and implemented and applied in the Service Generating Platform (SGP). Especially two theorems and the concept of similarity among variables not only can be used to construct business process in the SGP, but also could be applied in other modules of the SGP such as Template-based Service Matching, Service Inquiry and Recommendation System.

ACKNOWLEDGMENT

This research is supported by the Novel Mobile Service Control Network Architecture and Key Technologies (Grant No.2010ZX03004-001), the Fundamental Research Funds for the Central Universities (Grant No. 2009RC0507), "973" program of National Basic Research Program of China (Grant No. 2011CB302704), and The National Natural Science Foundation of China (Grant No. 61003067).

REFERENCES

[1] Alexandre Alves, Assaf Arkin, Sid Askary, Ben Bloch, Web Services Business Process Execution Language Version 2.0, http://www.oasisopen.org/committees/ download.php/19511/wsbpel-specification-draft-Aug01-2006.htm#_Toc142302863

[2] Erik Christensen,Francisco Curbera,Greg Meredith,Sanjiva Weerawarana, Web Services Description Language (WSDL) 1.1,http://www.w3.org/TR/WSDL

[3] Budan Wu, Zhi Jin,Bin Zhao. A Modeling Approach for Service-Oriented Application Based on Extensive Reuse. Proceedings of IEEE International Conference on Web Services (ICWS2008), Beijing, 2008, pp.754-757

[4] Liu Ying, Wang Li,IBM China Research Lab,An Intelligent Service Composer for Business-level Service Composition,The 9th IEEE International Conference on E-Commerce Technology and the 4th IEEE International Conference on Enterprise Computing,E-Commerce and E-Services(CEC-EEE 2007)

[5] Chao Ma, Yangxiang He, An Approach for Visualization and Formalization of Web Service Composition, 2009 International Conference on Web Information Systems and Mining

[6] S. Staab, et al., Web services: Been there, done that? IEEE Intelligent Systems, IEEE Computer Society, Volume 18, Issue 1, Page(s): 72 – 85, 2003.

[7] B. Srivastava and J. Koehler, Web Service Composition-Current Solutions and Open Problems. In Proc. of ICAPS 2003 Workshop on Planning for Web Services, Page(s): 28-35, 2003.

[8] F. Casati and M. Shan, Dynamic and adaptive composition of e-services. Journal of Information Systems, Elsevier Science Ltd., Volume 26, Issue 3, Page(s): 143-163, 2001.

[9] Q. Sheng, et al. Self-Serv: A Platform for Rapid Composition of Web Services in a Peer-to-peer Environment. In Proc. of the 28th International Conference on Very Large Data Bases, 2002.

[10] Thomas Erl, "Service-Oriented Architecture (SOA):Concepts, Technology, and Design." The Prentice Hall Service-Oriented Computing Series, ISBN 0-13-185858-0.

[11] A. Pnueli, The temporal logic of programs, Proc. 18th IEEE Symposium on Foundation of Computer Science, pp. 46–57, 1977.

[12] Y. Liu, S. Müller, K. Xu, "A Static Compliance Checking Framework for Business Process Models". Special Issue on Compliance Management, IBM Systems Journal, 46(2). (2007)

[13] Fu, Yuchen; Jin, Tao; Ling, Xinghong; Liu, Quan; Cui, Zhiming, A multistrategy semantic web service matching approach ,2007 International Conference on Convergence Information Technology, ICCIT 2007, pp. 253-256

[14] Burstein MH, Hobbs JR, Lassila O, Martin D, McDermott DV, McIlraith SA, Narayanan S, Paolucci M, Payne T, Sycara K. DAML-S: Web service description for the semantic Web. In Proc. of the International Semantic Web Conference. Springer-Verlag, Page(s): 348~363, Sardinia, 2002

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Aspect-Oriented MDA Approach for Non-Functional Properties of Distributed Cyber Physical Systems

Lichen Zhang
Faculty of Computer Science and Technology
Guangdong University of Technology
Guangzhou 510090, China

***Abstract*—Cyber physical systems have many non-functional requirements, which always crosscut the whole system modules. That may cause the code tangle and scatter, make the systems hard to design, reuse and maintain, and affect performance of systems badly. AOP is a new software development paradigm, which could attain a higher level of separation of concerns in both functional and non-functional matters by introducing aspect, for the implementation of crosscutting concerns. Different aspects can be designed separately, and woven into systems. In this paper, we propose a four-stage method of aspect-oriented MDA for non-functional properties to develop cyber physical systems. The model- based development, aspect-oriented approach and formal methods are integrated effectively for the development of the non-functional properties of distributed cyber physical systems. A case study illustrates the aspect oriented MDA development of cyber physical systems.**

Keywords- Non-Functional Properties; Aspect-Oriented; MDA

I. INTRODUCTION

A cyber-physical system (CPS) is a system featuring a tight combination of, and coordination between, the system's computational and physical elements. Today, a pre-cursor generation of cyber-physical systems can be found in areas as diverse as aerospace, automotive, chemical processes, civil infrastructure, energy, healthcare, manufacturing, transportation, entertainment, and consumer appliances.Embedded computers and networks monitor and control the physical processes, usually with feedback loops where physical processes affect computations and vice versa. In the physical world, the passage of time is inexorable and concurrency is intrinsic. Neither of these properties is present in today's computing and networking abstractions .

Model Driven Architecture (MDA)[1] is a development method which can generate useable software products directly by the model. It includes a series of standardized modeling, transformation rules and other relevant standards architecture. In MDA, the software development process is driven totally by models; model played a very important role. At first you should model the business logic of the system at a high level of abstraction, the initial model is not concern with any real platform and environment, that is platform-independent model (PIM) and then it will be transformed into PSM through the mapping mechanism according to the actual software operating environment, it is platform and technology-related business logic model. Finally, PSM become to practical code through a lower mapping. Developers do not need to consider platform and technology details, the specific code will be automatically generated from these models.

Aspect-oriented software development methods [2] make up object-oriented software development methods in system development needs of non-functional characteristics of the existing limitations question problem. Use separate technology of concerns separates all the crosscutting concerns of the system, and then analyzed, designed, modeled for each cross-cutting concerns, to address crosscutting concerns in object-oriented software development, the code tangling and scattering problems, enhancing the system's modular degree, lowering coupling between modules.

In this paper, we propose a four-stage method of aspect-oriented MDA for non-functional properties to develop dependable and distributed real-time systems.

II. APPLYING AOP AND MDA TO NON-FUNCTIONAL REQUIREMENTS

The distributed systems software development process based on aspect-oriented system is divided in five phases[3].

The first phase is a profound analysis of the requirements. The phase includes three steps:

Step one handles the non-functional requirements and then identifies which of those are crosscutting.

Step two performs a traditional specification of functional requirements, in this case, using an UML-like approach where the use case model is the main specification technique.

Step three starts by composing functional requirements with aspects; then it identifies and resolves conflicts that may arise from the composition process.

The concepts of overlapping, overriding and wrapping can be adopted to define the composition part of the model. Overlapping indicates the requirements of the aspect modifies the functional requirements they transverse. In this case, the aspect requirements may be required before the functional ones, or they may be required after them. Overriding indicates the requirements of the aspect superpose the functional requirements they transverse. In this case, the behavior described by the aspect requirements substitutes the functional requirements behavior. Wrapping indicates the requirements of the aspect encapsulate the

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.26

functional requirements they transverse. In this case, the behavior described by the functional requirements is wrapped by the behavior described by the aspect requirements.

In the design phase, the distributed system will be designed considering both the requirements and the constraints posed by the system and system. Using the MDA approach to produce the platform specific models includes five steps (see Fig. 1):

Step one: Create the PIM for the distributed system.

Step two: Select the target system and create the generic system aspects.

Step three: Transform PIM to enhanced PIM using the application converter.

Step four: Transform the generic aspects to enhanced aspects using the aspect converter.

Step five: Weave the enhanced aspects into the enhanced PIM to produce the PSM.

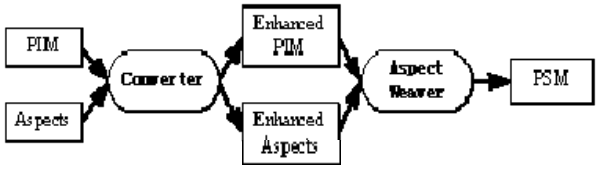


Figure 1. Process view of the PSM generation

The validation phase is in charge of validating the application design against both functional and non-functional models. Also in this phase, the system characteristics have to be considered since they can affect the application validation. Model-based analysis techniques can be used for validation purposes . Because it provides a way for the design-time analysis of distributed systems enabling rapid evaluation of design alternatives with respect to given performance measures before committing to a specific platform.

In the development phase, the source code of classes and aspect is generated, and the distributed system is built on top of the aspect-oriented system platform.

Since aspect may affect the behavior of one or more classes through advice and introduction. Traditional testing techniques, such as unit testing, integration testing, are not applicable to test aspect in the testing phase. Some aspect-oriented testing approaches, such as data-flow-based unit testing, state-based testing approach, and model-based testing approach can be used in this phase.

To address the system development, principled methods are needed to specify, develop, compose, integrate, and validate the application and system software used by DRE systems. These methods must enforce the physical constraints of DRE systems, as well as satisfy the system's stringent functional and non-functional requirements. Achieving these goals requires a set of integrated Model Driven System (MDM) tools [5]that allow developers to specify application and system requirements at higher levels of abstraction than that provided by low-level mechanisms, such as conventional general-purpose programming languages, operating systems, and system platforms. Different functional and systemic properties of DRE systems via separate system and platform-independent models are applied.[4] . Through profound study of model driven architecture and aspect-oriented technology[7], the idea of aspect-oriented is introduced into MDA, the existing three-stage MDA development process[8] to increase one aspect of the computation independent model stage (ACIM), proposed four-stage method of aspect-oriented MDA for system' non-functional properties to develop. Four stages are: ACIM stage, APIM stage, APSM stage, aspects of the code model stage. ACIM stage uses the UML use-case modeling techniques]to establish aspects of computation independent model that is also called the business model, which is a effective tool of communicating with personnel, more clearly stating the system non-functional properties, reducing the complexity of establishing APIM; Modeling language AOUML and AJUML are defined by the double expansion method based on UML[9], which combines the object constraint language OCL[10] so that they are used for establishing APIM, APSM. As the two-stage model description language is the level of the UML meta-model extensions, mapping each other between modeling language, therefore, use the element model conversion method for model conversion, the specific technology platform used aspect-oriented software development language-AspectJ; the last stage, can use auxiliary tools to convert APSM into code; vice versa ,code can be converted into APSM . the code itself is a model and the inverter can be achieved to the APSM.

III. CASE STUDY: INTELLIGENT TRANSPORTATION SYSTEMS

Intelligent Transportation systems(ITS)[6] – automotive, aviation, and rail – involve interactions between software controllers, communication networks, and physical devices. These systems are among the most complex cyber physical systems being designed by humans, but added time and cost constraints make their development a significant technical challenge. MDA approach can be used to improve support for design, testing, and code generation. MDA approach is increasingly being recognized as being essential in saving time and money. Transportation systems consist of embedded control systems inside the vehicle and the surrounding infrastructure, as well as, the interaction between vehicles and between vehicle and the

infrastructure. The dynamics of these interactions increases complexity and poses new challenges for analysis and design of such systems. ITS CIM mode is shown in Fig.2[6].

In cyber-physical systems (CPS) such as ITS, the passage of time becomes a central feature — in fact, it is this key constraint that distinguishes these systems from distributed computing in general. Time is central to predicting, measuring, and controlling properties of the physical world: ITS CIM The modeling process of non functional requirements time of ITS by aspect –oriented MDA is shown as Fig.3, Fig.4 , Fig 5., Fig 6,Fig 7and Fig.8.

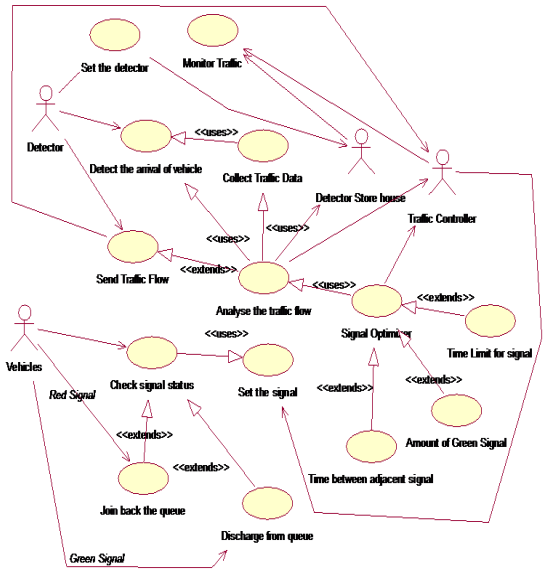


Figure.2 ITS CIM: Extended Use case Diagram

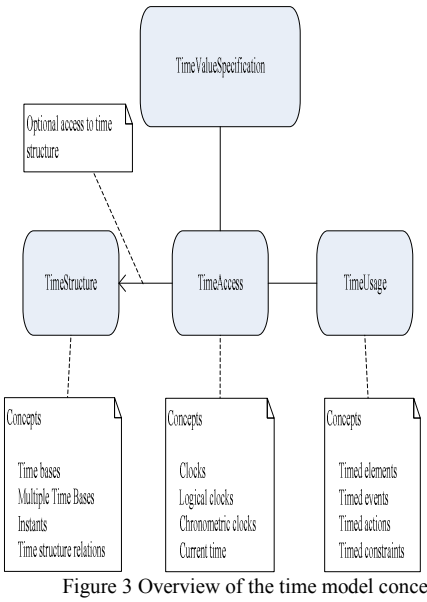


Figure 3 Overview of the time model concerns

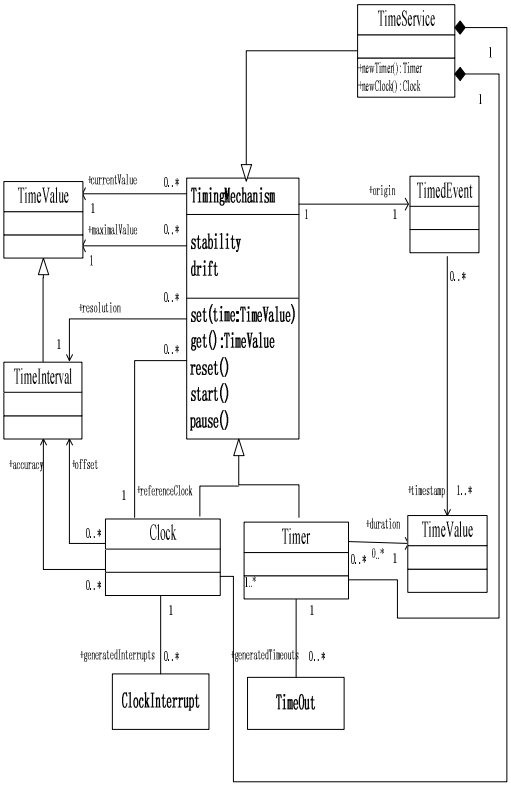


Figure 4 Time CIM Mode

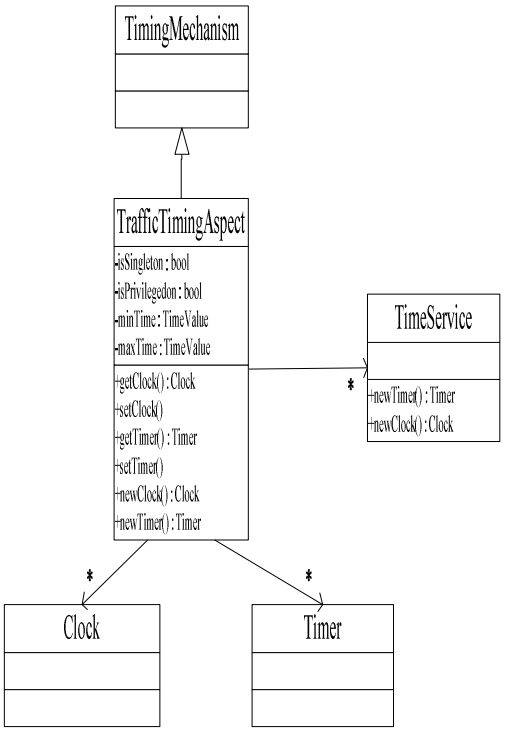


Figure 5 Time PIM Model

Context TrafficTimingAspect:

inv: *--the isSingleton is not equal isPrivileged*

if self.isSingleton then self.isPrivileged=false

else self.isPrivileged=true endif

self.isSingleton=not self.isPrivileged

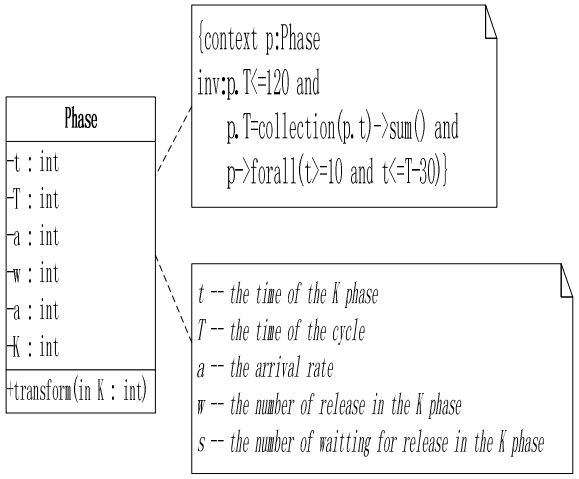
--maxTime always large than minTime

self.maxTime>=self.minTime

The time restriction of phase in road crossing is as following.

$$\begin{cases} 10 \le t_1^{ij} \le T-30 \\ 10 \le t_2^{ij} \le T-30 \\ 10 \le t_3^{ij} \le T-30 \\ 10 \le t_4^{ij} \le T-30 \\ T = t_1^{ij} + t_2^{ij} + t_3^{ij} + t_4^{ij} \\ T \le 120 \end{cases}$$

The time restriction of phase is specified by OCL as following:



Figue 6Ttime Restriction of phase

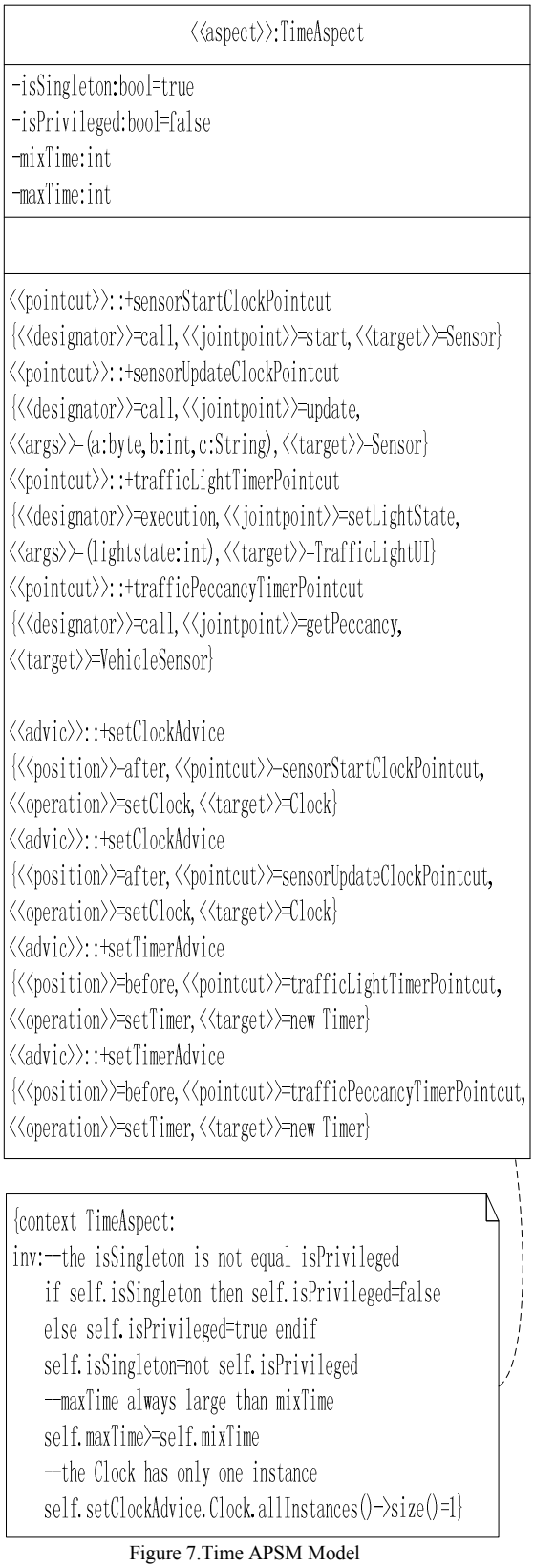


Figure 7.Time APSM Model

```
package com.aspect.its;

public aspect TimeAspect {
    private boolean isSingleton = true;
    private boolean isPrivileged = true;
    private int mixTime;
    private int maxTime;

    public boolean isSingleton() {
        return isSingleton;
    }

    public void setSingleton(boolean isSingleton) {
        this.isSingleton = isSingleton;
    }

    public boolean isPrivileged() {
        return isPrivileged;
    }

    public void setPrivileged(boolean isPrivileged) {
        this.isPrivileged = isPrivileged;
    }

    public int getMixTime() {
        return mixTime;
    }

    public void setMixTime(int mixTime) {
        this.mixTime = mixTime;
    }

    public void setMixTime(int mixTime) {
        this.mixTime = mixTime;
    }

    public int getMaxTime() {
        return maxTime;
    }

    public void setMaxTime(int maxTime) {
        this.maxTime = maxTime;
    }
   pointcut sensorStartClockPointcut():
       call(* Sensor.start(..));

   pointcut sensorUpdateClockPointcut(byte a, int b, String c):
       call(* Sensor.update(byte ,int,String))&&args(a,b,c);

   pointcut trafficLightTimerPointcut(int lightstate):
       execution(* TrafficLightUI.setLightState(int))&&args(lightstate);

   pointcut trafficPeccancyTimerPointcut():
       call(* VehicleSensor.getPeccancy(..));

   after():sensorStartClockPointcut(){
       Clock.setClock();
   }

   after(byte a, int b, String c):sensorUpdateClockPointcut(a,b,c){
       Clock.setClock();
   }

   before(int lightstate):trafficLightTimerPointcut(lightstate){
       new Timer().setTimer();
   }

   before():trafficPeccancyTimerPointcut(){
       new Timer().setTimer();
   }
}
```

Figure 8. time ISM Model

IV. CONCLUSION

In this paper, we proposed a four-stage method of aspect-oriented MDA for non-functional properties to develop real-time cyber physical systems. We illustrated the proposed method by the development of ITS and demonstrated four-stage method of aspect-oriented MDA that can be used for modeling complex system' non-functional characteristics, effectively reduce the complexity of software development and coupling between modules to enhance the system's modular.

The further work is devoted to developing tools to support the automatic generation of model and code.

ACKNOWLEDGMENT

This work is supported by the Major Program of National Natural Science Foundation of China under Grant No.90818008.

REFERENCES

[1] Object Management Group.OMG MDA guide v1.0.1[EB/OL].http://www.omg.org/docs/omg/03-06-01.pdf.

[2] Wehrmeister, M.A., Freitas, E.P., and Pereira, C.E., et al., "An Aspect-Oriented Approach for Dealing with Non-Functional Requirements in a Model-Driven Development of Distributed Embedded Real-Time Systems ", 10th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing, Santorini Island, Greece, May7-9, 2007, IEEE Computer Society, pp.428-432.

[3] Jingyong Liu, Yong, Zhong, Lichen Zhang, Yong Chen. Applying AOP and MDA to middleware-based distributed real-time embedded systems software process. 2009 Asia-Pacific Conference on Information Processing, APCIP 2009, pp 270-273, 2009.

[4] Liming Zhu,Gorton, I. ,"UML Profiles for Design Decisions and Non-Functional Requirements",Second Workshop on Sharing and Reusing Architectural Knowledge - Architecture, Rationale, and Design Intent, 20-26 May 2007 pp.8-9.

[5] D. S., "Model Driven Architecture:Applying MDA to Enterprise Computing", OMG Press.

[6] K.Ranjini, A.Kanthimath and Y.Yasmine. Design of Adaptive Road Traffic Control System throughUnified Modeling Language. International Journal of Computer Applications, Volume 14– No.7, February 201

[7] Clemente, P. J., Sánchez, F., and Perez, M. A. "Modelling with UML Component-based and Aspect Oriented Programming Systems",Seventh International Workshop on Component-Oriented Programming at European Conference on Object Oriented Programming (ECOOP). Málaga, Spain.,2002,pp.1-7.

[8] G. Madl, S. Abdelwahed. "Model-based Analysis of Distributed Realtime Embedded System Composition", Proceedings of the 5th ACM international conference on Embedded software, New Jersey, USA , 2005.

[9] Weis, T., Becker, C., Geihs, K., Plouzeau, N., "A UML Meta-model for Contract Aware Components", Springer LNCS 2185, 2001

[10] Lavazza, G. Quaroni,M. Venturelli, "Combining UML and formal notations for modeling real-time systems", ACM SIGSOFT Software Engineering Notes, vol. 26, Sep. 2001, pp.196-206

The design and implementation of HART intelligent instrument monitoring and performance analysis system

Liuxiaoyu, Wanghongyuan[i]* Pancao Gaojinshu Xvjing
School of Information science and engineering, Changzhou University, Changzhou, 213164
E-mail: hywang@cczu.eud.cn

ABSRTACT: HART intelligent instrument system complete Online monitoring and alarm through the com server, moreover, it also researches the locator's performance analysis and fault diagnosis. Field application shows that the system steady precision, dynamic response characteristics and etc can get the predetermined requirements.

Keywords: HART intelligent instrument Online monitoring and alarm

1 Introduction

With the development of the times, and the advancement of science and technology, Industrial production control mode is developed from a centralized control to a decentralized control. It is inevitable to make the site equipment of process industry digitizing and webifying, so the intelligent site instruments which are based on several micro controllers will be emerged. But intelligent instrument only use fieldbus for digital communication can display its outstanding performance. The fieldbus technology has many advantages such as managing wiring fee, operation flexible, the implementation of network management and interoperability, So it has become acknowledged direction in the process control. Users have input huge funds in this system structure which connects traditional simulation type field device and DCS, while the fieldbus control system cannot completely replace DCS system which are developed more perfect. So it is necessary to adopt a communication protocol which can use digital signal and is compatible with analog signals transitional communication protocol in the transitional period. HART

(Highway Addressable Remote Transducer) communication protocol can meet this need ,and It has become the industry standard . HART technology has strong vitality in the early 21st century .

2 System realization principle

2.1 system structure

This system mainly consists of intelligent instrument, HART agreement communication devices, HART Server communication software (including COM Server and FIOS communication software), database (including SQL Server and force charged with real-time database), PC management software (including application Server and client). Intelligent instrument is mainly based on HART agreement, such as discriminating

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.90

pressure transmitter, pressure transmitter, valves locator, flow instrument and etc. HART protocol communications device is responsible to collect and transmit digital signals in intelligent instrument, then the digital signals are transmitted to of the PC through Adam module, after that the COM server software analyze this data, and storing the collecting data in real-time database, and other automation system can remotely access the data of field instrument via FIOS communication software.The application server is to configure and monitor intelligent instrument, providing visualization for user interface which make user operations conveniently. The client is responsible to publish monitoring information and connections in internet.

2.2 parts of pneumatic or electric valves locator

(1) Input device.

In practical project, commonly using 4 - 20mA current as systems input signal, input devices will detect input signals and input the signals into comparison unit, sometimes electrical valves locator still need electrical switching device.

(2) Comparison unit.

Comparison unit compares the input signal and the valve position feedback signals, if it detected the difference of these two signals, comparison unit will drive valve actuator move until feedback signal matches to the input signal.

(3) Feedback device.

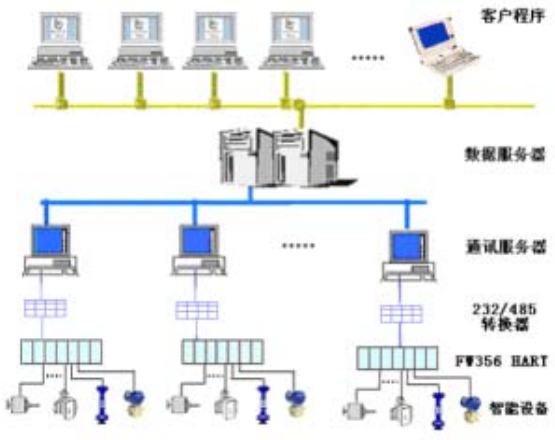
Feedback device detect the changes of valve position and transfer the position of valve rod to comparison unit when the signals through the sensor conversion.

2.3 The operating principle of intelligent locator

Valves locator is based on the principle of closed-loop control. It can regulate the valve according to compared deviation which got from the difference between control signals from sensor (4-20mA current signal) and the position signals of valve rod from feedback rod, thereby the position of valves core is more accurate and is correspond to the position of control signal.

3 System function

HART communication server software includes COM server and FIOS communication software. COM server realize the serial communication with HART Netbridges, and send the read command of basic information of instrument and process variables to HART Netbridges, at the same time receive and process instrument process variables and alarm information from HART Netbridges; It also support forward-down special HART agreement which can realize the configuration and monitoring for site HART instrument. And support the send special HART treaty order to HART Bridges to realize configure and monitor HART instruments. FIOS communication software can send the signal of HART instrument to the



real-time database. The system connection of HART agreement communication system as follows:

The explanation of the communication server interfaces shows in the table1. Table 1 Explanation for Module Interfaces

Interface function name	Explanation for interface function	Module name
GetDevicesInfo	Get the note list of devices' configuration information	IHartequip
SetEquiTSD	Set the equipments' ID, description, configuration date	IHartequip
SetTopLimi	Set the upper limit of range (put the equipments' current PV value as the upper limit of range)	IHartequip
SetPVTranFun	Set the transmitting function	IHartequip
SetPVScale	Set the range	IHartequip
Initialize	Initialize the bridge and the communication server	IHartequip
AddDevice	Notify the application server of adding equipments	IHartequip
NeedRefreshOnline	Notify the application server of refreshing on-line equipments' configuration	IHartequip
AlarmChange	Notify the application server of equipments' alarm state change	IHartequip
Transfunc	The special channel for transmitting HART commands	IHartequipEvent
Fire_OnDogErr	Notify the application server of whether the commands have been transmitted successfully	IHartequipEvent

HART agreement includes three orders, general command, common command and special orders. General command is applicable to the whole intelligent equipment that follow HART agreement, It includes reading manufacturer and equipment types, reading principal variable value and unit, reading dynamic variable value and the main variable current; General command is used for common operations, it is applicable to the most HART intelligent device, but not all, such as reading choice of process variables, writing principal variable range value, not principal variable overrun and etc. Special orders are based on specific applications. It is designed by equipment manufacturers. The most components interface of COM server is based on general command, and general commands and special commands can send and analysis directly via HartTrans component and Transfunc interface.

Transfunc interface:

03	1A	00 00	BF 06 60 00	39	41 3F A0 00	39	42 47 60 00	39	41 3F A0 00	27	41 95 00	00	D4

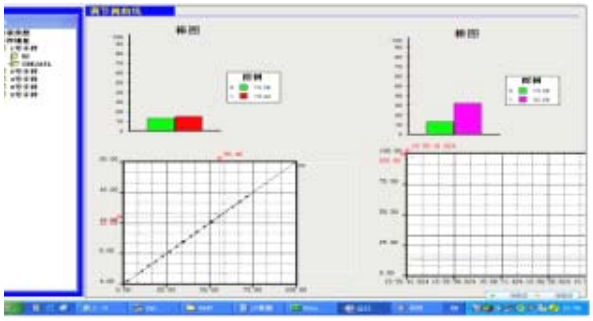
Transfunc interface can send command, receive data directly and analyze data in the end. The data collected from intelligent instrument is the binary code at the beginning. It necessary to transfer the data to decimalism according to the data link layer frame format and IEEE 754 standard, but each HART command format is not identical, so separate analysis for each command is inevitable. When the third command is send through special channels to instrument, the data as follows:

03 is command number, 1A is burst mode. The

bytes " 4247 60 00 ", "BF 06 60 00" and "41 3F A0 00 "which is behind to state bytes("0060") ,They indicate "SV ","TV " and "FV" , 39 is"%". (PV stand for principal variable which is principal variables in operation of HART instrument, such as the pressure measuring values of Pressure transmitter. Some HART protocol instrument can also provide the second variable SV, third variable TV and even fourth variables, such as the second variable of 3051 pressure transmitter is sensor temperature. The second, third, and fourth variable are designed based on different machine types. The first command only reads PV value, the third command reads all variables values).”413F A000” signify the current value which was 11.9766 after calculation, and "27" stand for mA. The result is the same as Hand hold device.

For intelligent locator, just get the main variables PV is far from enough, the working state of control valve will be got when SV got as well. Then we get relationship of the input current and PV, after test and verify, the result shows that the input current and feedback value keep a linear relationship.

The theory of locators' diagnosis and performance analysis : for each intelligent locator which will be put into use, we backup a standard curve at beginning, and set the standard curve as the future locator calibrations masterplate,period of time passed, if the locator's SV/PV curve is different from standard curve, the system will remind operators to overhaul the locator. The following is SV/PV curve, the horizontal axis is PV (the area from lower limit to upper limit), vertical axis is input current (the area is 4-20, the unit is mA).



4 Conclusion

HART intelligent instruments system has many advantages, such as the operation of it is sample and convenient, and the locator's performance analysis can found fault of control valve in time and raise the alarm , it ensure produce process safely and smoothly. Then the automation degree and office efficiency are improved in enterprise.

References

[1] HART-SMART Communications Protocol. Common Tables. Document Revision: 9.1

[2]Liu Zheng.The Design of Intelligent Valve Actuator Controller Based on ATmegal28.Engineering and Automation, Beihang University, Bejing l0,2007.

[3]Zhong shen-hui . Research and Development of HVP Series Smart Valve Positioner . PROCESS AUTOMATION INSTRUMENTATION Vol . 28 Supplement Issue September 2007.

[4] Fisher-Rosemount Systems Inc. HART-SMART Communications Protocol, FSK Physical Layer Specification. Document Revision: 7.0

[5] HART-SMART Communications Protocol. Data Link Layer Specification. Document Revision: 7.1

[6] Kenneth L. Holladay. Calibrating HART transmitters. USA: www.hartcomms.org, 2002

Software Design and Implementation for Obtaining ZigBee Network Structure Information

Wenxue Li[a], Zhaomin Zhu, Xiaofeng Gu[b]
Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education)
Jiangnan University
Wuxi, China
a. zaijiangda@163.com; b. xgu@jiangnan.edu.cn

Lei He
Suzhou Institute of Nano-tech and Nano-bionic
Chinese Academy of Sciences
Suzhou, China
lhe2009@sinano.ac.cn

Abstract—In this paper, we have developed a new method for ZigBee network structure to get timely information on the network, including the network address and the parent-child relationship of all the nodes. This method has been verified and implemented into the software program.

Keywords-ZigBee; distributed addressing scheme; network structure tree

I. INTRODUCTION

Recently, ZigBee technique has been extensively used in wireless sensor networks (WSN) due to its low power loss, low cost and high capacity. Single ZigBee network can contain up to 65,536 nodes. In real applications such as the streetlight control system and the environmental monitoring system, large numbers of nodes need to be deployed. This paper has developed a new algorithm for network structure. With the help of this new algorithm, we can obtain timely information of the network, including the network address, the parent-child relationship of all the nodes, and display it in the software program finally.

Network structure mentioned in this paper is based on the distributed addressing scheme of ZigBee protocol, as shown in Figure 1, which is different from the network topology. ZigBee network topology is the routing path of data transmission, including mesh, tree and star topologies. The tree structure illustrated in Figure 1 shows all the network nodes, where the top node is the parent of lower nodes. For example, nodes of 0x0001, 0x143E, 0x287B, etc. are the children of the coordinator node 0x0000, and nodes of 0x0002, 0x035F, etc are the children of the node 0x0001.

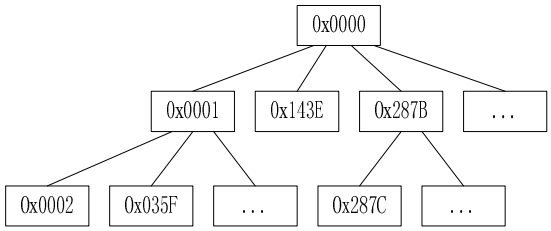


Figure 1. ZigBee network structure

From the above network structure obtained by our algorithm, we can visually see all nodes in the network and their relations, and configure and manage these nodes. When a node joining or leaving the network, the software program can automatically refresh the tree and update the network structure.

II. ZIGBEE DISTRIBUTED ADDRESSING SCHEME

There are three types of node equipments in the ZigBee network: coordinator, router and end-device. The coordinator is responsible for establishing and maintaining the network; the router nodes can transmit data, allowing the child nodes to join and assign network address; the end-device can only send and receive data, not allowing the joining of child nodes.

The value of the NIB's (Network layer Information Base) attribute nwkAddrAlloc is 0x00, where network addresses are assigned using a distributed addressing scheme that is designed to provide every potential parent with a finite sub-block of network addresses. These addresses are unique within a particular network and are given by a parent to its children. The entire network addresses are assigned by the following parameters:

1. The maximum depth of the network (L_m);
2. The maximum number of children that a parent may have (C_m);
3. The maximum number of routers that a parent may have (R_m).

We can compute $C_{\text{skip(d)}}$, essentially the size of the address sub-block being distributed by each parent at that depth to its router-capable child devices for a given network depth d, as follows:

$$C_{skip(d)} = \begin{cases} 1 + C_m(L_m - d - 1) & R_m = 1 \\ \dfrac{1 + C_m - R_m - C_m R_m^{L_m - d - 1}}{1 - R_m} & R_m \neq 1 \end{cases} \quad (1)$$

A parent assigns an address that is 1 greater than its own to its first router-capable child device. Subsequently assigned addresses to router-capable child devices are separated from each other by $C_{skip(d)}$. Addresses assigned to its end-device children by a parent are continuous. The n^{th} end-device child address is given by the following equation:

$$A_n = A_{parent} + C_{skip(d)d} R_m + N \ . \quad (2)$$

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.82

A_{parent} is the parent's address. After the values of the three parameters L_m, C_m, R_m are determined, ZigBee network will assign an address to each new node. In this paper, the following experimental parameters are used: L_m=5, C_m=20, R_m=20. By substituting the above data into equation (1), we get $C_{skip(0)}$=0x143D, $C_{skip(1)}$=0x35D, $C_{skip(2)}$=0x8D, $C_{skip(3)}$=0x15, $C_{skip(4)}$=0x01. The network structure is the same as shown in Figure 1.

III. ALGORITHM DESIGN

A. ZigBee network design

When the coordinator receives command of inquiring network structure through the serial port, it will give three actions: 1. it returns the network information, including the PAN (Personal Area Network) ID and the value of L_m, C_m, R_m; 2. It broadcasts the command to all the equipments in the network; 3. It receives response information from all the equipments, and transmits it to PC.

The flow chart of the ZigBee network is shown in Figure 2.

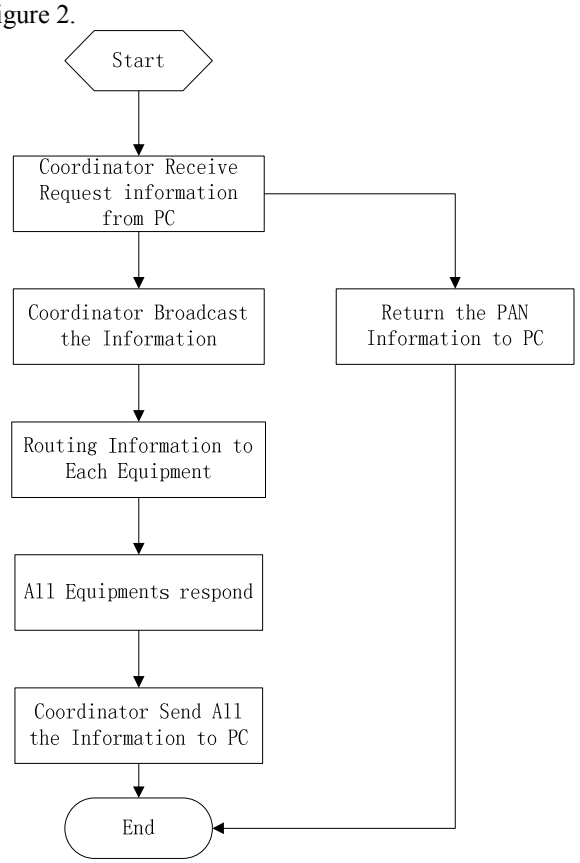


Figure 2. Flow chart of ZigBee network

B. Software Design

The PC sends command to the ZigBee coordinator through the serial port. The command format is shown in Table 1.

TABLE 1. Request command frame

SOP	CMD		Length	Destination address		Cluster ID		FCS
02	00	18	04	FF	FF	11	00	0D

0xFFFF expresses the destination address that is broadcast. Cluster ID defines the function of the command. FCS is used to ensure packet integrity, and this field is computed as an XOR of all the bytes in the message starting with CMD and through the last byte of data.

The coordinator responds network information, as is shown in Table 2.

TABLE 2. Response from coordinator

SOP	CMD		Length	PAN ID		L_m	C_m	R_m	FCS
02	10	18	05	12	13	05	14	06	1B

All the nodes in the network receive the command, and send a response message to the coordinator, which transmit the message to PC through the serial port. Response message format is shown in Table 3.

TABLE 3. Response from node

SOP	CMD		Len	Addr		Node type	Father addr.		Cluster ID		FCS
02	10	18	07	B9	3C	02	B8	3C	12	00	1E

This message includes the network address of the node itself, the parent address and node type. 0x02 stands for router, and 0x03 stand for the end-device.

The command and response message transmission is shown in Figure 3. The coordinator broadcasts after receiving command from PC. Then the router and end-device nodes respond. Routers can transmit the messages when the message can't be sent to the coordinator by one hop.

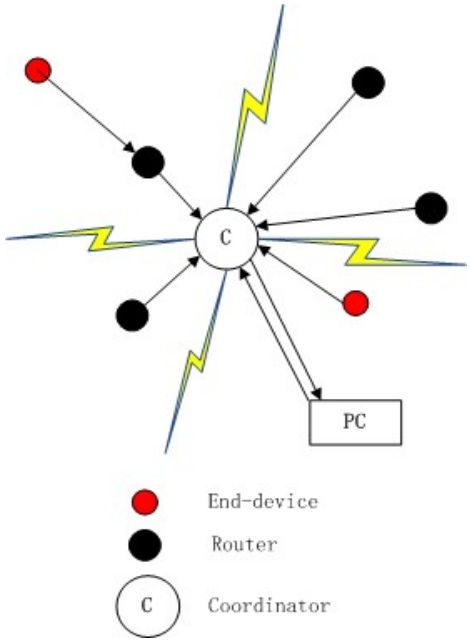


Figure 3. Message transmission process

The coordinator address is 0x0000, on the topmost layer of the network structure tree. PC starts to calculate and update information of the network tree after it received response from nodes.

Because the ZigBee device saves its information into memory when it joins the network first time, the child can exist independently in the network (even if parent leaves

the network or shut down). This algorithm will calculate the depth of the node based on the node information, and add the parent device that does not exist.

- If there's response from router node, determine whether the parent exists in the network. Connect this device to the next level of the parent which exists in the network. If its parent doesn't exist, calculate the depth.

Suppose the child address is N_c, and parent address is N_f. If $(N_c\text{-}N_f\text{-}1)=0$, child depth $D=N_c$; If $(N_c\text{-}N_f\text{-}1)\neq 0$, $(N_c\text{-}N_f\text{-}1)/C_{skip(d)}$, d=0,1,2,3,4. When $(N_c\text{-}N_f\text{-}1)/C_{skip(d)}$ is integer, child depth $D=d+1$, parent depth $D=d$. Then add parent to the network tree.

- If there's response from the end-device node, determine whether the parent exists in the network similarly. Connect this device to the next level of the parent if the parent exists. Otherwise, take the FLOOR of $[(N_c\text{-}N_f\text{-}1)/6]$ to get the size of the address sub-block being distributed by parent. That is to say, $C_{skip}(d)=[(N_c\text{-}N_f\text{-}1)/6]$. Child depth $D=d+1$, parent depth $D=d$, and add parent to the network tree.

The flow chart of PC software is shown in Figure 4.

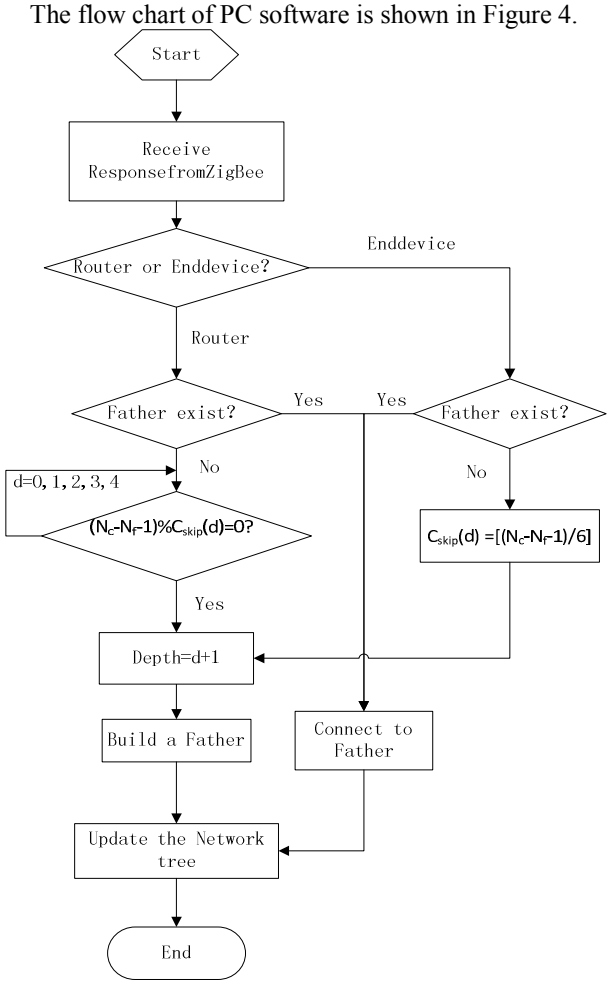


Figure 4. Flow chart of the software program on PC

C. *Results and analysis*

The Software program on PC, Z-Network Tool, is developed based on the above algorithm. In our experiment, the coordinator connects to the Internet through Serial to Ethernet Module. The software sends commands and accesses data through Ethernet. The software interface is shown in Figure 5.

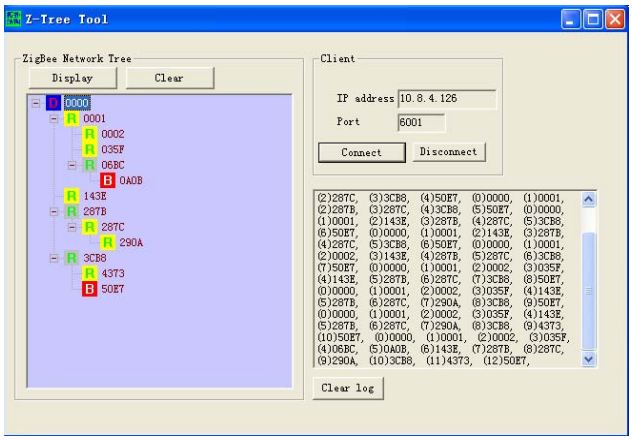


Figure 5. Software interface

In the experiment, we open the coordinator 0x0000, the router nodes 0x0001, 0x0002, 0x035F, 0x143E, 0x287C, 0x290A, 0x4373, and the end-device equipments 0x0A0B, 0x50E7. Finally we get the network structure tree shown in Figure 6.

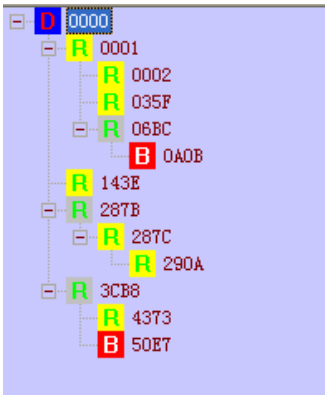


Figure 6. Software running result

From Figure 6, we can find that the nodes we opened are all shown in the network structure tree. Parents of the nodes 0x0A0B, 0x287C, 0x4373 and 0x50E7 do not exist, and then the software completes the father nodes, and shows them by gray icon. We can also find that coordinator is in the depth 0, 0x0001 and 0x143E are in the depth 1, 0x0002, 0x035F, 0x287C, 0x4373 and 0x50E7 are in the depth 2, 0x0A0B and 0x290A are in the depth 3, respectively. According to the distributed addressing scheme, we can verify that the result is correct.

IV. CONCLUSION

We customize the command frame that is sent to the ZigBee network and the response message returned to PC. Software program on PC receives the response from ZigBee and calculates the nodes' positions to generate the structure tree. The network structure tree can be refreshed timely to reflect the equipment changes in the network. The functions of this software can also be extended. For example, in the positioning system, we can configure the coordinates of a node with this software.

ACKNOWLEDGMENT

This work is supported by SRF for ROCS, SEM, the Fundamental Research Funds for the Central Universities (JUSRP20914, JUDCF10031) and the Foundation of Key Laboratory of Advanced Process Control for Light Industry (Jiangnan University), Ministry of Education, P. R. China.

REFERENCES

[1] ZigBee Alliance, Inc. ZigBee Specification: ZigBee Document 053474r17[S]. USA: ZigBee Standards Organization, 2007.

[2] L. Cui, F. Wang, H. Luo, et al. A pervasive sensor node architecture. In: NPC2004, Lecture Notes in Computer Science 3222, Springer Berlin/Heidelberg, 2004: 565-567.

[3] G. Ding, Z. Sahinoglu, B. Bhargava, P. Orlik, and J.Y. Zhang, "Reliable broadcast in zigBee networks", Sensor and Ad Hoc Communications and Networks (SECON 05), September. 26-29, 2005, pp. 510-520, doi: 10.1109/SAHCN.2005.1557103.

[4] S. Choi, H. Cha, and S.Cho, "A soc-based sensor node: evaluation of retos-enabled cc2430," Sensor, Mesh and Ad Hoc Communications and Networks (SECON 07), June 18-21, 2007, pp. 132-141, doi: 10.1109/SAHCN.2007.4292825.

[5] B. NEFZI and Y. Q. Song. Performance analysis and improvement of zigbee routing protocol. In 7th IFAC International Conference on Fieldbuses & Networks in Industrial & Embedded Systems, pages 351-369, France, 2007.

[6] M. Sveda, and R. Trchalik, "ZigBee-to-internet interconnection architectures," International Conference on System (ICONS 07), April 22-28, 2007, pp. 30-36, doi: 10.1109/ICONS.2007.58.

[7] Z. Yao, F. Dressler, "Dynamic Address Allocation for Management and Control in Wireless Sensor Networks". Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS'07), 2007.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Evolving Neural Network Structure by Indirect Encoding Based on BQPSO

Fang Bao[1] Jun Sun[2] and Wenbo Xu [2]

[1]Jiangyin polytechnic college. No.168, xicheng road,
Jiangyin Jiangsu, China 214405
[2]School of Internet of Things Engineering, SouthYangtz
university. No. 1800, Lihudadao, Wuxi Jiangsu, China 214122
baofang@mail.jypc.org

Abstract: This paper proposes a novel algorithm of neural network structure evolve. First, the algorithm designs an indirect encoding schema representing the structure of neural network, use joint seed representing the existence of connection in neural network. Then, creating and evolving the coordinates of the joint seed using Binary Quantum-behaved Particle Swarm Optimization (BQPSO), evolving the value of the joint seed using nine-palace evolving rule, by separately evolve the coordinates and value of the joint seed, the growing and pruning of the network structure is achieved. The experimental results show that the algorithm has stable complexity when dealing with different scales of neural network. By the proposed indirect encoding schema and separated coordinates and value evolving, the algorithm solves the problem of geometrically growing structure-evolving complexity successfully.

Keywords: neural network structure evolve, indirect encoding, joint seed, BQPSO

I. INTRODUCTION

The perfectly-designed structure of Feed-forward Neural Network (FNN) is the foundation of successful application of it. Recently, the research has been concentrated on the design of FNN structure using evolutionary algorithm. The design of optimal structure of FNN can be regarded as an optimization problem in the corresponding search space, where each particle represents a kind of FNN structure. So, the key point becomes how to translating the FNN structure into the particle of the evolutionary algorithm, mean the encoding schema of FNN.

Mainly, there are two kinds of encoding schema, direct encoding and indirect encoding. Direct encoding proposed by Miller[1] is well known, which is based on the codification of all possible connections in FNN, thus the length of the encoding schema (mean the length of the particle) is the square of all possible neurons in network. Direct encoding has little flexibility, the length of the structure encoding will achieve geometrically large when the FNN is bigger, the evolving is impossible on such particle. Direct encoding is suitable for small architectures.

In order to reduce the length of the particle and make the encoding schema more scalable, indirect encoding had been proposed in recent years. Merril[2] introduced a fractal method of indirect encoding, use fast simulated annealing method for the evolution of the FNN structure. Compare with the direct encoding, the research space is relatively small in the indirect encoding. The limitation is that in the optimizing process, each particle must be decoded to illustrate an untrained FNN, and be trained by traditional training method, so the evolving speed is slow.

QPSO[3] is derived from Particle Swarm Optimization (PSO)[4] that proposed by Kennedy and Eberhart in 1995. PSO is a colony-intelligent evolutionary algorithm. The algorithm is motivated by the behavior of animal organization such as migration. Individual make the decision of next step according to the best experience of itself and the best experience of the colony, thus moving to the optimized position step by step. QPSO defines the particle in a quantum space determined by the probability-density function, thus the search space of particle is greatly extended.

QPSO initializes a group of random particles, 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 stop 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 when compared with PSO[5][6]. BQPSO[7] is the discrete binary version of QPSO.

In this paper, a novel algorithm of FNN structure evolving is proposed. When dealing with different scales of FNN, the length of the encoding schema in the proposed algorithm changes not too much, and the convergent performance is stable.

The following of the paper is organized as: In section 2, the indirect encoding schema is introduced. Section 3 illustrates the evolving of the coordinates and value of the joint seed. Section 4 gives the experiment results. The paper is concluded in section 5.

II. INDIRECT ENCODING SCHEMA

Joint seeds are used to represent the existence of connections in FNN in this algorithm, and a 2-dimention space is utilized to position the joint seed.

$$M = Dimx \times Dimy \tag{1}$$

where Dimx is the sum of all input and output neurons in FNN, Dimy represents the number of maximum possible neurons in hidden layer. Let b to be the sum of all input and output neurons, then b^2 is the number of maximum possible neurons in hidden layer, and a $b*b^2$ space can be used to represent the FNN structure.

In case that the number of neurons in input layer is n,

978-0-7695-4415-1 2011

DOI 10.1109/DCABES.2011.41

define the joint seed in position (i,j) as:

If (i<=n), represents a connection from the ith input neuron to the jth hidden neuron.

If (i>n), represents a connection from the jth hidden neuron to the (i-n)th output neuron.

If the value of the joint seed is 1, the connection exist, there is no connection if the value is 0. If there is no joint seed in position (i,j), call the position is inactive, else it is active whether the joint seed's value is 1 or 0.

Now, the structure of FNN which has 1 hidden layer can be represented by such a 2-dimension space in fig 1. Coordinates of joint seeds in the space will be created and evolved by BQPSO, value of them will be evolved by the special nine-palace evolving rule. The structure of FNN is evolved step by step until the final stable situation is achieved.

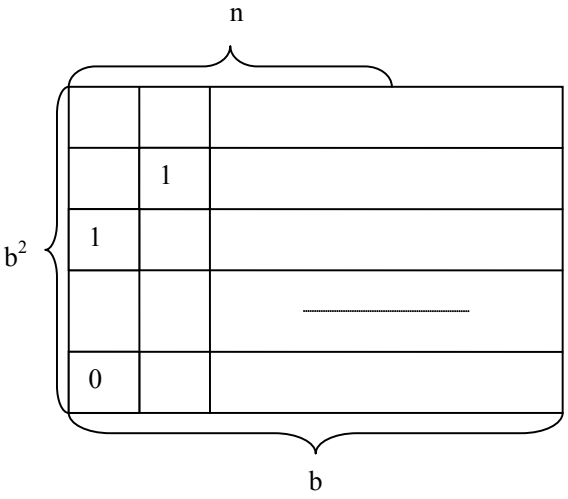


Figure 1. 2-Dimension Space Representing FNN Structure

For such purpose, dimension of the BQPSO particle is designed to be 3, the first dimension represents the row-coordinates of the joint seed(which is b), the second dimension represents the first multiplier of column-coordinates(which is b^2) of joint seed, and the last multiplier of b^2 is represented by the third dimension. The length of each dimension is determined by the binary string that could represent b. Coordinates of joint seeds is made up of such 3 binary strings, meanwhile the length of the FNN structure encoding is also ascertained.

For instance b=8, the sum of input and output neurons is 8, then the maximum possible neurons in hidden layer is 8*8=64. Three 3-bit binary strings are used to represent the coordinates of the joint seed. The length of the structure encoding is 9. Even when the sum of input and output neurons is 256, mean b=256, the maximum possible neurons in hidden layer is 256*256, however, a binary string of three 8-bit can be used to represent the coordinates of the joint seed. The length of the FNN structure encoding is 24. Complexity of FNN structure encoding is not changed so much.

III. EVOLVING OF STRUCTURE

The FNN structure is obtained by evolving the coordinates and value of the joint seed separately.

A. Evolution of Joint Seed's Coordinates

Coordinates of joint seed in the space are represented by BQPSO particle, it is created and evolved by BQPSO. Like other evolutionary algorithm, fitness function of BQPSO is the gist of individual evolution, as well as the terminating condition of the algorithm.

In this algorithm, fitness function of BQPSO could not simply be the approximate error of FNN. It ought to be the minimum network structure that could reach the desired approximate error more quickly. So, sum of connections, training generations, desired approximate error and practical approximate error of network should all be taking into account.

Define the fitness function as follows:

$$F = \frac{1}{nc * rc} * \frac{e_d}{e_r} \qquad (2)$$

where e_d is the desired approximate error and e_r is the practical approximate error, nc is the sum of connections in current network, rc is the current training generations of BQPSO.

Such fitness function represents the effect of the proposed algorithm, means how to reach the desired approximate error by less connections and training generations. If the value of fitness function is stable in some generations, it means the proportion of e_d, e_r, nc and rc is fixed, the algorithm is finish, no more coordinates of joint seed is evolved.

At the beginning of this algorithm, according to the length of the structure encoding schema, certain numbers of particle that has such length are initialized by BQPSO, their value are the coordinates of the initial joint seeds. Every next generation, particles are evolved according to above fitness function, meanwhile the coordinates of the next generation of joint seeds are obtained. Until the fitness value of BQPSO is stable, the algorithm finish, no more particles are generated, means no more joint seeds needed to be generated.

B. Evolution of Joint Seed's Value

Once the coordinates of joint seed is created and evolved by BQPSO, the value of it is evolved in a nine-palace which center is the joint seed.

In neural network, neurons work independently. The output of neuron at one time is only determined by the neurons that connected to it at that time. Other neurons which are not connected to it at that time have no effect. So the nine-palace evolving rule can be used here to change the

value of joint seed, mean the connections in network.

2 nine-palace evolving system is used, one is called growing system and the other is called pruning system.

The growing system is used to form an almost full-connected network. Joint seed is called growing seed now. Coordinates of several initial growing seeds is created by BQPSO, let g_s be the value of it, we define the nine-palace evolving rule of growing system as:

If the position is inactive, but there exist growing seeds in at least 3 continuous grids in other 8 grids of its nine-palace, which implies large probability of neuron-connection in this local area, then one growing seed g_s is filled into this position.

The growing system grows until there is no position could satisfy above condition.

The almost full-connected network is not accord with the fundamental of FNN structure designing, pruning system is imported to delete some connections.

Define the initial state of pruning system to be the final state of growing system plus the new pruning seeds created by BQPSO. Let c_s be the value of pruning seed, if there already has a value in the position that created to be pruning seed, replace it with c_s. Define the nine-palace evolving rule of pruning system as:

If the position has g_s, and in other 8 grids of its nine-palace, there exist growing seeds in at least 2 continuous grids that lies in same row or column, and exists one pruning seed that connected to them but is not lies in the same row or column, which implies large probability of pruning in this local area, then one pruning seed is filled into this position.

The pruning system works until there is no position could satisfy above condition.

Integrate 2 systems, first, the coordinates of several growing seeds are created by BQPSO, the growing seeds grow. According to the growing seeds and newly created pruning seeds, the pruning seeds grow. The final pruning system represents the current FNN structure. Replace the growing seeds with 1 and the pruning seeds 0, the current network structure could be obtained.

C. Training of Current FNN

Current network is trained using float-point QPSO. 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. FNN can be trained using QPSO.

In this structure design algorithm, because current network structure has been ascertained, so we can train it with QPSO algorithm.

D. Summary of FNN Structure Evolution

The complete steps of indirect BQPSO–based FNN structure evolve algorithm could be described as follows:

1. According to the indirect structure encoding schema, the length of each particle in BPQSO is ascertained.

2. Ccertain numbers of particle are initialized by BQPSO, their value are the coordinates of the initial joint seeds-growing seeds.

3. Growing system and pruning system run in sequence, and current network structure is obtained.

4. Current network is trained by QPSO.

5. The next generation of BQPSO particles are evolved according to the fitness function, meanwhile the coordinates of the next generation of growing seeds are obtained.

6. Go back to step 3 until the fitness value of BQPSO is stable, the algorithm finish.

IV. EXPERIMENT RESULTS

The proposed indirect BQPSO-based FNN structure evolve algorithm is tested on many domains. The goal is to obtain optimized FNN structure that could solve the given problem. In this section, 2 typical domains are introduced, then, the results are compared with those provided by direct encoding schema.

The direct encoding schema used in our experiment is designed as: a 2-dimension matrix as in formula(1) is used to represent the current FNN structure, at each point, value 1 means the connection exist while 0 means no connection. Such matrix is regarded as one particle of BQPSO, and the structure of FNN is evolved by BQPSO directly. So the length of the structure encoding is $b^* b^2$. The fitness function and training of current FNN is the same as the proposed indirect schema. The direct schema is called direct algorithm, while the indirect schema is called indirect algorithm in following sections.

A. Parity Problem

Parity problem evaluates parity character of input model, an N-bit binary string. This is a highly complicated problem because of the similar form of input models. It is widely used to test the efficiency of structure design and training algorithm of FNN. Current research[8] show that, when there exist full connections between input and hidden layer, hidden and output layer in FNN, the number of sufficient hidden layer neurons is N/2+1 if N is even and (n+1)/2 if N is odd. In this experiment, 7-bit and 8-bit parity problem are adopted, parameters are design as follows:

TABLE 1 Parameters of Parity Problem Using Indirect Algorithm

	length of encoding	number of particles in BQPSO
7-bit	3*3=9	18
8-bit	4*3=12	24

TABLE 2 Parameters of Parity ProblemUsing Direct Algorithm

	length of encoding	number of particles in BQPSO
7-bit	8*64=512	36
8-bit	9*81=729	48

B. Approximation of HERMIT Polynomial

HERMIT Polynomial:

$$f(x) = 1.1(1 - x + 2x^2)e^{(-\frac{x^2}{2})} \quad (3)$$

FNN algorithms performing this function-approximate have been discussed a lot. The most common FNN structure is one there is 1 input neuron, 7 hidden neurons and 1 output neuron[9].

In the experiment, the initial structure is set as: 1 input neuron, 4 hidden neurons and 1 output neuron.

TABLE 3 Parameters of HERMIT Polynomial Using Indirect Algorithm

space size	length of encoding	number of particles in BQPSO
2*4	1*3=3	5

TABLE 4 Parameters of HERMIT Polynomial Using Direct Algorithm

space size	length of encoding	number of particles in BQPSO
2*4	2*4=8	10

C. Experiment Result

Figure.2 and Figure.3 show FNN structures obtained using indirect algorithm for the 7-bit parity problem and 8-bit parity problem.

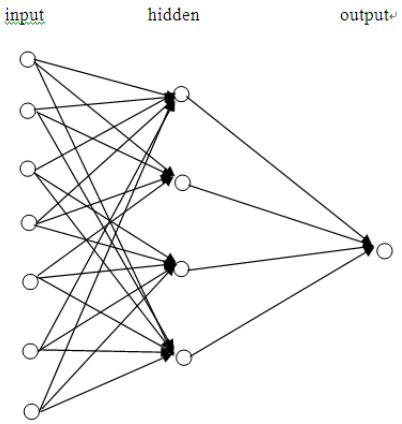


Figure 2. FNN Structure of 7-bit Parity Problem Using Indirect Algorithm

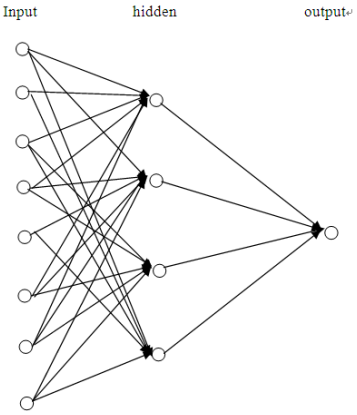


Figure 3. FNN Structure of 8-bit Parity Problem Using Indirect Algorithm

Compared with the research in [8], when using indirect algorithm, numbers of hidden neurons are the same, but some connections are eliminated, mean the network is not full connected, the structure obtained by the proposed algorithm is simpler than [8], the validity of this algorithm is proved.

Figure.4 and figure.5 show the structures obtained using direct algorithm for the same problem.

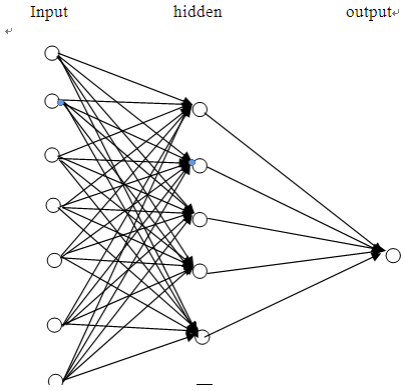


Figure 4. FNN Structure of 7-bit Parity Problem Using Direct Algorithm

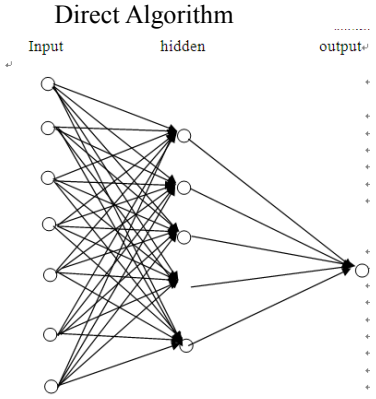


Figure 5. FNN Structure of 8-bit Parity Problem Using Direct Algorithm

Compared with the result using indirect algorithm, numbers of hidden neurons is larger, as well as the number of connection, the final structure obtained is more complex using direct algorithm.

Figure.6 shows the FNN structure obtained using indirect algorithm for the approximation of HERMIT Polynomial and figure.7 shows the structure obtained using direct algorithm.

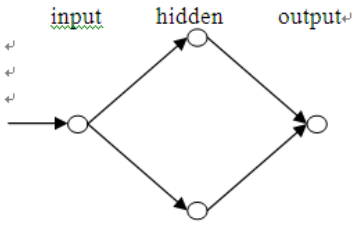


Figure 6. FNN Structure of HERMIT Using Indirect Algorithm

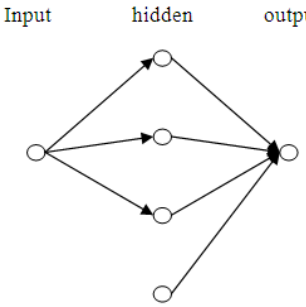


Figure 7. FNN Structure of HERMIT Using Direct Algorithm

As can be observed, the structure obtained by the proposed indirect algorithm is much simple than most commonly used FNN structure for solving this problem in [9], and is also simple than the structure obtained using direct algorithm.

When applying the proposed structure design algorithm to other domains, whether the initial FNN is simple or complex, all the final FNN structure achieved is simpler than usually used one.

For both indirect and direct algorithm, the average change trend of fitness value between every two generations for solving the parity problem and HERMIT are shown in figure .8 and figure.9. In these figs, Y-coordinate represent the change of fitness value between every two generations, abscissa represents the training generations.

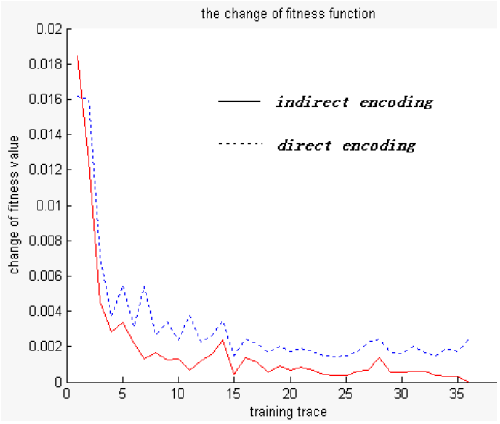


Figure 8. Change of Fitness Value for Parity Problem

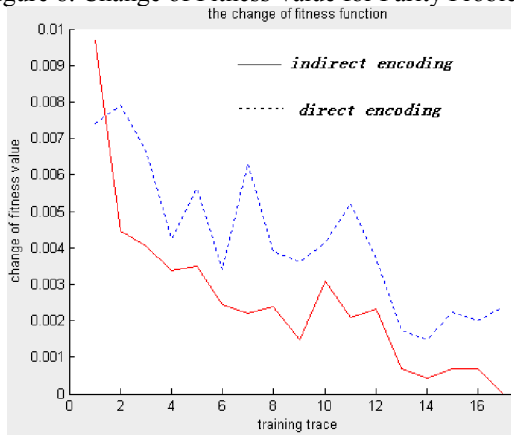


Figure 9. Change of Fitness Value for HERMIT

From the change of fitness function, we can see, indirect algorithm has stable convergent character, even when the scale of FNN changes, it can achieve a stable change rate in about 15 generations. When the change of fitness value between every two generations reduces to 0, it means the value of fitness function is stable now, and means the proportion of e_d, e_r, nc and rc in formula (2) is fixed. So, the final FNN structure is achieved at the end, it means the algorithm is finish.

In direct encoding schema, the change of fitness value between every two generations could not reduce to 0, it will surge at about 0.0003 at the end, this means the value of fitness function could not stable, structure design could not stop.

V. CONCLUSIONS

According to the Feed-forward Neural Network, a novel

algorithm of neural network structure evolve by indirect encoding based on BQPSO is proposed. The algorithm disngs joint seed representing the existence of connection in neural network. By separately evolve the coordinates and value of the joint seed, the growing and pruning of the network structure is achieved. With specially designed fitness function, the coordinates of the joint seed is evolved by BQPSO, the value of the joint seed is evolved by nine-palace evolving rule, current network is trained by float-point QPSO. The final stable structure is found step by step.

The experimental results show that the algorithm has stable convergent capability when dealing with different scales of FNN. The proposed algorithm can solve the problem of geometrically growing encoding length that commonly used structure design algorithm can't deal with.

REFERENCES

[1] Miller.G.F,Todd.P, Hegde.S. Designing neural networks using geneticalgorithm. Proceedings of the 3th International Conference on GAs[C].MorganKanfmann,1989.379～384.

[2] John.W.L.Merril. Fractally Configured Neural Network. Neural Netwoks.1991,Vol.4,53-60.

[3] Jun Sun, bin Feng, wenbo Xu. Particle Swarm Optimization with Particles Having Quantum Behavior[C]. Piscataway: IEEE int. Conf. on evolutionary computation,2004. 325-331.

[4] J.Kennedy, R.Eberhart. Particle Swarm Optimization. IEEE int. Conf. on Neural Network1995. 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. Pro- ceedings of the International Symposium on distributed Computing and Application to Business, Engineering and Science 2006. 767-770.

[7] Jun Sun, wenbo Xu, Wei Fang, ZhiLei Chai. Quantum-Behaved Particle Swarm Optimization with Binary Encoding. The 8th International Conference on Adaptive and Natural Computing Algorithms.2007, 376-385.

[8] Soniag E. D. Feedforward Nets for Interpolation and Classification[J]. Journal of Computer and System Science,2003,Vol 45

[9] Xu L,Krzyzak A,Oja E. Rival Penalized Competitive Learning for Clustering Analysis RBF-Net and Curve Detection. IEEE Transactions on Neural Networks[J]. 1999, 4(4):636-649.

Contraction-Expansion Coefficient Learning in Quantum-Behaved Particle Swarm Optimization

Na Tian, Choi-Hong Lai, Koulis Pericleous
School of Computing and Mathematical Science
University of Greenwich
London, UK

Jun Sun, Wenbo Xu
School of IOT Engineering
Jiangnan University
Wuxi, China

***Abstract*—**Quantum-behaved particle swarm optimization was proposed from the view of quantum world and based on the particle swarm optimization, which has been proved to outperform the traditional PSO. The Expansion-Contraction coefficient is the only parameter in QPSO, which has great influence on the global search ability and convergence of the particles. In this paper, two parameter control methods are proposed. Numerical experiments on the benchmark functions are presented.

Keywords-*Quantum-behaved Particle Swarm Optimization; Contraction-Expansion coefficient; cosine function; annealing function;*

I. INTRODUCTION

Particle Swarm Optimization (PSO) algorithm is a population-based optimization technique originally introduced by Kennedy and Eberhart in 1995 [1]. A PSO system simulates the knowledge evolvement of a social organism, in which individuals (particles) representing the candidate solutions of an optimisation problem traverse through a multi-dimensional search space in order to determine the optima or sub-optima. The position of each particle is evaluated as according to the objective function, and particles in a local neighbourhood share memories of their "best" positions. These memories are used to adjust the particles' own velocities, and their subsequent positions. It has already been shown that the PSO algorithm is comparable in performance with and may be considered as an alternative to the Genetic Algorithm (GA) [2].

In the original PSO system with M particles, 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 the 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:

$$V_i(k+1) = \omega V_i(k) + c_1 r_1 (P_i(k) - X_i(k)) + c_2 r_2 (P_g(k) - X_i(k)) \quad (1)$$

$$X_i(k+1) = X_i(k) + V_i(k+1) \quad (2)$$

where $i = 1, 2, \cdots, M$, $j = 1, 2, \cdots, D$, ω is the Inertia Weight, c_1 and c_2 are called Acceleration Coefficient. r_1 and r_2 are random numbers distributed uniformly in (0,1), that is $r_1, r_2 \sim U(0,1)$. Vector $P_i = (P_{i1}, P_{i2}, \cdots, P_{iD})$ is the best previous position (the position giving the best objective function value) of particle *i* called personal best 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 position.

The main disadvantage of the original PSO algorithm is that it is not guaranteed to be global convergent [3]. Concepts of a quantum-behaved particle swarm optimization (QPSO) was developed to address the disadvantage and first reported by Sun in 2004 [4].

Trajectory analysis in [5] demonstrated the fact that the convergence of the PSO algorithm may be achieved if each particle converges to its local attractor, $p_i = (p_{i1}, p_{i2}, \cdots p_{iD})$ defined at the coordinate:

$$p_i(k) = \varphi P_i(k) + (1-\varphi) P_g(k) \quad (3)$$

where P_i is personal best position of particle i, P_g is the global best position of all particles, $\varphi \in (0,1)$. It can be seen that p_i is a stochastic attractor of particle i that lies in a hyper-rectangle with P and P_g being two ends of its diagonal and moves following P_i and P_g.

In quantum world, the velocity of the particle is meaningless, so in QPSO system, position is the only state to depict the particles, which moves according to the following equation:

$$X_i(k+1) = p_i(k) \pm \alpha \left| C_i(k) - X_i(k) \right| \ln(1/u) \quad (4)$$

where C, called Mean Best Position, is defined as the mean value of the *pbest* positions of all particles. That is

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.32

$$C(k) = (C_1(k), C_2(k), \cdots, C_D(k)) = \left(\frac{1}{M}\sum_{i=1}^{M} P_{i1}(k), \frac{1}{M}\sum_{i=1}^{M} P_{i2}(k), \ldots, \frac{1}{M}\sum_{i=1}^{M} P_{iD}(k) \right) \quad (5)$$

where M is the population size and P_i is the personal best position of particle i. The parameter α in (4) is known as the Contraction-Expansion Coefficient, which can be adjusted to control the speed of convergence.

The QPSO method is different from the original PSO method in that the iterative update of the former method is given by (4) ensuring particles appear in the entire D-dimensional search space during each of the iteration steps, while the particles in the latter method can only move in a bounded space. Using the global convergence criterion in [3], one can conclude that the QPSO method is a global convergent algorithm whereas the original PSO method is not. Moreover, unlike the original PSO method, the QPSO method does not require velocity vectors for the particles at all and has fewer parameters to control, making the method easier to implement. Experimental results performed on some well-known benchmark functions show that the QPSO method has better performance than the original PSO method [6-8].

In order to improve the search ability of the particles, a lot of works have been done to select the Inertia Weight. In [9-10], Shi and Eberhart depicted the linearly decreasing Inertia Weight from 0.9 to 0.4, which is found to be able to accelerate the convergence rate. A self-adaptive method for adjusting Inertia Weight is proposed in [11], which is defined as a function of particle fitness, swarm size and dimension size.

In this paper, the Expansion-Constriction coefficient in QPSO, which has the similar meaning to the Inertia Weigh in PSO, is investigated. A new method is proposed to adjust the coefficient during the search process, which is able to enhance the global search ability of the particles in QPSO.

The rest of this paper is organized as follows. Section 2 reviews the commonly used coefficient selection methods and gives our proposed method. The numerical results of benchmark tests are presented in section 3. Finally in section 4, the conclusion and future works are given.

II. LEARNING OF CONTRACTION-EXPANSION COEFFICIENT

The Contraction-Expansion coefficient α, the only parameter in QPSO, influences the trade-off between global and local exploration ability of the particles. A larger α facilitates global exploration (searching new areas) while a smaller α tends to facilitate local exploitation to fine-tune the current search area. Suitable selection of α can provide a balance between global and local exploration ability and thus save searching time to find the optimum. $\alpha \le 1.7$ must be satisfied to guarantee the convergence [12]. The linearly decreasing Expansion-Constriction from 1.0 to 0.5 is commonly used in the literature, , this is because the larger α at the beginning helps to find good seeds and the later smaller α facilitates fine-tune search.

$$\alpha = 1.0 - 0.5\frac{k-1}{k_{\max}-1} \quad (6)$$

A constant value $\alpha = 0.75$ is analyzed by Sun in [13] that it has better performance in simple unimodal problems. Many adaptive selection methods were proposed in order to enhance the performance of QPSO. In [12], an adaptive method on individual level is proposed as (7), and proved to outperform the linearly decreasing method.

$$\alpha(z) = \begin{cases} 0.6, & z > 0 \\ 0.7, & -2 < z \le 0 \\ 0.6 + 0.1\times k, & -k-1 < z \le -k \\ 1.0 + 0.2\times(k-4), & -k-1 < z \le k \\ 1.8, & z \ge -8 \end{cases} \quad (7)$$

where $z = \ln(\Delta F)$, ΔF is defined as follows:

$$\Delta F = \frac{(F_i - F_{gbest})}{MIN(ABS(F_i), ABS(F_{gbest}))} \quad (8)$$

where F_i is the fitness of the ith particle, F_{gbest} is the global best fitness of the swarm. However, we have to compute z according (8) before we obtain the value of α, which is time consuming and not easy to use.

In [11], Dong defined a self-adaptive Inertia weight function for a particle in terms of its fitness value, the swarm size and the dimension size as:

$$\omega = \frac{1}{3 - \exp(-\frac{M}{200}) + (\frac{R}{8D})^2} \quad (9)$$

where M is the swarm size, D is the dimension size of the solution space and R denotes the fitness rank of the given particle. But the value of ω is always smaller than 0.5, which is not suitable for QPSO.

Inspired by Simulated Annealing, the Contraction-Expansion coefficient α may be adjusted according to the annealing function:

$$\alpha = \alpha_0 * (CR)^k \quad (10)$$

where CR is chosen as according to the maximum iteration.

For the multimodal problems, it's easy to get trapped into the local optimum because of the particles' premature. So here, we propose another parameter control method in which α adjusts according to the cosine function as follows:

$$\alpha = 0.5\cos(\frac{\pi k}{2k_{\max}}) + 0.5 \quad (11)$$

where k is the current generation, $k_{\max}$ is the maximum generation. In Fig. 1, one can note that the cosine function decreases slower than the linearly decreasing function and annealing function. For the complex multimodal, QPSO with cosine decreasing α may perform better in the global exploration.

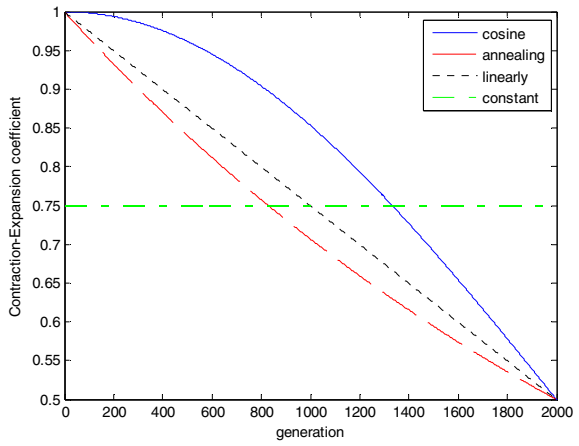


Figure 1. Contraction-Expansion coefficient decreases in different way

The procedure of QPSO algorithm is described as follows:

1) Initialize the population with uniformly distributed random numbers. Initialize the personal best position and global best position.
2) Calculate the Contraction – Expansion coefficient by Eq. (6), (10) or (11).
3) Compute the mean best position by Eq. (5).
4) Update the position of each particle by Eq. (4).
5) Calculate the fitness of each particle.
6) Update the personal best position and global best position.
7) If the termination criterion is satisfied, then stop. Otherwise, go to step 2).

III. BENCHMARK TESTS

A. Benchmarks and Parameter Settings

Five well-known benchmark functions given in Table 1 were used to evaluate the performance of the QPSO with different parameter selection methods, both in terms of the optimum solution and the rate of convergence to the optimum solution. These benchmark functions were widely used in evaluating performance of evolutionary methods. All the functions are all minimization problems and have the global minimum at the origin or very close to the origin.

Function f_1 is simple unimodal problem, f_2 is simple multimodal problem, and f_3 - f_5 are highly complex multimodal problems with many local minima. All swarms were randomly initialized in an area equal to one quarter of the feasible search space that was guaranteed not to contain the optimal solution. Initialization ranges for all functions can be also found in Table 1.

TABLE 1. BENCHMARK FUNCTIONS

Function	Expression	Range
Sphere	$f_1(x)=\sum_{i=1}^{D}x_i^2$	$(-100,100)^D$
Rosenbrock	$f_2(x)=\sum_{i=1}^{D}[100(x_{i+1}-x_i^2)^2+(x_i-1)^2]$	$(-30,30)^D$
Rastrigin	$f_3(x)=\sum_{i=1}^{D}[x_i^2-10\cos(2\pi x_i)+10]$	$(-5.12,5.12)^D$
Griewank	$f_4(x)=\frac{1}{4000}\sum_{i=1}^{D}x_i^2-\prod_{i=1}^{D}\cos(\frac{x_i}{\sqrt{i}})+1$	$(-600,600)^D$
Ackley	$f_5=20+e-20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D}x_i^2}\right)-\exp\left(\frac{1}{D}\sum_{i=1}^{D}\cos(2\pi x_i)\right)$	$(-32,32)^D$

As in [10], for each function, three different dimension sizes are tested, which are 10, 20 and 30. The maximum number of generations is set as 1000, 1500 and 2000 corresponding to the dimensions 10, 20 and 30, respectively. Population sizes of 20, 40 and 80 are used with different dimensions. A total of 50 runs were carried out and the average optimal value and the standard deviation are presented.

Four kinds of Contraction-Expansion coefficient α selection methods are tested. In the results tables, the abbreviations QPSO-CON, QPSO-LIN, QPSO-COS and QPSO-ANN stand for QPSO with constant α, QPSO with linearly decreasing α, QPSO with cosine decreasing α and QPSO with annealing decreasing α respectively.

B. Results from Benchmark simulations

The mean best values and standard deviation for 50 trials on the five benchmark functions are listed in Tables 2-6. For the unimodal function f_1, QPSO-CON outperforms others for all the dimension sizes and converges fast to the global optimum. For the simple multi-modal function f_2, QPSO-CON performs better in the dimension sizes 20 and 30 except $D=10$. Rastrigin function f_3 is a fairly difficult problem due to its large number of local minima, on which, QPSO-COS works better. Because the cosine function decreases more slowly than the linearly decreasing function and annealing function, the particles have more time to explore the big area before fine-tuning the small area. So the QPSO-COS can prevent the particles from being trapped into the local optimum and the global search ability is enhanced. For Griewank f_4, QPSO with annealing α QPSO-ANN outperforms others, either in terms of the optimum and convergence. Within finite number of iterations e.g. 2000, QPSO-ANN converges faster than QPSO-COS and QPSO-LIN to a minimum. QPSO-COS and QPSO-ANN has competitive performance on f_5, QPSO-COS works better for the large dimension size, e.g. $D=30$.

Comparison of the convergence between all the algorithms over the five benchmark functions is shown in Figs. 2-6. In Figs. 2-3, QPSO-CON converges fastest to a minimum on the unimodal function and simple multimodal function. In Fig. 4, the convergence of QPSO-COS is the slowest, but the superiority appears at the later stage of the searching process. QPSO-ANN shows the best performance in Fig. 5. From Fig. 6, we can note that QPSO-COS has the best convergence, which may illustrate the ability of QPSO-COS in solving the complex problems.

TABLE 2. SPHERE FUNCTION

M	D	k_{max}	MEAN BEST VALUE (STANDARD DEVIATION)			
			QPSO-CON	QPSO-LIN	QPSO-COS	QPSO-ANN
20	10	1000	**5.78E-76** (1.40E-75)	1.85E-40 (1.29E-	8.44E-36 (5.90E-35)	2.24E-42 (1.30E-
	20	1500	**8.39E-57** (3.04E-56)	6.67E-21 (4.62E-	4.43E-19 (2.93E-18)	9.35E-22 (2.78E-
	30	2000	**1.82E-45** (6.27E-45)	2.04E-15 (4.28E-	8.19E-12 (3.15E-11)	6.65E-14 (1.92E-
40	10	1000	**2.66E-87** (1.15E-86)	5.30E-73 (3.64E-	1.94E-59 (1.36E-58)	2.13E-76 (1.47E-
	20	1500	**2.00E-71** (9.72E-71)	1.50E-43 (6.14E-	1.32E-34 (3.76E-34)	5.31E-42 (3.60E-
	30	2000	**1.24E-58** (3.22E-58)	1.50E-30 (3.95E-	3.02E-22 (1.48E-21)	2.92E-30 (8.39E-
80	10	1000	1.39E-97 (5.81E-97)	1.02E-100	3.86E-79 (1.34E-78)	**2.28E-109**
	20	1500	**1.39E-85** (4.67E-85)	2.97E-70 (1.11E-	6.45E-53 (3.05E-52)	9.20E-72 (5.18E-
	30	2000	**3.48E-72** (9.35E-72)	1.57E-50 (5.28E-	5.05E-37 (2.06E-36)	2.22E-51 (1.31E-

TABLE 3. ROSENBROCK FUNCTION

M	D	k_{max}	MEAN BEST VALUE (STANDARD DEVIATION)			
			QPSO-CON	QPSO-LIN	QPSO-COS	QPSO-ANN
20	10	1000	41.09 (67.08)	**11.75** (27.37)	14.79 (42.50)	28.60 (72.62)
	20	1500	**25.30** (31.36)	44.70 (44.34)	47.59 (68.83)	71.03 (87.45)
	30	2000	**64.56** (53.22)	72.15 (79.69)	85.34 (86.64)	69.90 (74.20)
40	10	1000	9.59 (16.06)	7.55 (11.07)	6.95 (5.21)	**5.10** (4.54)
	20	1500	**17.06** (26.66)	39.67 (31.26)	37.16 (31.75)	37.99 (31.50)
	30	2000	**28.35** (30.37)	47.56 (37.84)	48.46 (31.30)	53.72 (37.24)
80	10	1000	6.36 (3.91)	4.11 (3.37)	**3.82** (3.92)	4.17 (3.52)
	20	1500	**11.56** (17.16)	27.65 (26.76)	33.02 (30.05)	27.83 (27.54)
	30	2000	**31.02** (33.40)	41.28 (27.90)	34.18 (23.51)	41.36 (28.90)

TABLE 4. RASTRIGRIN FUNCTION

M	D	k_{max}	MEAN BEST VALUE (STANDARD DEVIATION)			
			QPSO-CON	QPSO-LIN	QPSO-COS	QPSO-ANN
20	10	1000	23.93 (13.12)	5.74 (3.23)	**4.64** (3.79)	5.20 (2.92)
	20	1500	67.69 (27.27)	19.06 (10.44)	**16.22** (8.10)	19.70 (11.17)
	30	2000	113.38 (34.69)	38.77 (14.36)	**38.52** (21.25)	46.10 (19.48)
40	10	1000	11.81 (9.16)	3.42 (2.00)	**3.29** (2.16)	3.52 (1.42)
	20	1500	33.27 (19.99)	13.03 (7.30)	**12.06** (5.13)	12.32 (4.28)
	30	2000	71.00 (30.14)	23.49 (8.81)	**23.07** (10.85)	23.31 (7.97)
80	10	1000	6.50 (4.90)	2.29 (1.40)	**2.14** (1.32)	2.26 (1.31)
	20	1500	24.35 (13.23)	9.48 (3.21)	**8.60** (6.42)	9.00 (2.98)
	30	2000	43.18 (19.42)	17.44 (5.86)	**16.83** (5.82)	17.85 (6.42)

TABLE 5. GRIEWANK FUNCTION

M	D	k_{max}	MEAN BEST VALUE (STANDARD DEVIATION)			
			QPSO-CON	QPSO-LIN	QPSO-COS	QPSO-ANN
20	10	1000	6.84E-02 (4.32E-02)	7.17E-02 5.22E-02	9.52E-02 (8.19E-02)	**5.70E-02** (4.25E-02)
	20	1500	2.27E-02 (1.96E-02)	2.09E-02 (2.03E-02)	2.45E-02 (3.10E-02)	**2.02E-02** (1.52E-02)
	30	2000	1.46E-02 (1.89E-02)	1.68E-02 (2.68E-02)	1.13E-02 (1.20E-02)	**9.37E-03** (1.31E-02)
40	10	1000	5.40E-02 (2.96E-02)	6.05E-02 (4.40E-02)	8.22E-02 (8.45E-02)	**4.47E-02** (2.74E-02)
	20	1500	1.78E-02 (1.74E-02)	2.08E-02 (1.91E-02)	1.85E-02 (2.02E-02)	**1.76E-02** (1.46E-02)
	30	2000	1.21E-02 (1.52E-02)	1.24E-02 (1.21E-02)	1.15E-02 (1.34E-02)	**1.06E-02** (1.40E-02)
80	10	1000	5.50E-02 (3.29E-02)	4.07E-02 (3.94E-02)	6.50E-02 (4.93E-02)	**3.78E-02** (2.69E-02)
	20	1500	1.59E-02 (1.59E-02)	1.29E-02 (1.40E-02)	1.61E-02 (1.55E-02)	**1.22E-02** (1.60E-02)
	30	2000	1.05E-02 (1.51E-02)	9.61E-03 (1.18E-02)	8.45E-03 (1.09E-02)	**7.88E-03** (1.07E-02)

C. Results Disscusions

The methods proposed to control the parameter in QPSO, have been shown to be successful especially on the complex Rastrigin, Griewank and Ackley functions. Beyond our expectations, QPSO-ANN performs better on the Griwank function. Since QPSO-COS converges slowly, it maybe need more iterations to achieve the global minimum. Here, we set the number of generation $k_{max} = 5000$ for $D = 30$ with 80 particles to compare the performance of different methods on the Griewank function. The results are shown in Table 7 and

Fig. 7, from which we can note that QPSO-COS performs better than other methods although it converges slowly.

TABLE 6 ACKLEY FUNCTION

M	D	k_{max}	MEAN BEST VALUE (STANDARD DEVIATION)			
			QPSO-CON	QPSO-LIN	QPSO-COS	QPSO-ANN
20	10	1000	17.66 (6.45)	15.20 (8.54)	**13.19** (9.46)	13.99 (9.16)
	20	1500	19.59 (4.00)	19.17 (4.75)	19.11 (4.83)	**18.74** (5.53)
	30	2000	20.61 (6.68E-02)	20.15 (2.88)	**20.11** (2.87)	20.56 (9.68E-02)
40	10	1000	17.21 (6.94)	13.98 (9.15)	14.38 (8.97)	**12.40** (9.70)
	20	1500	19.97 (2.85)	19.54 (3.80)	18.69 (5.51)	**18.29** (6.10)
	30	2000	20.19 (2.89)	19.30 (4.88)	**20.00** (2.93)	20.12 (2.88)
80	10	1000	15.18 (8.54)	13.58 (9.32)	**13.56** (9.31)	15.17 (8.53)
	20	1500	17.89 (6.61)	19.07 (4.82)	**17.85** (6.51)	19.09 (4.82)
	30	2000	20.13 (2.88)	19.25 (4.87)	**18.44** (6.05)	18.88 (5.49)

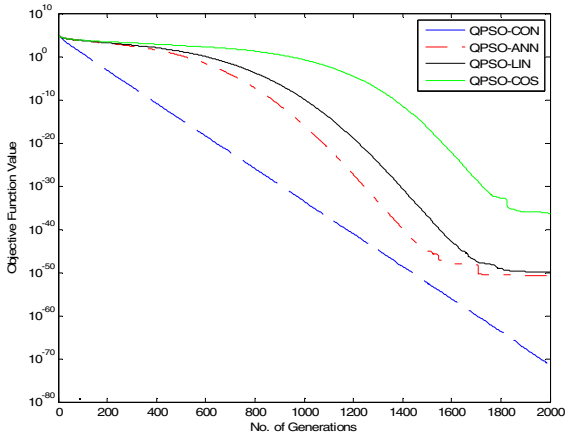


Figure 2. f_1 Sphere function

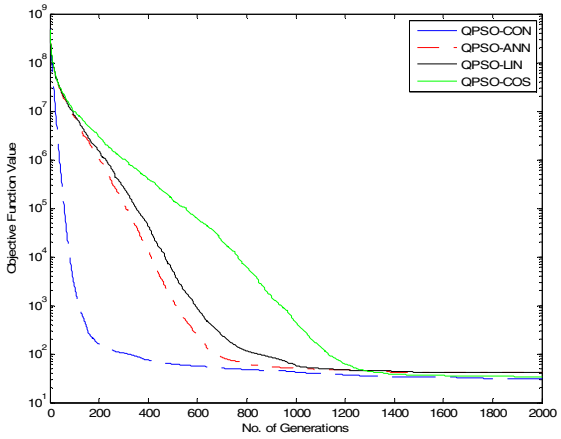


Figure 3. f_2 Rosenbrock function

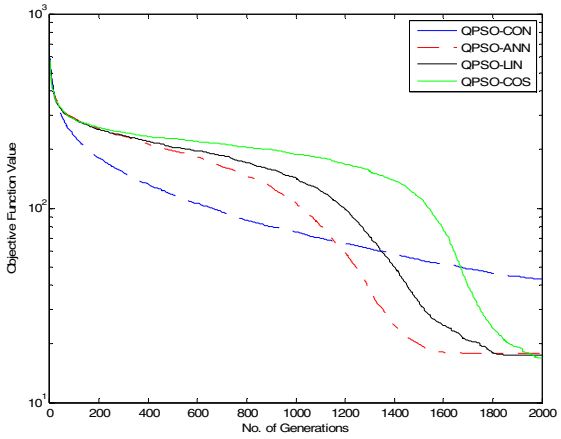


Figure 4. f_3 Rastrigrin function

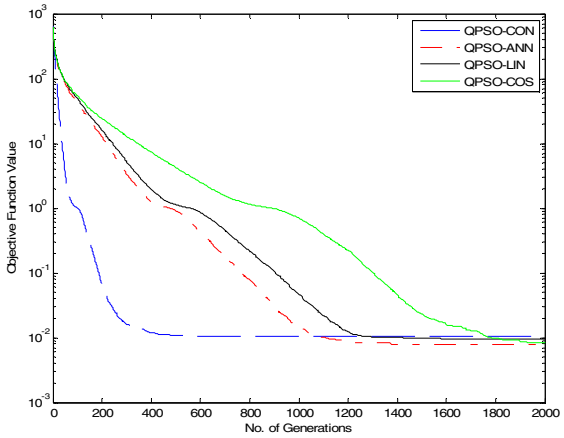


Figure 5. f_4 Griewank function

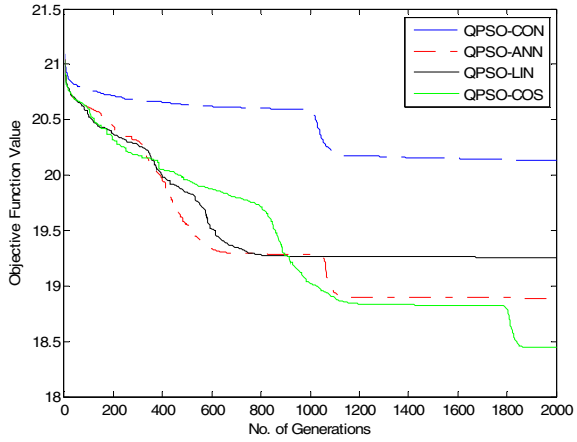


Figure 6. f_5 Ackley function

Table 7 GRIEWANK FUNCTION WITH 5000 ITERATIONS

M	D	k_{max}	MEAN BEST VALUE (STANDARD DEVIATION)			
			QPSO-CON	QPSO-LIN	QPSO-COS	QPSO-ANN
80	30	5000	6.75E-03 (7.37E-03)	7.28E-03 (1.15E-02)	6.03E-03 (9.37E-03)	8.46E-03 (1.25E-02)

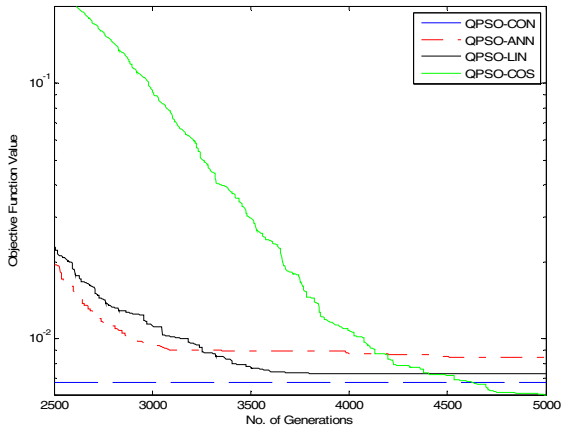


Figure 7. f_4 Griewank function with 5000 iterations

IV. CONCLUSION AND FUTURE WORKS

In this paper, we reviewed previous works on the selection of the Contraction-Expansion coefficient, which has significant impact on the performance of QPSO. We also proposed two new methods to control α, which are simple to use and can improve the global search ability and avoid the premature convergence. The comparison with other coefficient selection methods is presented by the benchmark tests. In our future work, we will be devoted to find an efficient and easily used self-adaptive coefficient selection method.

ACKNOWLEDGMENT

This work is partially supported by Bursary of the University of Greenwich, "the Fundamental Research Funds for the Central Universities" (Project Number: JUSRP21012), the innovative research team project of Jiangnan University (Project Number: JNIRT0702), and National Natural Science Foundation of China (Project Number: 60973094).

REFERENCE

[1] J. Kennedy and R.C. Eberhart, Particle swarm optimization, IEEE Int. Conf. Neural Networks, Perth, Australia, 1995.

[2] R.C. Eberhart and Y. Shi, Comparison between genetic algorithm and particle swarm optimization. Evolutionary Programming VII, Lecture Notes in Computer Science 1447, Springer Berlin, Heidelberg, pp. 611-616, 1998.

[3] F. Van den Bergh, An analysis of particle swarm optimizers, Ph.D. diss., University of Pretoria, South Africa, 2001.

[4] J. Sun, B. Feng and W.B. Xu, Particle swarm optimization with particles having quantum behaviour, Cong. Evolution. Comp., Portland, USA, 2004.

[5] 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 , pp. 58-73, 2002.

[6] J. Sun, W.B. Xu and J. Liu, Parameter selection of quantum-behaved particle Swarm optimization, Lecture Notes in Computer Science, 3612

[7] J. Sun, W.B. Xu and B. Feng, A global search strategy of quantum-behaved particle swarm optimization, IEEE Conf. on Cybernetics and Intelligent Systems, Singapore, 2004.

[8] J. Sun, W.B. Xu and B. Feng, Adaptive parameter control for quantum-behaved particle swarm optimization on individual level, IEEE Int. Conf. Systems, Man and Cybernetics, Hawaii, 2005.

[9] Y. Shi and R.C. Eberhart, Parameter Selection in particle swarm optimization, Int. Conf. Evolutionary Programming, Springer-Verlag London, UK, 1998.

[10] Y. Shi and R.C. Eberthart, Empirical study of particle swarm optimization, IEEE Cong. Evolutionary Computation, Piscataway, NJ, 2000.

[11] C. Dong, G.F. Wang, Z.Y. Chen and Z.Q. Yu, A method of self-adaptive inertia weight for PSO, Int. Conf. Computer Science and Software Engineering, vol. 1, pp. 1195-1198, 2008.

[12] J. Sun, W.B. Xu and B. Feng, Adaptive parameter control for quantum-behaved particle swarm optimization on individual level, IEEE Int. Conf. Systems, Man and Cybernetics. Piscataway, NJ, pp. 3049-3054, 2005.

[13] J. Sun, W. Fang, X.J. Wu, Z.P. Xie and W.B. Xu, Quantum-behaved particle swarm optimization: analysis of the individual particle's behavior and parameter selection, Evolutionary Computation (MIT Press), in press.

Research on the node localization based on quantum particle swarm optimal algorithm for WSNs

YANG Jian-bin，
School of Internet of Things Engineering，
Southern Yangtze University,
Wuxi，Jiangsu，China
magicyang168@163.com

XU Wen-bo,
School of Internet of Things Engineering，
Southern Yangtze University,
Wuxi，Jiangsu，China
xwb@jiangnan.edu.cn

***Abstract*- In the applications of wireless sensor networks, the effects of data reflection largely depend on the accuracy of localization of the information collection node. Therefore, the position information of nodes is very important. Without it, events can not be sensed and measured, and value of application will be lost. To improve the location accuracy of nodes in wireless sensor networks, the paper puts forward research of node localization model in WSNs based on quantum particle swarm optimal algorithm. The results of the simulation show that the method can get significantly higher localization accuracy than the genetic algorithm and get the convergence of optimal solutions faster than the standard particle swarm algorithm under the same conditions.**

Keywords： wireless sensor networks(WSNs)； node localization； particle swarm optimization(PSO)； Quantum particle swarm optimization (QPSO)

I. INTRODUCTION

Wireless sensor network is made of the micro-sensor nodes which are low cost, low power consumption, with sensing, data processing, storage and wireless communication capabilities through self-organization [1]. Wireless sensor network is a collection of sensor technology, distributed information processing technology, embedded technology and wireless communication technology together, interact with the outside world through sensors, to complete data collection, communications transmission, data processing and other functions, and which is widely used in environmental monitoring , industrial control, military, health care, intelligent home, intelligent urban traffic areas. Node localization as one of the three key technologies of wireless sensor network, and is charged with not only the perception of data had occurred, but also for the related problems of wireless sensor network routing, not only for network management, load balancing, preventing the emergence of the energy hole, but also for target tracking, real-time monitoring. In the node deployment, each node is installed on the global positioning system, will result in a significant increase in cost, capacity consumption increases, volume increases, so that the advantages of wireless sensor will no longer exist, and it does not seem to suit the occasion realistically used in real life. In order to meet the majority of occasions, and maintain high precision positioning of the nodes, it is a good idea that the use of intelligent optimization algorithm processing the positioning of wireless sensor network node problem. On the one hand to save costs, save energy, and make full use of the advantages of wireless sensor networks, on the other hand, QPSO has the advantages of quickly find the optimal solution, and it can play a quick and easy solution advantage of intelligent optimization algorithms

II. RELATED WORK

Node localization algorithm for wireless sensor networks is divided into Range-Based and Range-Free two location algorithm based on it is whether or not need to measure the actual distance of nodes in localization process, The former need to measure the actual distance or orientation between adjacent nodes, and use the distance between the nodes to compute the unknown node location. Such as RSSI (Received Signal Strength Indicator) algorithm, TOA (Time Of Arrival) algorithm, TDOA (Time Difference Of Arrival) algorithm and AOA (Angle Of Arrival) algorithm, The latter do not need to measure the absolute distance or orientation between the nodes, but to use the estimated distance between nodes to compute

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.79

node positions, For example, the centroid algorithm, DV-Hop algorithm and APIT (Approximate Point-In-Triangulation Test) algorithm. The former is more accurate positioning, but has a larger computation and communication costs; the latter's positioning accuracy is low, but all aspects of costs are also low, so it is appropriate for low-power, low-cost application fields. This paper is more attention to the field of sensor network applications in precise positioning, so this research is based on the Range-Based high precision positioning mechanism.

In the Range-Based positioning mechanism, the sensor nodes in the network can be divided into two types: one node is the ordinary nodes that do not have the capacity of measuring their coordinates, also called Unknown Nodes, which is the main part of networks; other nodes are equipped with positioning tool like GPS to measure the coordinates of their own, general are referred to as Beacon Node, and can be configured by GPS or manual. They are scattered across the sensor network deployment area.

A trilateral positioning method

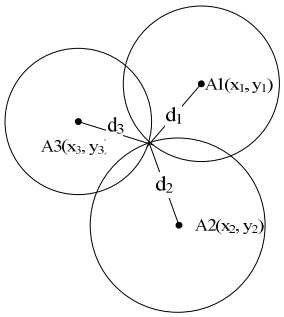


Figure 1 trilateral positioning method

$$\begin{aligned} d_1 &= \sqrt{(x-x_1)^2+(y-y_1)^2} \\ d_2 &= \sqrt{(x-x_2)^2+(y-y_2)^2} \\ d_3 &= \sqrt{(x-x_3)^2+(y-y_3)^2} \end{aligned} \quad (1)$$

The actual coordinates value of three beacon nodes A1, A2, A3, and their distance to the unknown node A, we use A1, A2, A3 three beacon nodes as the center respectively, and use the distance of the unknown node to beacon node A as the radius for three circles, then the intersection of three circles will be the coordinates of unknown nodes.

B The quantum behavior of the particle swarm (QPSO) algorithm[3]

In the classical PSO algorithm, the particles are achieved search in a limited search space and a certain line, and the particle velocity and displacement will be limited by a variety of tracks, like the birds in the sky in search of food, will decrease the speed and shorten the distance because of various external factors. But in the quantum world, particles' trajectory is not limited, and only evolves over time, the evolution observes Schrodinger equation, so X_i and V_i of the particles can not be determined simultaneously. In PSO system, if a single particle has the quantum behavior, PSO algorithm will follow certain rules to make calculations. In QPSO algorithm, the quantum state of a microscopic particle is described by the wave function φ (x, k). When φ (x, k) is confirmed, the probability distribution of any particles' average and measuring value of mechanical volume are fully determined.

By the Monte Carlo method, the equation which describes the movement of particle swarm is introduced, where p, β, $Mbest_i$ are respectively given by equation (4), (2) and (3), u is a random number belongs to [0,1].

$$x_i(t+1) = p \pm \beta \times \left| Mbest_i - x_i(t) \right| \times \ln(1/u) \quad (2)$$

The formula (4)-(5) are introduced , β is called the shrinkage factor changes by t in formula (4), β changes with the change of t, β_1 and β_2, are respectively called the start and end values shrinkage factor changes by t. the variable of Maxiter is maximum number of iterations. Generally, $\beta_1 = 1$, $\beta_2 = 0.5$, so in most cases, the value of β always lies between 0.5-1.0.

$$\beta = \frac{(\beta_1 - \beta_2) \times (Maxiter - t)}{Maxiter + \beta_2} \quad (3)$$

Formula (4) $Mbest_i$ is the average of the optimal location that all particles have searched in the local, pid is the best position that each particle has searched in the local; In type (4a), $Mbest_i$ represents the average of optimal position that each of particles has found in the global search, p_{gd}^{j} represents the optimum position which all particles have found in the global search.

$$Mbest_i = \frac{1}{M}\sum_{d=1}^{M} p_{id} \quad (4a)$$

$$Mbest_i = \frac{1}{M}\sum_{d=1}^{M} p_{gd}^{i} \quad (4b)$$

Local search ensures convergence of PSO algorithm, and the parameter P in formula (2) played a decisive role, such as the type (5) below.

$$P = \frac{(c_1 * p_{id} + c_2 * p_{gd}^{i})}{c_1 + c_2} (c_1 + c_2 = 1) \quad (5)$$

2) Quantum particle swarm optimization algorithm process

The QPSO algorithm process as follows:

a) Initialize each particle's position vector;

b) Calculate the objective function value of each particle;

c) Update the local best position of each particle p_{ij} and related variables;

d) Update the global best position pgj and other relevant variables;

e) According to formula (4a) and (4b) to calculate mBest;

f) According to formula (5) to calculate PP_{ij};

g) According to formula (2) to calculate x_{ij};

h) Repeat steps b) - g), until it meets the number of iterations

III. BASED ON THE QUANTUM BEHAVIOR OF THE PARTICLE SWARM WITH NODE LOCATION MODEL AND ITS IMPLEMENTATION

Algorithm main steps are as follows:

Step 1 unknown node (x, y) and beacon nodes NodeP are deployed in the 5m * 5m two-dimensional area. Beacon nodes known their exact coordinates.

Step 2 the beacon node can measure the distance to the unknown node by sending data packets to the them. Simulation is based on the following formula:

$$d_i = \sqrt{(x - x_i)^2 + (y - y_i)^2} \quad (6)$$

Step 3 For each unknown node, we initialize the optimal position, X = 5 * rand, Y = 5 * rand, and initialize the value of each parameter of QPSO algorithm. pBest is initialized to the current value of the particle.

Step 4 Use QPSO algorithm to find the coordinate of (X, Y) which can make the formula (7) have minimum. According PSO algorithm, we can update particle's velocity and position to find the optimal estimates.

Fitness function as:

$$f(x,y) = \frac{1}{M}\sum_{i=1}^{M} (\sqrt{(x - x_i)^2 + (y - y_i)^2} - \hat{d}_i)^2 (M = 3) \quad (7)$$

Step 5 Use QPSO algorithm to find (x,y) which can make the error minimize, use vector record the optimal value and iteration number every time.

Step 6 The second step to the fifth step continuously updated to find the final optimal solution, that is, to minimize errors and obtained the optimal location coordinates of each node.

In the iterative process, each node choose the closest three beacons node to locate, it can get the minimum distance, thus reduce the error caused the distance. And it makes the estimated value of the coordinate closer to the true value, achieve optimal results.

A Simulation results and analysis

1) Simulation Environment

In order to verify the quantum particle swarm algorithm solve the question of node localization in wireless sensor networks is the effectiveness, we use MATLAB R2009a to estimate and simulate quantum particle swarm optimization algorithm for WSN nodes' localization. Simulation experiments use conventional genetic algorithm (GA), the standard particle swarm optimization (PSO) and the quantum particle swarm optimization (QPSO) algorithm to compare.

Population size is set to 100, all nodes are deployed in 5m * 5m square region. As is shown in Figure 2, beacon node deployment coordinates B1 [0 1], B2 [1 3], B3 [2 1], B4 [3 4.2], B5 [4 3], B6 [1.2 3.9] and B7 [3.4 2.8]; unknown node deployment coordinates A1 [1.2 2.4], A2 [2.5 1.4], A3 [0.4 3], A4 [2.5 3.2] and A5 [3.6 3.5].

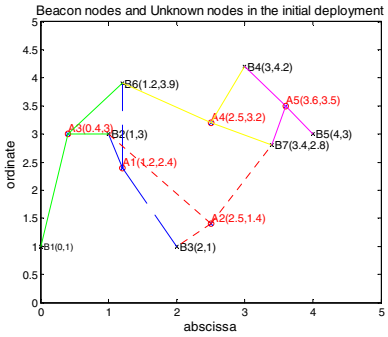


Figure 2 Initialize deployment of Beacon Node(x) and Unknown Node(o)

In the initialization parameters of genetic algorithm [8], the length of the encoding is 10, crossover probability pc = 0.60; mutation probability pm = 0.1; the maximum number of iterations is 1000. The initialization parameters of standard particle swarm optimization [5-7,9]: the initial inertia weight w = 0.9, termination of inertia weight w = 0.4, acceleration of weight $C_1 = C_2 = 2$, the maximum speed $V_{max} = 2$. Quantum particle swarm optimization initialization parameters for the maximum of β is 1, the minimum is 0.5.

2) Experimental results

TABLE I SOLVING COORDINATES OF THE NODE WITH GA

(1000 ITERATIONS)

Unknown Node	*True coordinates*	*Solving coordinates*
A1	(1.2000, 2.4000)	(1.2023, 2.3949)
A2	(2.5000, 1.4000)	(2.4731, 1.3782)
A3	(0.4000, 3.0000)	(0.4007, 2.9618)
A4	(2.5000, 3.2000)	(2.4926, 3.2160)
A5	(3.6000, 3.5000)	(3.6070, 3.4701)

TABLE II SOLVING COORDINATES OF THE NODE WITH PSO AND QPSO (1000 ITERATIONS)

Unknown Node	*True coordinates*	*Solving coordinates*
A1	(1.2000, 2.4000)	(1.2000, 2.4000)
A2	(2.5000, 1.4000)	(2.5000, 1.4000)
A3	(0.4000, 3.0000)	(0.4000, 3.0000)
A4	(2.5000, 3.2000)	(2.5000, 3.2000)
A5	(3.6000, 3.5000)	(3.6000, 3.5000)

As the table shows, the relative error of genetic algorithms to solve is larger, and in the case of ranging error small infinitely, the accuracy of particle swarm algorithm and quantum particle swarm optimization algorithm close to 100%, which indicates they have the advantage of high accuracy for solving.

Three algorithms find the optimal solution first compared the number of iterations shown in Table III.

TABLE III COMPARISON OF SOLVING RESULTS FOR THREE ALGORITHMS (100TIMES)

Algorithms	*First find the optimal*	*Meaning Generations of stop*	*Fitness Value*
GA	Nothing	1000	0.0135
PSO	624	657	0
QPSO	266	336	0

Table III shows that the generations of quantum particle swarm optimization algorithm to find the optimal solution of node localization in wireless sensor network is about 300, while the standard particle swarm optimization is more than 600 generation, it reflects the quantum particle swarm optimization algorithm solve the problems quickly, thus saving limited energy for sensor network nodes to maintain the persistence of life cycle, so wireless sensor networks can ensure longer and better data collection and transmission at a certain extent.

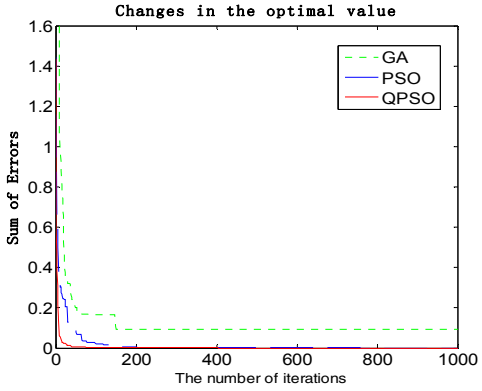


Figure 3 Iterative process of solving the Localization Errors Of three algorithms(1000times)

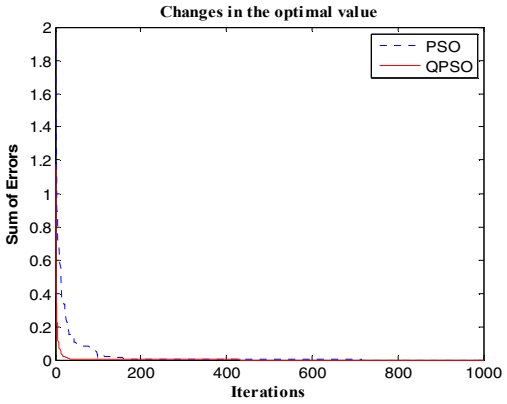


Figure 4 Iterative process of solving the Localization Errors of PSO and QPSO algorithm (1000times)

Figure 3 shows changes of the sensor nodes' location error in QPSO algorithm, GA algorithm and PSO algorithm. It can be seen from Figure 3: In the case of infinitely small errors, PSO algorithm and QPSO algorithm can achieve smaller value of fitness function than GA algorithm. Figure 3, Figure 4 and Table III show that: QPSO algorithm can be found the optimal solution in less generation than PSO algorithm. Figure 4 shows the QPSO algorithm can find the ideal solution for this problem in about 100 generations, and can find the optimal solution about 300 generations.

When the ranging error is in a certain range, as the figure shows above, PSO and quantum particle swarm algorithm also can find the optimal solution , but there are some errors can be seen from the above, in the case of ranging error is 0.1, the sum of five points ranging error is around 0.5, but on the basis of the convergence bending of iterative curve, quantum particle swarm optimization obtain optimal solution faster than the standard PSO, while the genetic algorithm only can obtain solution close to the optimal solution.

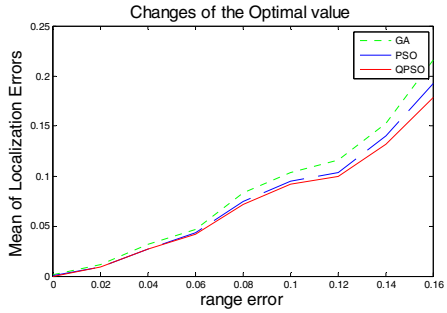


Figure 5, node localization error of three algorithms in the same conditions of ranging error (300 times)

As simulation results show that: in the same case, the node localization algorithm based on QPSO algorithm obtained better results than GA algorithm and PSO algorithm，It uses features of the fast convergence itself and avoiding premature convergence to local optimum solution, with little cost for iterating to find optimal solution, and significantly improves the positioning accuracy and the speed of positioning.

IV. CONCLUSION

QPSO algorithm has features of faster convergence, shorter computation time, and higher solution accuracy. QPSO algorithm is applied to wireless sensor node localization to find the solution in this article, it can complete the optimal positioning of WSNs effectively and quickly, and largely improve the positioning accuracy of wireless sensor nodes, which plays an important role in extending the life cycle for the entire wireless sensor network.

REFERENCES

[1] Sun Li-ming , Li Jian-zhong, Chen Yu . Wireless Sensor Networks [M]. Beijing : Tsinghua University Press, 2005.

[2] Kennedy J, Eberhart R. Particle swarm optimization[C]//Proc IEEE International Conference on Neural Networks , IV Piscataway, NJ, USA. 1995 : 1942−1948.

[3] Sun J, Feng B, Xu Wb. Particle swarm optimization with particles having quantum behavior[C]//Proceedings of 2004 Congress on Evolutionary Computation. 2004 : 325−331.

[4] Sun J. A global search strategy of quantum-behaved particle swarm optimization[C] //Proc 2004 Congress on Cybernetics and Intelligent Systems.

[5] K. S. Low, H. A. Nguyen, and H. Guo, A localization of a wireless sensor network[C]// particle swarm optimization approach for the Proc. IEEE Int. Symp.Ind. Electron. 2008, pp. 1820−1825.

[6] A. Gopakumar and L. Jacob, Localization in wireless sensor networks using particle swarm optimization[C]// Proc. IET Int. Conf. Wireless, Mobile Multimedia Netw. 2008, pp. 227−230.

[7] Mao G, Fidan B, Anderson B. Wireless sensor network localization techniques [J]. Computer Networks, 2007,51(10) : 2529−2553.

[8] Zhang Lei, Duan Li-li, Qian Zi-juan, Huang Guang-ming, WSN Node Localization Technology Based on Genetic Algorithm [J]. Computer Engineering, 2010, 36(10) : 85-87.

[9] Wang Xiao-le, Xu Jia-ping, Research on node localization based on particle swarm optimization for WSNs[J] . Journal of Computer Applications, 2009, 29(2) : 494-495.

[10] WANG Fu-Bao;SHI Long;REN Feng-Yuan. Self-Localization Systems and Algorithms for Wireless Sensor Networks [J]. Journal of Software, 2005, 16(5) : 45-49.

QPSO and FISH Algorithm Apply in the Wireless Sensor Network Coverage

WANG Zhi
School of Internet of Things Engineering，
Southern Yangtze University,
Wuxi，Jiangsu，China
wz_jnedu@sina.com

XU Wen-bo
School of Internet of Things Engineering，
Southern Yangtze University,
Wuxi，Jiangsu，China
xwb@jiangnan.edu.cn

Abstract-The applying of QPSO-FISH algorithm in the wireless senor network are mainly studied in this paper.Wireless sensor network is an energy-constrain network, coverage efficiency and power consumption are two important performance indicators. Maximizing the network coverage and minimizing the number of working nodes are network optimization goals.Establish an optimal model for coverage in wireless sensor network. Utilize the characteristics of parallel,fast coveragence and fast local search speed, propose a coverage optimization strategy based on QPSO-FISH algorithm.The results of simulation show that the algorithm can get an optimal set of nodes and improve the real-time capability of nodes scheduling.

Key words: wireless sensor networks; AFSA algorithm;QPSO algorithm ;optimal coverage

I.INTRODUCTION

Wireless sensor network is composed of a number of single node,each node iteracts with the environment by controling parameters or sensing.Node coverage is an important issuein sensor network. Optimization coverage of wireless sensor network has great significance in rational allocation of network resources, complete environmental awareness well and other tasks[1].The coverage problems of the wireless sensor networks is divided into deterministic coverage and random coverage[5]. Deterministic coverage mainly researches how to cover the known environment with the least nodes.Random coverage monitors the unknown areas of environment by adopting random deployment[6].In general, the network nodes that randomly deployed can not be recycled or supplied the energy.Therefore,how to ensure covering the adequate monitoring area and extending the life of network is a serious problem[7].Currently,the mainly researching method is using redundant coverage of nodes to extend the lifetime of network,adopting the energy-efficient coverage solution that rotates the active nodes and dormant nodes,which can extend the survial time of network efficiently[9]. In order to solve this problem,the adopted algorithm at present are generic algorithms, QPSO, AFSA. AFSA has the ability to overcome the local extremum and obtains good global extremum[1].In the initial of optimization,it has fast convergence,but with the complexity and size of optimization problem continues to expand,algorithm is difficult to obtain accurate optimal solution,only can find a satisfactory solution domain.QPSO has faster convergence rate,but all the particles are moving to the optimal direction,it makes particles tending to homogenization[2].When it optimizes complex and high-dimensional terrain functions , algorithm are easy to fall into local optimum,post-shock easily,it will cause precocious phenomenon and evolution

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.70

stagnation[4].This article combines the advantages of QPSO and AFSA,use the global converagence of AFSA quickly find a satisfactory solution domain,then use QPSO for quick local search..It makes the mixture algorithm not only have the quick local search speed,but also have the global convergence.

II. Cover Problem Description

This article assumes that the area of network is two-dimensional rectangular region plane,network is consisted of N nodes.The set of all sensor nodes is $S=\{s_i,i=1,2….N\}$,coverage model of each sensor node can be expressed as a circle whilch the coordinates is its center ,Rs is its radius[1].Each node can be expressed as $s_i=(x_i,y_i,R_s)$.Digital discrete set of monitoring area is divided into m×n pixels.the area of each pixel as a unit,set an arbitrary pixel p(x,y),the event that target pixel is covered by the sensor nodes s_i is r_i,i=1,2…N,the probability of this event is $P(r_i)$,sensor nodes use the Boolean sensor cover model.then the probability is two-valued function,

$$P(r_i)=P(x,y,s_i)=\begin{cases}1 & d(s_i,p)\le R_s\\ 0 & otherwise\end{cases}$$

for the pixel(x,y), if this pixel is coverd by any pixel of node set C ,then we think it is coverd.Recording the probability that the pixels (x,y) is covered by the selected node set C is P(x,y,C)

$$P(x,y,C)=P(\bigcup_{i=1}^{n} r_i)=1-\prod_{i=1}^{n}(1-P(x,y,s_i))$$

the sum of pixel area that set of sensor work nodes covered is the sum of pixels covered by all the work nodes in working nodes set recorded as area(C),

$$area(C)=\int_0^m\int_0^n P(x,y,C)d_x d_y$$

define a Boolean control vector X=(a1,a2….aN),the control vector describes the state of sensor network nodes.$a_i = 1$ denotes the sensor node i is active,$a_i =0$ denote s the sensor node i is dormant.Sign sensor network coverage f1(x),node utilization f2(x).The coverage rate and node utilization rate:

$$f_1(X) = area(C)/(m\times n)$$

$$f_2(X) = \sum_{i=1}^{N} a_i / n$$

The coverage optimization of the wireless sensor network,one aspect is maximizing network coverage,the other aspect is minimizing the node utilization[2]. Therefore,the coverage control of wireless sensor network is a multi-objective optimization problem,we can form a linear combination of the target by the weighted ,change multiple original sub-objective optimization function into a single target objective optimization function.Objective optimization function is defined as:

$F(X) = w_1 f_1(X)+w_2(1-f_2(X))$

$0<w_1<1$,$w_1 + w_2 = 1$,W1,W2 are the corresponding weights of sub-objective function,and the value of them depends on the requirements of designers to the network indicators.

III. AFSA

AFSA is a stochastic optimization algorithm based on swarm intelligence , in order to achieve optimization,the algorithm simulates the foraging,cluster, rear-end of fish by constructing an aritificial fish[8]. The state of artificial fish can be expressed as vector $X=(x_1,x_2….x_n)$, x_i(i=1,2…n) are variables that need optimization.The concentration of the current location which artificial fish stays is expressed as Y=f(x).Y is the objective function value.The distance between artificial fishes is expressed as $d_{i,j}=||X_i-X_j||$.Visual representatives the sensing range of artificial fish,Step representatives moving steps of artificial fish, δ representatives congestion factor.AFSA

initializes a group of artificial fish, searches for optimal solutions by iteration,in each iteration,the artificial fishes update themselves by foraging,clustering and rear-end[8].

The behavior of artificial fish describes as follow:

1)Foraging behavior.It is a behavior that the fish swim to the place where the food is more.Set the current state of artificial fish Xi,select a state Xj randomly in its view,then calculate the value of objective function and compare them.If $Y_j < Y_i$ in the minimization problem,(or $Y_i<Y_j$ in the maximization problem.),then move a step to this direction;otherwise, re-selected a state X_j randomly in its view,judge whether meets the conditions.After it repeats trynumber times,if it still can not satisfied the conditions,then move one step randomly and make the X_i to a new state.

2)Clusters act.Artificial fish X_i searches the number of partners n_f within the field of current vision ,if there is more food in the center of the partners and not too crowded,then X_i move a step toward the center of partner ,otherwise executive the foraging behavior.

3)Rear-end behavior.When one or several fish of the fish swarm find food ,its neighboring partners will trail it quickly to the point that there is more food.Let the state of artificial fish X_i searching for the best state X_{max} within its view,if there is more food and not too crowded near the X_{max} then X_i move a step to the center of partner ,otherwise executive the foraging behavior..

4)Random behavior.Select a state randomly in view of artificial fish and then move to it,it is a default behavior of foraging behavior.In foraging behavior,if try_number is small,increase the diversity of population,in order to jump out of local optimum.

5)Bulletin board.It is used to record the state of the best individual of artificial fish.When each artificial fish completes its operation,compares its current state to the bulletin board,if its state better than bulletin board,then update the bulletin board with its state,otherwise maintain the bulletin board ,thus the bulletin board records the best states of history.

IV. AFSA-QPSO

AFSA algorithm is not sensitive with initial value and parameter,it has good capacity to overcoming local extremum and get global extremum.[8]But convergence is slow lately,it can only find a satisfactory soluntion domain. and it is difficult to obtain accurate optimal solution.Particles of QPSO algorithm dynamically adjust the current speed and current location according to their own best position pbest and the best position that the swarm experienced gbest,it has fast convergence speed.However ,late in the algorithm,particles tend to the same,make the algorithm very easy to fall into local optimum,causing significant prematurity.

If organically combine the two algorithms to retain the advantages of two algorithms:first using the global convergence of AFSA quickly find a satisfactory solution domain,then use QPSO for fast local search ,make the mixture algorithm not only has a fast local search prime rate,but also ensure global convergence performance. well

V. the step of AFSA-QPSO solving the wireless sensor network coverage problem :

1)Initialization of artificial fish scale N,the location of each fish pos,vision visual,moving step step,the congestion factor δ,the maximum number of repeat trynumber,particle acceleration coefficients c1 and c2,AFSA iterations,QPSO iterations and so on.

2)Calculate the fitness of each artificial fish,and

compared them with the bulletin board ,if it is better,then assign it to the bulletin board.

3)Each fish updates its position by foraging,clusters and rear-end.

4)AFSA termination conditions.If it reaches the preset number of evolutin times,then update the bulletin board with optimal value,optimal location,go to 5) otherwise go to 2)

5)Assign the location information where obtain the optimal value to particles.

6)Assign the best location ,best value of bulletin board to pbest and gbest.

7)Calculate the fitness of each particle.

8)For each particle compares its fitness value with the best position that it experienced ,if it is better,then let it become the current best position pbest.

9) For each particle,compare its fitness value with the global best position that swarm experienced ,if it is better,then update global best position gbest..

10)Update particle position.

11)Check the termination condition,if it fits the termination condition,then output the optimal value, terminate the algorithm, otherwise go to 7).

VI. simulation

Experimental environment settings are as follows:

Monitoring area is located 100m × 100m,the sensing radius of each sensor node R_s = 9m,use a 2.2GHZ computer,use VC to execute simulation.Use the setting of parameters in article [1],set the number of artificial fish 100,set the maximum number of AFSA iterations 300,set the vision of artificial fish 5,retries 3,congestion factor 0.6.Conducted 20 tests,use 50 nodes to optimize the coverage of the monitoring area.

In order to verify the effectiveness of the algorithm,we use 2 algorithm to execute simulation.Compare the results of 2 methods to prove the effectiveness of AFSA-QPSO.

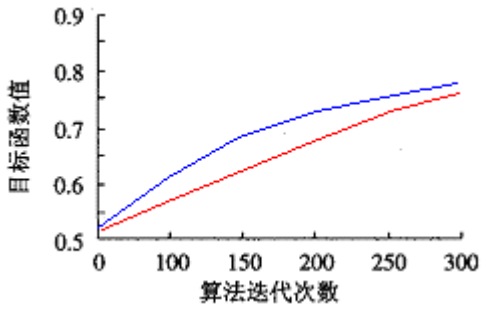
Place n sensor nodes randomly in the monitoring area,use AFSA and AFSA-QPSO to optimize the coverage of monitoring region.Set the number of artificial fish 100,the maximum iteration 300,artificial fish vision 5,retries 3,the congestion factor 0.6.When the number of nodes are taken 60,carring out 30 tests to get an average results.The comparison is as follows:

figure(1)

comparision of AFSA and AFSA-QPSO

iteration	AFSA-QPSO coverage	AFSA coverage	nodes
100	81.24	78.57	60
150	85.34	84.55	60
200	91.28	89.33	60
250	94.55	93.23	60
300	97.65	96.73	60

With the increase of iterations,the coverage of algorithm increased.The coverage of AFSA-QPSO are larger than the AFSA. Therefor,AFSA-QPSO can get better performance.



blue line representative AFSA-QPSO

red line representative AFSA

6.Conclusion

AFSA-QPSO algorithm combines the advantages of AFSA and QPSO .It has good ability to overcome local extremum and obtain global extremum.Meanwhile,it also has a faster convergence rate and the ability to obtain exact optimal solution in the latter.

In this paper,I propose a coverage optimization strategy based on AFSA-QPSO to the coverage problem in wireless sensor network.Simulation results show that AFSA-QPSO can solve the approximate of the optimal coverage with low cost quickly and efficiently.Thereby reduce the network redundancy,improve network efficiency.

References；

[1]L.M.zhou K.H.yang “wireless sensor network coverage optimization strategy based on AFSA algorithm “ in appliction research of computer 2010.6

[2]SLIJEPCEVIC S,POTKONJAK M.”Power efficient organization of wireless sensor networks[C]//Proc of International Conference on Communications. Helsinki:IEEE Communication Society 2001:472.476

[3]Y.jiang,Y.wang “Optimal multiuser detection with artificial fish swarm algorithm[C]//Pro of International Conference on Intelligent Computing. Berlin, Hedlberg:SpringerVerlag,2007:1084-1093

[4]Heo Nojeong and Varhney P K An Intelligent Deployment and Clustering Algorithm for a Distributed Mobile Sensor Network[C]//IEEE Internarional Conference on Systems,Man and Cybernetics,2003.

[5]Koushanfar S Meguerdichian,Potkojak F,Srivatava M B,Coverage Problem in Wireless Ad-hoc Sensor Networks[c]//IEEE Inforcom,2001,380-1387.

[6]Zhang Honghai,CHou Jennifer:Maintining Sensing Coverage and Connectivity in Large Sensor Networks[J],Wireless Ad-hoc and Sensor Networks 2005(1-2):89-123

[7]Mauro Leoncini,Giovanni Resta.Paolo Santi,Partially Controlled Deployment Strategies for Wireless Sensors[J].Ad Hoc Networks,2009(7):1-23.

[8]FARZI S.Efficinet job scheduling in gird computing with modified artificial fish swarm algorithm[J].International Journal of Computer Theory and Engineering,2009,1(1):13-18

[9]Shakkottai S,Stikant T,and Shroff N.Unreliable sensor grids: coverage connectivity and diameter[C].Proceedings of IEEE IN FOCOM’03 SanFrancisco, CA,USA,Mar.30-Apr.3.2003 1073-1083

[10]Ram S S,Manjunath D,Lyer S K,and YogeShwaran D.On the path coverage properties of random sensor networks[J].IEEE Transactionson Mobile Computing, 2007,6(5):494-506

Hybrid-search quantum-behaved particle swarm optimization algorithm

Zhou Chao
Institute of IOT engineering
Southern Yangtze University
Wuxi,China
zhouchao3634@sohu.com

Sun Jun
Institute of IOT engineering
Southern Yangtze University
Wuxi,China
sunjun_wx@hotmail.com

***Abstract*—Quantum-behaved particle swarm optimizafion algorithm(QPSO) can improve the search quality of particle swarm optimizafion algorithm(PSO) in a certain extent.But it still shows that its precision of searching is low and its capability of local searching is weak. Hybrid-search quantum-behaved particle swarm optimizafion algorithm(HSQPSO) has introduced the Chaos search mechanism which based on tent map.It doesn't change the search mechanism of QPSO,and it re-joins the chaos search mechanism to compose the hybrid-search mechanism based on the original.Through comparing the optimal values of two search mechanisms in the iterative process,the global optimum will be obtained. results show that the HSQPSO not only retains the fast convergence of QPSO,but also has higher search efficiency and search precision and isn't easy to be trapped in the local optimal value.**

Keywords-quantum-behaved particle swarm optimization(QPSO); chaos serch;Tent Map; particle swarm optimization(PSO)

I. INTRODUCTION

During the last decade, the development in optimization theory sees the emergence of swarm intelligence, a category of stochastic search methods for solving global optimization (GO) problems. Ant colony (AC) and particle swarm optimization (PSO) are two paradigms of this kind of methods.

Particle Swarm Optimization(PSO), first introduced by Kennedy and Eberhart in 1995 [1], is a stochastic optimization algorithm based on swarm intelligence,due to be inspired by social behavior of birds. PSO is initialized with a swarm of random solutions (particles). Each particle flies in D-dimensional problem space with a velocity, which is adjusted at each time step. The particle flies towards a position, which depends on its own past best position and the position of the best of its neighbors. The quality of a particle position depends on a problem-specific objective function (fitness) [2]-[5]. The algorithm has the features of simple and efficient.But there are also has the shrtcomings of easy to fall into local optimization which will lead to low precision and easy-to-divergence [6].

Quantum particle swarm algorithm(QPSO) is a new research hotspot in recent years. In [7], Jun Sun et al. introduce quantum theory into PSO and propose a Quantum-behaved PSO (QPSO) algorithm. The experiment results indicate that the QPSO works better than standard PSO on several benchmark functions and it is a promising algorithm. It is a new evolutionary algorithm with the capability of high efficient global search.It has the advantage of few adjust parameters,easy to implement,strong convergence,fast convergence and it has been a hot research field of swarm intelligence.

However,QPSO's local search ability is not strong.So this paper propose a new algorithm,we called it hybrid-search quantum-behaved particle swarm optimizafion algorithm(HSQPSO).

II. THE ALGORITHM OF QPS0

QPSO thinks that particles have quantum behavior and there is a DELTA potential whose centre is Pbest.In the quantum space,the particles' nature to meet the aggregation is completely different.The particles have no derermined moving trajectory. This allows particles to explore and find the global optimal solution in the whole feasible solution space.So the global search abililty of QPSO is much better than the classical PSO.

The QPSO algorithm is given below [8]:

Step 1:Initialize an array of particles with random positions inside the problem space.

Step 2：Determine the mean best position among the particles by means of

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

Where M is the population size and p_i is the best position recorded by particle i.

Step 3：Evaluate the desired objective function (e.g. minimization) for each particle and compare with the previous best values. If the current value is better than the previous best value,then set the best value to the current value, i.e. if $f\left(X_i\right) < f\left(P_i\right)$,then $X_i = P_i$.

Step 4：Determine the current global position minimum among the particle's best positions, i.e.

978-0-7695-4415-1 2011

DOI 10.1109/DCABES.2011.50

$$g = \arg\min_{1 \le i \le M} \left(f\left(p_i \right) \right) \quad (2)$$

where M is the population size.

Step 5：Compare the current global position wth the previous global best position. If the current global position is better than the previous global position, set the global position to the current global position.

Step 6：For each dimension of the particle, obtain a stochastic point between P_{ij} and P_{gj}

$$p_j = \varphi^* P_{ij} + (1-\varphi)^* P_{gj}, 0 < \varphi < 1 \quad (3)$$

Step 7：Update the position by the stochastic equation

$$x_{i,j}(t+1) = p_{i,j}(t) \pm \alpha \cdot \left| C_j(t) - x_{i,j}(t) \right| \cdot \ln\left[1/u_{i,j}(t) \right], u_{i,j}(t) \sim U(0,1) \quad (4)$$

where u is a random number uniformly distributed in $(0,1)$, α is the only parameter of the algorithm, is called the contraction–expansion coefficient and can be tuned to control the convergence speed of the algorithms.

Step 8：Repeat steps 2–7 until a stop criterion is satisfied or a pre-specified number of iterations are completed.

III. CHAOS SEARCH

Chaos widely exists in natural phenomena and social phenomena,and it is a much more common phenomena exits in nonlinear system.Chaotic motion has stochastic,ergodicity and regularity in a certain range.That makes it range between deterministic and stochastic and has a wealth of space and time dynamics.Which most important is chaotic motion can't Repetitive traverse all states.Optimizing search mechanism with these characeristics of chaotic motion can make the algorithm jump out of local optimum,maitain population diversity and improve the capability of global serch of the algorithm.

Chaos search iterative generats chaos variable by chaos map,then magnificates the range of chaos variable and maps it to theoriginal range,and thus iterative optimizes the new generated variables.Different chaotic map will lead to different optimization effect of chaos optimization process. Chaos variables are usually generated by Logistic map.But the researchers found that the Tent Map has better traversal uniformity than Logistic Map.

The Tent Map defines as follow [9]:

$$z_{n+1} = \mu\left(1 - 2\left| z_n - 0.5 \right|\right), 1 \le z_0 \le 1 \quad (5)$$

Where $\mu \in [0,1]$, n=0,1,2…….

IV. THE PROPOSED HSQPSO

Hybrid-search quantum-behaved particle swarm optimizafion is a algorithm conbinated the advantages of chaos search mechanism and the QPSO's original search mechanism.In order to make the chaos search have high search efficiency and high search accuracy,we intruduce the intervention mechanism. According to the definition of chaotic search,we can appropriately intervent the process of chaos search and that will make we get better results [3].The expression of intervention mechanism is defined as[9]:

$$Z_k' = (1-\beta)\psi^* + \beta Z_k \quad (6)$$

$$\psi^* = \left(X' - X_{\min} \right) / \left(X_{\max} - X_{\min} \right) \quad (7)$$

$\psi*$ is the value which current optimal value maps to the [0,1]. $X' = \left(x_1' ..... x_D' \right)$ is the value which $X = \left(x_1 ..... x_D \right)$ maps to [0,1]. X represent the particle's position. $Z = \left(z_1 ..... z_D \right)$ is chaos vector generated by Tent Map. $Z' = \left(z_1' ..... z_D' \right)$ is a chaos vector relate to the X where the intervention mechanism intruduced. β is a parameter. $0 \le \beta \le 1$ .Here, $\beta = 1 - k/k_{\max}$, $k_{\max}$ is the maximum number of iterations.

The HSQPSO algorithm is given below:

Step 1: Initialize the particles of QPSO's search mechanism(define as search 1) and the particles of chaos search mechanism(define as search 2) with random positions inside the problem space respectively.Using Tent Map （here, μ =1） to generate chaotic variables.Rewrite （5）as follow：

$$z_j^{i+1} = 1 - 2\left| z_j^i - 0.5 \right| \quad (8)$$

Where z_j denotes the j th chaos variable, and i denotes the chaos iteration number. Evaluate the mean best position among the particles of search 1 according to (1).

Step 2:Evaluate the desired objective function (e.g. minimization) for each particle in search 1 and search 2 respectively and compare search 1's value with search 2's value.If search 1's value is better than search 2's value,then compare search 1's value with the previous best values.Else compare search 2's value with the previous best values. Then set the best value to the current value.

Step 3:Update the posision of search 1's particles according to (4) and update the posision of search 2's particles according to (9).

$$x_{i,j} = x_{\min,j} + z_j^i \left(x_{\max,j} - x_{\min,j} \right) \quad (9)$$

Step 4:If the globle optimum has not been updated for several consencutive iterations and the iteration number $k \le 2/3 k_{\max}$ （ $k_{\max}$ is the maximum number of iteration）, then use the intervention mechanism.Update chaos variable z according to (6) and (7).

Step 5: Repeat steps 2–4 until a stop criterion is satisfied or a pre-specified number of iterations are completed.

V. EXPERIMENT RESULTS AND DISCUSSION

To test the performance of HSQPSO, seven benchmark functions are used here for comparison with classical PSO and QPSO.These seven funtions and their domain are listed

in table 1 and table 2 .The fitness value is set as function value. We had 50 trial runs for every instance. In order to investigate the scalability, different population sizes M are used for each function with different dimensions. The population size is 20. The value of α for QPSO and HSQPSO is 1->0.5.Because HSQPSO has one more comparation than the other two algorithms,the generation of HSQPSO is set as 1000 and the generation of the other two algorithms are set as 2000.

Table 1: Mathematical Expressions of the Benchmark Functions

Sphere Function	$f_1(X)=\sum_{i=1}^{N} x_i^2$
Rosenbrock Function	$f_2(X)=\sum_{i=1}^{N-1}(100\cdot(x_{i+1}-x_i^2)^2+(x_i-1)^2)$
Rastrigrin Function	$f_3(X)=\sum_{i=1}^{N}(x_i^2-10\cdot\cos(2\pi x_i)-10)$
Griewank Function	$f_4(X)=(1/4000)\sum_{i=1}^{N}x_i^2-\prod_{i=1}^{N}\cos\left(x_i/\sqrt{i}\right)+1$
Ackley Function	$f_5(X)=-20\exp\left\{-0.2\sqrt{(1/N)\sum_{i=1}^{N}x_i^2}\right\}-\exp\left\{(1/N)\sum_{i=1}^{N}\cos(2\pi x_i)\right\}+20+e$
Shaffer's Function	$f_6(X)=0.5+\frac{\left(\sin\left(\sqrt{x_1^2+x_2^2}\right)\right)^2}{\left(1.0+0.001\left(x_1^2+x_2^2\right)^2\right)}$
Schwefel Function	$f_7(X)=\sum_{i=1}^{N} x_i\sin(\sqrt{\lvert x_i\rvert})$

Table 2: Up and Lower Bounds of the Benchmark Functions

f_1	f_2	f_3	f_4	f_5	f_6	f_7
±100	±30	±5.12	±600	±32	±100	±500

Table 3:The experiment result of Benchmark Functiion

Function	Method	M	Mean	R
f_1	PSO	10	3.1073e-057	9.8260e-058
		20	1.5231e-004	2.8743e-005
		40	385.4261	691.1253
	QPSO	10	6.1632e-078	6.4667e-078
		20	1.9221e-029	3.4667e-029
		40	2.4111e-007	1.0693e-006
	HSQPSO	10	1.4433e-074	2.7967e-074
		20	1.1725e-029	2.1933e-029
		40	2.0150e-009	6.4831e-009
f_2	PSO	10	6.8961	5.4197
		20	34.0915	40.7582
		40	2.4791e+004	4.7990e+004
	QPSO	10	3.8353	0.3870
		20	21.6347	17.7944
		40	148.2459	211.8704
	HSQPSO	10	5.1722	0.3056
		20	15.6184	0.3809
		40	58.9014	33.0893
f_3	PSO	10	9.6517	3.9255
		20	46.2702	9.6803
		40	141.0483	32.3937
	QPSO	10	4.6049	2.7890
		20	9.3858	1.5487
		40	60.0924	18.7886
	HSQPSO	10	6.3441	3.1326
		20	11.2032	3.9616
		40	74.0513	35.6363
f_4	PSO	10	0.1116	0.0684
		20	0.2795	0.3274
		40	4.1094	4.8599
	QPSO	10	0.0413	0.0268
		20	0.0352	0.0164
		40	0.0065	0.0085
	HSQPSO	10	0.1113	0.0589
		20	0.0166	0.0438
		40	0.0046	0.0109
f_5	PSO	10	0.5112	0.6749
		20	3.9207	2.0106
		40	11.3531	2.2402
	QPSO	10	2.6645e-015	0
		20	4.7073e-014	9.4800e-014
		40	6.8507e-005	7.8395e-005
	HSQPSO	10	2.6645e-015	0
		20	1.1546e-014	8.2473e-015
		40	6.3421e-006	9.5734e-006
f_6	PSO	10	0.0058	0.0050
		20	0.0049	0.0051
		40	0.0043	0.0049
	QPSO	10	0.0049	0.0054
		20	0.0039	0.0048
		40	0.0019	0.0039
	HSQPSO	10	1.4706e-006	3.5607e-006
		20	1.9519e-004	0.0014
		40	1.9514e-004	0.0014
f_7	PSO	10	-4.3927e+069	3.0450e+070
		20	-2.3834e+069	1.6078e+070
		40	-2.2406e+068	7.7581e+068

	QPSO	10	-3.5012e+003	300.6398
		20	-6.5826e+003	257.1356
		40	-1.1085e+004	1.8959e+003
	HSQPSO	10	-3.4977e+003	388.466
		20	-5.4892e+003	1.1590e+003
		40	-7.6974e+003	3.1820e+003

Figure1-Figure7 show performances of algorithms for 40-dimensional f1-f7 functions.We record mean best fitness values for 50 runs of each functions in table 3 define as Mean. R is the standard deviation of Mean in 50 runs. By compare the results, it is easy to see that QPSO and HSQPSO have better results than PSO,and HSQPSO showed a distinct advantage when the problem's dimension increases. In addition, HSQPSO algorithm has better stability than the other algorithm.

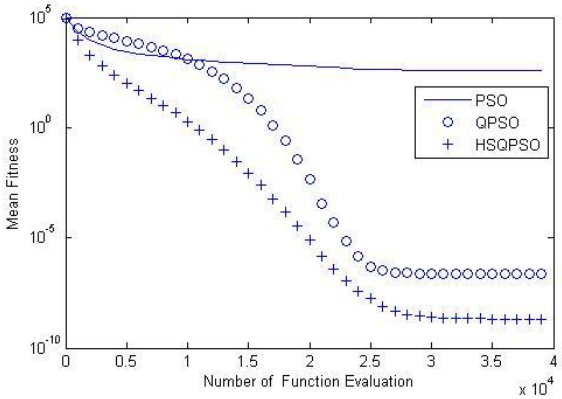


Figure1. Performance of the algorithms for 40-dimensional f1 function

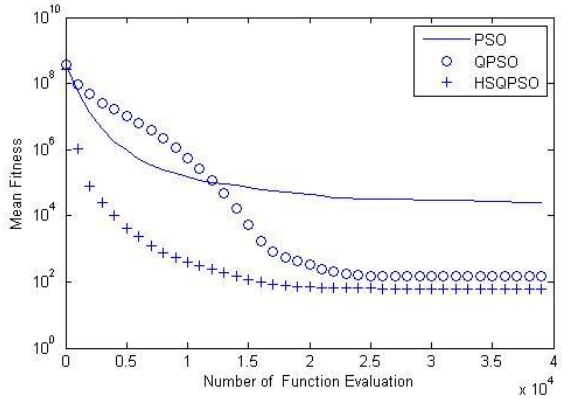


Figure2. Performance of the algorithms for 40-dimensional f2 function

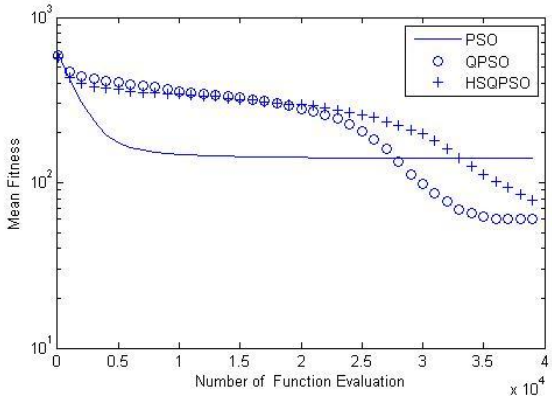


Figure3. Performance of the algorithms for 40-dimensional f3 function

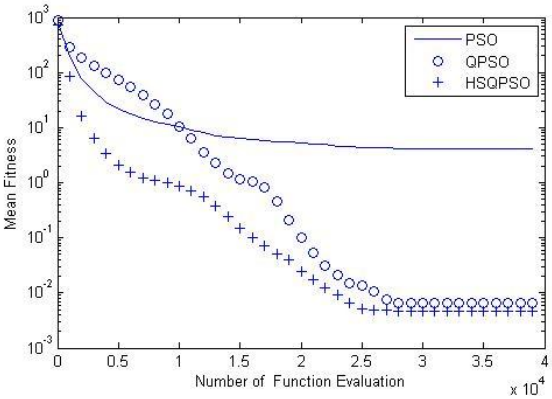


Figure4. Performance of the algorithms for 40-dimensional f4 function

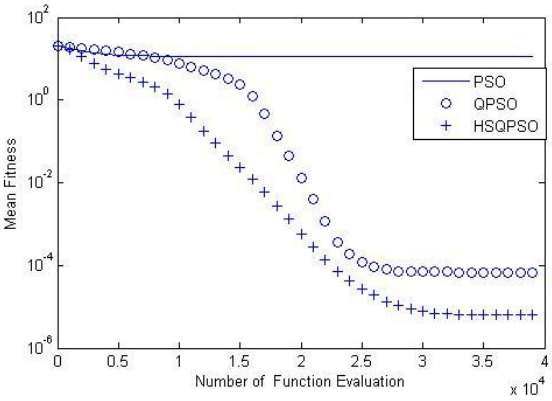


Figure5. Performance of the algorithms for 40-dimensional f5 function

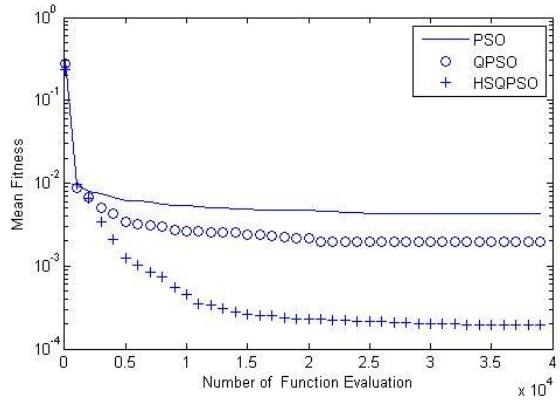


Figure6. Performance of the algorithms for 40-dimensional f6 function

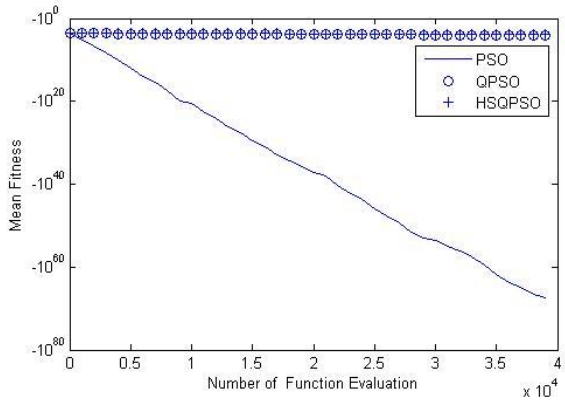


Figure7. Performance of the algorithms for 40-dimensional f7 function

VI. CONCLUSION

In this paper, based on the QPSO,we propose the HSQPSO. The algorithm is a probability searching technique, or says it accurately, an uncertainty searching algorithm. The advantage of PSO is it has ability of fast convergence. The feature of QPSO is it can protect the species diversity and have good local search ability. Because of introducing chaos search ,HSQPSO can search more precise in the global optimal solution region and find the optimal solution. And while strengthening the global search ability,HSQPSO also remains the excellent local search ability of QPSO. To some extent, HSQPSO is superior to QPSO and PSO.

REFERENCES

[1] J. Kennedy and R. Eberhart,,"Particle swarm optimization," Proc. IEEE int. Conf. On Neural Network, 1995: 1942-1948J.

[2] Kennedy. and R. Eberhart, "Particle swarm optimization," Proc. IEEE int. Conf. On Neural Network, 1995: 1942-1948

[3] J. Kennedy,"The behavior of particles,"The seventh annual conf. on evolutionary computation, 1997, pp303-308

[4] J. Kennedy,"The behavior of particles,"The seventh annual conf. on evolutionary computation, 1997, pp303-308

[5] J. Kennedy,"The particle swarm: social adaptation of knowledge, "IEEE int. Conf. On Evolutionary Computation, 1997, pp. 303-308

[6] Shi. Y., Eberhart, R.,"Empirical Study of Particle Swarm Optimization,"Proc. Congress on Evolutionary Computation. Piscataway, NJ (1999) 1945-1950

[7] Shi, Y., Eberhart, R.C,"A Modified Particle Swarm. Proc," 1998 IEEE International Conference on Evolutionary Computation. Piscataway, NJ (1998) 69-73

[8] J. Sun, B. Feng and W. Xu,"Particle Swarm Optimization with Particles Having Quantum Behavior," IEEE Proc. of Congress on Evolutionary Computation, 2004.

[9] J. Sun, J. Liu and W. Xu," Using quantum-behaved particle swarm optimization algorithm to solve non-linear programming problems, "International Journal of Computer Mathematics ,Vol. 84, No. 2, February 2007, 261–272

[10] Ying Song, Zengqiang Chen, and Zhuzhi Yuan,"New Chaotic PSO-Based Neural Network Predictive Control for Nonlinear Process," IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 18, NO. 2, MARCH 2007

[11] J. Sun, B. Feng and W. Xu,"A Global Search Strategy of Quantum-behaved Particle Swarm Optimization,"Proc. 2004 IEEE Conference on Cybernetics and Intelligent Systems, Singapore (2004) 111-115

[12] Clerc, M."The Swarm and Queen: Towards a Deterministic and Adaptive Particle Swarm Optimization,"Proc. 1999 Congress on Evolutionary Computation. Piscataway, NJ (1999) 1951-1957

[13] M.Clerc,J.Kennedy,"The Particle Swarm: Explosion, Stability and Convergence in a Multi-dimensional Complex Space,"IEEE Transactions on Evolutionary Computation, Vol. 6, No. 1. Piscataway, NJ (2002) 58-73

[14] Suganthan, P.N,"Particle Swarm Optimizer with Neighborhood Operator,"Proc. 1999 Congress on Evolutionary Computation, Piscataway,NJ(1999)1958-1962

Study on Automatic Test Generation of Digital Circuits Using Particle Swarm Optimization

Gu Yuan-liang
IoT Engineering
Jiangnan University
Wuxi, Jiangsu, China
guyuanliang2004@sina.com

Xu Wen-bo
IoT Engineering
Jiangnan University
Wuxi, Jiangsu, China
xwb@sytu.edu.cn

Abstract: The development of digital integrated circuit has put forward urgent demands for test technology. Test technology has become a bottleneck in the application of LSI/VLSI. Especially for sequential circuits, it is still a problem which is not resolved completely in theory. By making use of the structure information of circuits, a method of automatic test generation for sequential circuits based on Particle swarm optimization is presented, which is performed by two steps, initialization and fault detection. Experimental results show that the approach can achieve high fault coverage, and CPU times needed for test generations are very short, which shows that it is a method deserving research.

Key words: Automatic Test Generation; Sequential Circuits; Particle Swarm Algorithm

I. Introduction

For decades, scholars from various countries have done a lot of work in the sequential circuit test generation, but the timing circuit test is still a recognized problem.

According to the existing timing circuit test generation algorithms, deterministic algorithm is suitable only for small-scale processing circuit, such as HITEC[1] ; algorithm based on symbols and the state table has relatively poor fault coverage, the time of test generation is longer, but it can be a very simple test set, such as Symbat[2] ; on the other hand simulation-based algorithm, for sequential circuits of all sizes for test generation can get a better fault coverage, the problems faced is the lack of activation of unpredictable failures and dissemination of the necessary corresponding fault information, so often results in longer test generation, such as CRIS[3] ,GATEST[4].

Comprehensive comparison of above algorithms, simulation-based test generation is better than the other two algorithms.

Here primarily study the new algorithm of simulation-based automatic test generation of sequential circuits: apply the particle swarm algorithm to sequential circuit test pattern generation.

Particle Swarm Optimization (PSO) algorithm is similar to other evolutionary algorithms, but also use a "group" and "evolution" concept, according to the individual's fitness value, guide the search through cooperation and competition among particles produced by groups of intelligent. PSO algorithm is simple and easy to implement, requires less parameters to adjust, its unique memory function so that it can dynamically track the current search conditions to adjust the search strategy. It is a more efficient parallel evolutionary algorithm and is widely used in function optimization, pattern recognition, signal processing and other fields.

II. Automatic test generation based on PSO algorithm

A. Algorithm description

PSO algorithm is co-sponsored by a social psychologist from the United States and an electrical engineer, James Kelledy and Russell Eberhart, the basic idea is subject to the inspiration of the results which they early modeled and simulated the behavior of many bird groups. Birds use simple rules to determine their flight direction and flight speed, when a bird flew away from the birds and habitat, the result of other birds flying habitat, making falls more and more habitat for birds until the whole flock fall habitat. The bird is abstracted as a particle which do not have mass and volume,

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.84

these particles can still flight at a certain speed in the search space like the birds, and the flight speed of the particle dynamically adjust by its own and the group's flying experience, All particles have a decision by the fitness value of the objective function, and know their current location and the best position which are found so far, also know the best position which the whole particles of the entire group have found till now. Particles will use the following information to change their current position: (1) Current position (2) Current speed (3) the distance between current location and the best position of their own (4) the distance between current location and the best position of the group. The optimization search of PSO algorithm is that a group of particles search the best position by the iterative way, until fly to the best position and get optimal solution.

B. Evolvement regulation

Set $X_i=(x_{i1}, x_{i2},\ldots, x_{in})$ as the position of the particle i; set $V_i=(v_{i1},v_{i2},\ldots, v_{in})$ as the current flying speed of the particle i; set $P_i=(p_{i1}, p_{i2},\ldots,P_{in})$ as the best position it has passed, which is called individual best position "pbest"; set $G_i=(g_{i1}, g_{i2},\ldots, g_{in})$ as the best position all particles have passed, which is called global best position "gbest".

Global best position "gbest" improves the convergence speed of the algorithm at the expense of cost the robustness, the whole algorithm makes "gbest" as the attractor and pulls all the particles to it.so that all particles converge on the location. If during the evolutionary process, "gbest" is not effectively update, it will be premature. "Pbest" uses the multi-attractor to replace the single attractor in "gbest", thus prevent "gbest" from being possibly premature.

The equation of the evolution of the basic PSO algorithm is as follows:

$$v_{ij}(t+1)=v_{ij}(t)+c_1r_1(p_{ij}(t)-x_{ij}(t))+c_2r_2(g_{ij}(t)-x_{ij}(t)) \quad (2\text{-}1)$$

$$x_{ij}(t+1)=x_{ij}(t)+v_{ij}(t+1) \quad (2\text{-}2)$$

There subscript "i" denotes particle "i", "j" shows j-dimension of the particle, "t" represents the t generation, cognitive factors c1 and c2 are to adjust the weight of the particle's own and the group's experience, r1 and r2 are the two random number.

In the article the PSO algorithm is used for test generation of the digital integrated circuits, circuit input can only take discrete values 0 or 1, and therefore should use the discrete PSO algorithm type, make a few changes in the evolution equation of the basic PSO algorithm, can obtain the evolution equations of PSO algorithm of discrete variables with the inertia weight:

$$w=w_0-(w_0-w_f)\bullet t/T \quad (2\text{-}3)$$

$$v_{ij}(t+1)=w\bullet v_{ij}(t)+c_1r_1(p_{ij}(t)-x_{ij}(t))+c_2r_2(g_{ij}(t)-x_{ij}(t)) \quad (2\text{-}4)$$

$$sig(v_{ij}(t+1))=1/(1+\exp(-v_{ij}(t))) \quad (2\text{-}5)$$

$$x_{ij}(t+1)=\begin{cases} 0 & rand \geq sig(v_{ij}(t+1)) \\ 1 & other \end{cases} \quad (2\text{-}6)$$

Here W is the inertia weight, W_0 is the initial value of W, W_f is the final value of W, T is the maximum number of iterations. Rand is random number between (0,1), the meaning of other variables is the same as with the basic PSO algorithm, and the probability the particle i to sig (vij (t)) is 1, the probability the particle i to 1- sig (vij (t)) is 0.

C. Parameter analysis

The parameters which affect convergence of PSO algorithm are inertia weight W, cognitive coefficients c1 and c2, group size and the number of iterations.

Inertia weight W used to balance the global search ability and local search capabilities, so that the particles keep moving inertia and the trend has extended the search space. The smaller W is, the smaller V is, which benefits in the current solution space to find a good solution, so that the stronger the ability of the local search is; On the contrary the greater W is, the larger V is, which is useful for the particles search more space, so that the stronger global search ability is. Therefore, the value of W can be adjusted to control the impact the current speed on the previous speed, and thus make the global search and local search to achieve a balance. Generally set W higher in the previous search, so that particles search larger space to find a more outstanding individual; Set W smaller in the late search, so that the particles can search in the better regional. In the article let W linearly change with the evolution of algebraic manipulation (2-3), set w_0=0．9 and W_f=0.4.

Cognitive coefficients c1 and c2, c1 is the best place to adjust the particle toward the direction of its own steps, which is its own experience, c2 to regulate global best particle position toward the direction of the step, which is a group experience. When the individual found that the experience of partner is

better, it will make adjustments, we will progress towards the direction of the common understanding at the same time. If c1 = 0, then the particle does not have their own experience, only the group experience, and its convergence rate may be faster, but in dealing with more complex problems, easily fall into local optimal solution; If c2 = o, then the particle does not have group experience, and only has their own experience, there is no sharing of information between individuals, then M contains the movement of groups of individuals is equivalent to the M individuals separate exercise behavior, so little access to the most optimum solution; If c1 = c2 = O, then the particles can be used with no experience, the movement appears to be disorganized. Based on previous experience in all the simulations below are taken c1 = c2 = 2.

Group size and method of initial population will also affect the PSO algorithm convergence speed and power. Large group size ensures that there are some differences between individuals, expanding the search space, help in finding the optimal solution, but the time of each iteration is longer , the two trade-offs need to be considered. A common method to initialize the population is random generation. A common method to initialize the population is random generation, in order to accelerate the convergence to the optimal solution, save the search time, and increase the search efficiency, the initial group should be purposely set based on the issues optimized. Number of evolution generation is more conducive to find the global optimal solution, but increase the amount of computation, make running time longer, the maximum evolution generation, as the termination of a prescribed condition, is determined on the general level of difficulty according to the specific problem.

D. Algorithm implement

Optimization problems of PSO algorithm are as follows:

Step 1: Set the group size, according to the characteristics of the possible solutions domain of the problem needed to optimize, mapped the expression of the problem domain to the space of the particle swarm;

Step 2: Initially set the position and speed of the particle swarm;

Step 3: According to the specific optimization problem, set the fitness function, calculate the fitness of each particle; then initialize pbest and gbest;

Step 4: For each particle, calculate its fitness value, and compare with the fitness value of the best position “pbest” which the particle has experienced, if better, make it as the current best position of particle;

Step 5: Compare the fitness of each particle to the global best position “gbest” which the particle has experienced, if better, make it as the current global best position;

Step6: Update the velocity and position of the particle, according to the Evolution equation of the particle;

Step 7: If fail to reach the final condition (Usually is a fitness value which is good enough or reach a preset maximum pre-set), then return (4), or jump into (8);

H Get the global best position, which is the best optimal solution, mapped the solution back to the problem solution space, output the solution of the problem to be solved.

III. Test results

The object of the simulation of the algorithm used the sequential circuit iscas'89 of the international standards and several synchronous sequential circuits.

Table 1 is based on the initial results of PSO algorithm.

Table 2 is based on particle swarm optimization results of the automatic test generation and some typical results of the test generator.

Here HITEC is very representative of the test generator based on deterministic algorithm; CRIS[3]、GATEST[4] is the test generator based on GA; DIGATE[6]、STRAGHTGATE[7] is the test generator which crossed GA and the deterministic algorithm.

Table 1 Initialization results

Circuit Name	Trigger Quantity	Initialization Quantity	Sequence length	Time consumed（s）
S298	14	14	2	0.00
S344	15	15	2	0.00
S382	21	21	1	0.017
S444	21	21	1	0.017
S526	21	21	2	0.00
S641	19	19	1	0.017

S713	19	19	1	0.033
S820	5	5	1	0.033
S832	5	5	1	0.017
S1196	18	18	1	0.017
S1238	18	18	1	0.017
S1269	37	31	4	0.167
S1423	74	74	3	0.033
S1488	6	6	1	0.050
S1494	6	6	1	0.033
S5378	179	179	14	0.467
S35932	1728	1728	1	2.617

Table 2 Test generation results

Circuit name	Total number of faults	HITEC number of detected faults	GATEST number of detected faults	CRIS number of detected faults	DIGATE number of detected faults	STRATEGATE number of detected faults	PSO	
							number of detected faults	CPU Time(s)
S298	308	265	264	253	264	(265)	(265)	0.35
S344	342	328	(329)	328	(329)	(329)	(329)	2.9
S382	399	363	347	273	363	(364)	(364)	6.233
S444	474	414	405	397	420	(424)	(424)	5.717
S526	555	365	417	428	446	(454)	(454)	33.967
S641	467	(404)	(404)	398	(404)	(404)	(404)	3.85
S713	581	(476)	(476)	475	(476)	(476)	(476)	4.633
S820	850	813	517	451	621	(814)	474	82.6
S832	870	817	539	370	606	(818)	438	83.15
S1196	1242	(1239)	1232	1180	1236	(1239)	1236	82.6
S1238	1355	(1283)	1274	1229	1281	1282	1279	58.217
S1488	1486	(1444)	1392	1355	1378	(1444)	1432	34.283
S1494	1506	(1453)	1416	1357	1354	(1453)	1439	26.627
S5378	4603	3238	3175	3029	3447	(3639)	3233	388.88
S35932	39094	34902	35009	34481	35100	35100	(35101)	342.38
Am2910	2391	2164	2163	-	2195	(2198)	(2198)	238.38
Mult16	1708	1640	1653	-	1664	1665	(1666)	28.438
Div16	2147	1665	1739	-	1802	(1814)	1767	255.23
Pcont2	11300	3354	6826	-	6837	(6837)	6795	22.633
Piir8o	19920	14221	15013	-	(15072)	15071	(15072)	177.06
Piir8	29689	11131	-	-	18140	(18206)	18141	271.65

The value marked with () in Table 2 is the highest value of fault coverage of the six algorithms. Comparing with the results of the table, PSO has obvious advantages in the small-scale circuit. For s298, s344, s382, s444, s526, s641, s713, PSO also gives high fault coverage, better than GATEST, CRIS, HITEC, DIGATE, has achieved the very high fault coverage and is very close to the results of STRATEGATE of the highest fault coverage, indicating that the particle swarm algorithm has the outstanding advantage of test generation of the small-scale circuit. In some large-scale circuits it also has the highest fault coverage, such as s35932, mult16, piir8o. For s35932, mult1 even has received better results than that the other Builders can get. For the unpredictable fault circuit s444, s526, it also has the best fault coverage. In addition, the speed of PSO algorithm of test generation is very fast, Table 2 also shows the CPU time which the particle swarm algorithm is used for test generation. In some large-scale circuits also has the highest fault coverage, such as s35932, mult16, we can see that the time of particle swarm algorithm of test generation are all second level, the longest time is no more than 7min, compare to the time of test generation of the calculation based on the unit of minute even hour, the algorithm takes a very short time.

IV. Conclusion

From the experimental results we can see that automatic test generation of Sequential Circuits based on particle swarm algorithm is full use of the circuit structure information, in the small-scale circuits it can get very high fault coverage, in some large-scale circuits it can make very good results and the speed of the test generation is very fast. However, the generated test vectors are longer than that generated by the other five generators, in the next step should

compress the vectors by using of dynamic or static compressive method.

References

[1] T.M.Niermann,J.H.Patel. HITEC:A test generation package for sequential circuits. Proc.European Conf. Design Automation (EDAC),2007,214～218.

[2] G.Cabodi,P.Camurati,S.Quer. Symbolic exploration of large circuits with enhanced forward/backward traversals. Proc.EURODAC,2004,22～27.

[3] D.G.Saab,Y.G.Saab,J.A.Abraham. CRIS: A test cultivation program for sequential VLSI circuits. Proc.Int.Conf.Computer-Aided Design,2002,216～219.

[4] E.M.Rudnick,J.H.Patel,G.S.Greenstein,T.M.Niermann. Sequential circuit test generation in a genetic algorithm framework.Proc.Design Automation Conf.,2004,698～704.

[5] Li Zhi, Xu Chuanpei, Chen Guang. Based on Particle Swarm initialization of Synchronous Sequential Circuits. Electronic Measurement and Instrument, 2002,(4):33～38.

[6] M.S.Hsiao,E.M.Rudnick,J.H.Patel. Automatic test generation using genetically-engineered distinguishing sequences. Proc.VLSI Test Symp.,2006,216～223.

[7] M.S.Hsiao,E.M.Rudnick,J.H.Patel. Dynamic state traversal for sequential circuit test generation. Proc.VLSI European Design and Test Conf.,2007,22～28.

[8] Naoki Shiramatsu. Improvement in measuring the luminance distribution of the CRT electron beam spot. SID(DIGEDT),2008,624～627.

[9] Nijs C. Van der Vaart. Anovel electron source for CRTs. Information Display(SID),2005,18(6):14～17.

A suboptimal CI algorithm and its application to distributed target tracking

Hai-Xiao Cui, Xiao-Jun Wu, Xiao-Qing Luo
School of IOT engineering
Jiangnan university
Wuxi, China
haixiaocui@hotmail.com
wu_xiaojun@yahoo.com.cn
qingqingchong@126.com

***Abstract*—When the fused variables are independent or the statistics of variables are known perfectly , Kalman filter is rigorous and yields minimum mean squared error estimate. But in most situations, it is impossible to guarantee that the fused variables are independent and even might be highly correlated. In that case, it is possible to "over estimate" the statistics.CI algorithm can achieve a consistent estimation, when the correlation between the fused variables are unknown. However the caculation of optimal ω of CI costs lots of time. A simple way to calculate a suboptimal ω is proposed to reduce the time complexity and get a suboptimal estimation. According to the geometrical explanation of CI, a simple calculation equation of ω is given. And the functions of three kinds of fusion algorithm are illustrated in an application of decentralized estimation with circle topology, where it is impossible to consistently use a Kalman filter or other algorithms that need independent constraints.**

Keywords-convariace intersection; simple convex combination; distributed system; data association

I. INTRODUCTION

The main function of a filter is to get a reliable estimation according to data fusion algorithms from measurements corrupted by noise. Kalman filter [1] is one of the classic methods in target tracking area, which has wide applications under linear Gaussian condition. Probability Data Association (PDA) is developed on the basis of Kalman filter to deal with multi-measurement, which can be used in clutter environment for single target tracking. Both Kalman and PDA algorithm [2][7] can get optimal estimation result under the assumption of independence or the accurate statistical properties of data available. Simple Convex Combination algorithm [3] of trace fusion also needs information independence. However, in practice this assumption is not necessarily held. The Bar Shalom-Campo fusion algorithm [4] was proposed to deal with the correlation problem between data, which takes the correlation extent into consideration. The calculation of data correlation is complex, even unrealizable. Assumption that the data are independent can get a sub-optimal estimation result or even lead to divergence. In order to solve the correlation problem between data, Simon J. Julier and Jeffrey K. Uhlman proposed the Covariance Intersection Algorithm(CI) [5][6], which can, by just ignoring the correlation between data , get a consistent estimation and avoid complex calculation . But calculating the optimal ω needs plenty of time, this paper proposes a simple way to get a sub-optimal ω with reduced time complexity but acceptable accuracy, here we call the sub-optimal calculation method "simple covariance intersection"(SCI) algorithm.

This paper is organized as follows. First CI algorithm is given in Sec.2 and then suboptimal CI method is proposed in Sec.3. The main simulation results of different data fusion algorithms are presented in Sec.4. Finally, some conclusions are drawn in Sec.5

II. THE DESCRIPTION OF CI ALGORITHM

We consider the problem that two pieces of information, labeled X and Y, are to be fused together to get an estimation Z. X and Y could be from different sensor measurements or X could be an estimation from a fusion model and B could be sensor measurement. Using random variables **a**, **b** and covariance A, B represent X and Y respectively. We assume that the true statistics of these variables are unknown. The only information we can get are consistent estimates of the means and covariance of variables **a** and **b**, labeled as $\overline{a}, \overline{b}$ and $\overline{A}, \overline{B}$. The consistency constraint of CI algorithm is as follows:

$$A - B \geq 0$$
$$B - \overline{B} \geq 0$$

This definition conforms to the standard definition of consistency[8], i.e. A≥B （ or $A - B \geq 0$ ） if and only if $A-B$ is positive semidefinite. The fused output Z, represent by $\{c, C\}$, satisfies this consistent constraint:

$$C - \overline{C} \geq 0$$

where

$$\overline{A} = E\left[\tilde{a}\tilde{a}^T\right]$$
$$\overline{B} = E\left[\tilde{b}\tilde{b}^T\right]$$

$\tilde{a} = a - \overline{a}$ and $\tilde{b} = a - \overline{b}$ represent the residual error of X and Y respectively.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.16

The covariance intersection characterized by the convex combination is given by

$$p_{cc}^{-1} c = \omega p_{aa}^{-1} a + (1-\omega) p_{bb}^{-1} b \quad (1)$$

$$p_{cc}^{-1} = \omega p_{aa}^{-1} + (1-\omega) p_{bb}^{-1} \quad (2)$$

where the optimal ω, constrained between 0 and 1, is calculated by (2) according to different optimization strategy. CI algorithm can maintain a conservative estimation without considering the correlation between variables i.e. the covariance calculated through CI algorithm is bigger than the actual one which is almost impossible to obtain.

III. THE CALCULATION METHOD OF SUB-OPTIMAL ω

The most common data fusion algorithms compute a linear combination of the means, and then analytically determine the covariance of the result. The Kalman filter uses a linear update rule of the form

$$\bar{c} = W_a \bar{a} + W_b \bar{b} \quad (3)$$

and the covariance is

$$p_{cc} = W_a p_{aa} W_a^T + W_a p_{ab} W_b^T + W_b p_{ba} W_a^T + W_b p_{bb} W_b^T \quad (4)$$

The intuition of CI approach arises from a geometric interpretation of (4). If we plot out the covariance ellipses[1] of P_{aa}, P_{bb} and P_{cc}, labeled respectively as $T(P_{aa})$, $T(P_{bb})$ and $T(P_{cc})$, for all choices of P_{ab}, the trace of P_{bb} always lies within the intersection of P_{aa} and P_{bb}.

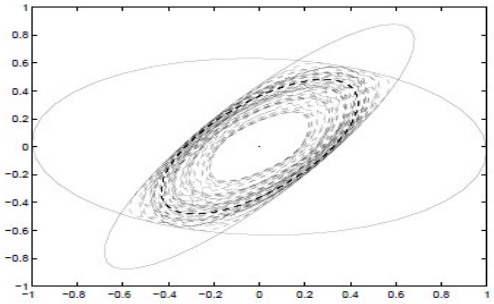


Figure 1. The variances of P_{aa} and P_{bb} are the solid ellipses. Different values of P_{cc} which arise from different choice of P_{ab} are shown as dashed ellipses.[5]

This interpretation suggests a fusion approach: if P_{cc} lies within the intersection of P_{aa} and P_{bb} for any choice of P_{ab}, then an update strategy which finds a P_{cc} which always encircles the intersection region without considering P_{ab}, and convex combination can satisfy this requirement. CI algorithm, which finds an optimal ω to minimize the ellipse of P_{cc}, is proposed.

From Fig 1, we think that the data fusion process of CI algorithm is equal to the process of calculating a minimized ellipse trace, which encloses two given ellipses' intersection region:

$$T(C) = \omega * T(A) + (1-\omega) * T(B) \quad (5)$$

where $T(X)$ represents the ellipse trace of matrix X.

Finding an ω, constrained between 0 and 1, minimizes $T(C)$. Obviously, there are two special cases, one is when $T(A) << T(B)$, ω is close to 1; the other is when $T(A) = T(B)$, ω equals to 1/2. The two cases can be described as:

$$\dot{\omega} * T(A) - (1-\dot{\omega}) * T(B) = 0 \quad (6)$$

where $\dot{\omega}$ represents a sub-optimal value of ω. The (6) can also be expressed as:

$$\dot{\omega} = T(B) / (T(A) + T(B)) \quad (7)$$

Obviously, there are $T(A) \geq 0$ and $T(B) \geq 0$, so the value of $\dot{\omega}$ gained from (7) always falls in constrained range. For the ellipse track of covariance here is not an computable object, here we use matrix determinant instead. Determinant value of a matrix is proportional to the size of its corresponding elliptical orbit. So (7) can be written as

$$\dot{\omega} = \det(B) / (\det(A) + \det(B)) \quad (8)$$

where $\det(X)$ represents the determinant of matrix X.

And this sub-optimal calculation method of $\dot{\omega}$ has advantage over the optimal method under certain condition. When matrix A equals or closes to B, the value of ω is 1 according to the optimal computing method, that means one of the two states information is simply abandoned. In contrast, using the sub-optimal method the value is 1/2, and both of state information is used in data fusion process. The latter is more reasonable, and gets a better estimated accuracy.

IV. EXPERIMENTAL MODEL AND RESULTS

Given a distributed system with four-node ring topology, each node has its own independent observation and fusion function. Information transfers between adjacent nodes (according to counter-clockwise). Node 1 and Node 3 have the location measurement of object, while Node 2 and Node 4 have velocity measurement. For simplicity, we assume that the dynamic model is the same for 4 nodes, as

1 For a covariance matrix P, its ellipse is the locus of points $\{ x^T P^{-1} x = c \}$, where C is a constant.

$$x_i(k+1) = Fx_i(k) + Gv_i(k+1)$$

where

$$F = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } G = \begin{bmatrix} T^2/2 & 0 \\ 0 & T \\ T^2/2 & 0 \\ 0 & T \end{bmatrix}$$

The observation model is

$$Z_i(k) = H_i x_i(k) + w_i(k)$$

when i=1 or i=3

$$H_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

when i=2 or i=4

$$H_i = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$v_i(k)$ is an uncorrelated, zero-mean Gaussian noise with variance $\sigma_v^2 = 4$. The variance of noise w_i is σ_w^2=1.

First, we predict the state of node i at time k+1, according to its measurements, using PDA algorithm [1] to get the distributed estimate. Then node i use state fusion algorithm to fuse the distributed estimate, transferred from other nodes, with its own. The specific steps are as follows:

1) According to the state of node i at time k and , predict the state and covariance of node i at time k +1:

$$\hat{x}_i(k+1|k) = F\hat{x}_i(k|k)$$

$$p_i(k+1|k) = FP_i(k|k)F^T + GG^T\sigma_v^2$$

2) According to the tracking gate, identify the effective echo at time k+1 from the measurements $z_{i,(k+1)}$, the number of effective measurement is denoted as $m_{i,(k+1)}$.

3) Calculate the association probability of each effective echo $\beta_{i,j}$, where j=1,..., $m_{i,(k+1)}$, the distributed estimate can be gained by the follow equation:

$$\hat{x}_i^*(k+1|k+1) = \hat{x}_i(k+1|k) + W_i(k+1)v_i(k+1)$$

where $W_i(k+1)$ is Kalman gain; $v_i(k+1)$ is innovation information which is computed as

$$v_i(k+1) = \sum_{j=1}^{m_{i,(k+1)}} \beta_{i,j} v_{i,j}(k+1)$$

And the corresponding covariance is calculated as

$$p_i^*(k+1|k+1) = \beta_{i,0} p_i(k+1|k) + [1-\beta_{i,0}]P^c(k+1|k+1) + \bar{P}(K+1)$$

where

$$\bar{P}(k+1) = W_i(k+1)[\sum_{j=1}^{m_{i,(k+1)}} \beta_{i,j} v_{i,j}(k+1)v_{i,j}^{'}(k+1) - v_{i,j}(k+1)v_{i,j}^{'}(k+1)]W_i^{'}(k+1)$$

$$P^c(k+1|k+1) = [1 - W_i(k+1)H_i]P_i(k+1|k)$$

4) Node i sends its distributed estimate to the nodes connected with it, and also accepts data from other nodes. Node i yields the new estimate $\hat{x}_i(k+1|k+1)$ and covariance $p_i(k+1|k+1)$ using state fusion algorithm.

Experiments with different data fusion algorithm are done respectively:

1. Using CI algorithm to calculate the optimal ω, the estimation results are given, and Fig.2 through Fig.5show the fusion error of four nodes as follows:

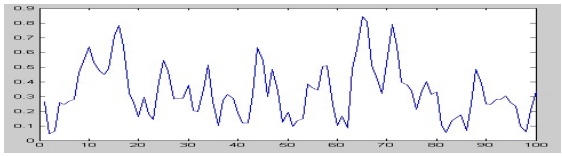


Figure 2. estimation result for node 1

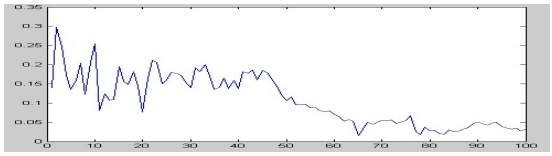


Figure 3. estimation result for node 2

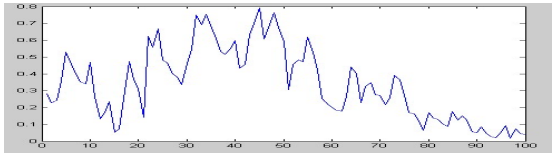


Figure 4. estimation result for node 3

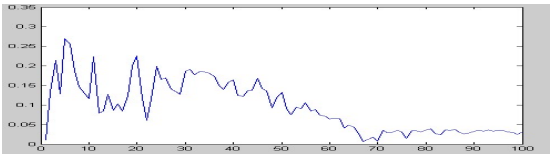


Figure 5. estimation result for node 4

Fig.2 through Fig.5 show the estimation result of CI algorithm. Unlike the approach above, all the nodes are consistent and observable. And especially for node 2 and node 4 the results are convergence as the propagation. This means CI algorithm works when the fused variables are correlated.

2. Using the sub-optimal SCI algorithm to calculate ω, the estimation results are given, in Fig.6 through Fig.9 ,which show the fusion error of the four nodes as follows:

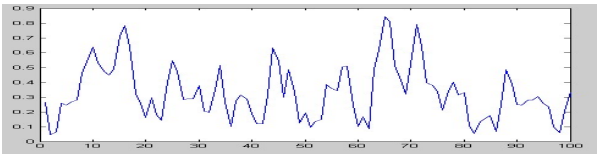


Figure 6. estimation result for node 1

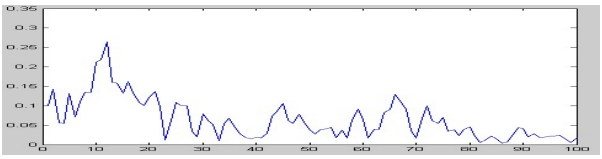


Figure 7. estimation result for node 2

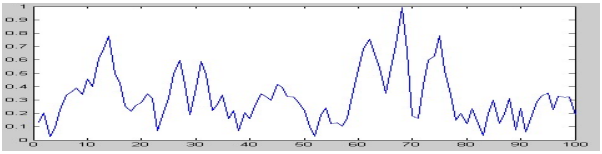


Figure 8. estimation result for node 3

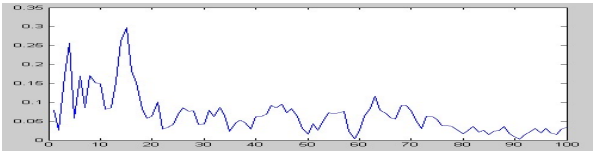


Figure 9. estimation result for node 4

As we can see from Fig.6 to Fig.9, the results of SCI are also consistent for four nodes. Although the squared error of each node is not as good as that of CI algorithm, it is still acceptable especially, when the time complexity is taken into account.

Table 1 shows the mean squared error of each node and the overall computational time using different data fusion algorithm after sampling 100 times.

TABLE I. THE COMPARISON OF FUSION RESULTS OF EACH NODE USING DIFFERENT DATA FUSION ALGORITHM.

Project / algorithm	Node 1	Node2	Node 3	Node 4	Run time(t)
CI	0.0593	0.0127	0.0390	0.0316	7.637057s
SCI	0.3367	0.0183	0.1882	0.0338	2.285537s

The data from Table 1 show that optimal CI does get better estimation result than sub-optimal SCI algorithm, especially for node 1 and node 3 whose variances of observation noise are larger than those of node 2 and node 4. Because the observation noise's variances of nodes 2 and node 4 are relatively small, the improvement of accuracy is not as obvious as the other two nodes. In spite of the decreasing of accuracy, SCI algorithm gains great improvement in time complexity against CI algorithm, using 29.93% as much time as CI algorithm. On the whole, SCI has certain advantages in practical applications.

ACKNOWLEDGMENT

This work was supported in part by the following project: Program for New Century Excellent Talents in University of China (No. NCET-06-0487), National Natural Science Foundation of China (No.60973094), Program for Innovative Research Team of Jiangnan University (No. JNIRT0702).Fundamental Research Funds for the Central Universities (Grant No. JUSRP31103).

REFERENCES

[1] press.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Bar-Shalom Y , Tse E. Tracking in cluttered environment with probabilistic data association. Automatica , 1975, 11(9):451-460.

[3] Ng Gee. Wah and Yang Rong, "Comparison of Decentralized Tracking Algotithms", Proceedings of the Sixth International Conference of Information Fusion,2003,177314.

[4] Chang K C, Saha R K, Bar-shalom Y. On optimal track-to-track fusion. IEEE Transaction on Aerospace and Electronic Systems,1997,33(4):1271-1276.

[5] Simon J. Julier , Jeffrey K. Uhlmann. A Non-divergent Estimation Algorithm in the Presence of Unknown Correlations. In Proceedings of the1997American Control Conference. Piscataway,NJ,USA:IEEE,1997.2369-2373.

[6] S. J. Julier and J. K. Uhlmann, "General decentralized data fusion with Covariance Intersection (CI)",in Handbook of Multisensor Data Fusion, D.Hall and J.Llians, Eds. Boca Raton, FL:CRC Press,2001,ch. 12, 12_1-12_25.

[7] Thiagalingam Kirubarajan and Yaakov Bar-Shalom. Probabilistic Data Association Techniques for Target Tracking in Clutter. PROCEEDINGS OF THE IEEE, VOL.92, NO.3, MARCH 2004.

[8] Jazwinski, A.H., Stochastic Processes and Filtering Theory, Academic Press, New York,1970.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

CSHFt: A Composite Fault-tolerant Architecture and Self-adaptable Hierarchical Fault-tolerant Strategy for Satellite System

Hao Zhou, Jingfei Jiang
College of Computer Science
National University of Defense Technology
Changsha, China
e-mail: javy329@hotmail.com

***Abstract*—Nowadays, building parallel system with high-performance commercial off-the-shelf (COTS) chips becomes the main way to improve satellite system performance greatly. System reliability is a tough issue which needs to be solved by more effective fault-tolerant scheme. Centralized fault-tolerant scheme has the risk of single point of failure (SPOF), while distributed fault-tolerant scheme is much complex and introduces large overhead. Both of these traditional methods have their own drawbacks. This paper proposes a composite self-adaptable hierarchical fault-tolerant (CSHFt) scheme which effectively integrates and expands the ideas of centralized and distributed fault-tolerant methods. It constructs a composite and symmetrical system architecture supporting both fault-tolerant methods simultaneously and realizes the self-adaptable hierarchical fault-tolerant strategy. CSHFt scheme executes system fault tolerance in the sequence of 'first centralized then distributed'. System switches its fault-tolerant mode actively according to its performing history. By combining and completing two traditional methods, CSHFt scheme eliminates the risk of SPOF, reduces the overall complexity and overhead of system fault tolerance and enhances system's real-time feature. System failure only occurs when all the system nodes are broken, which makes system highly reliable. Based on the prototype system, we verify the practical CSHFt scheme. Its performance is also analyzed and evaluated.**

Keywords-satellite system; fault-tolerant architecture; composite architecture; fault-tolerant strategy; hierarchical strategy; self-adaptable

I. INTRODUCTION

Satellite system operates in the complex and volatile deep space, where many negative factors may affect system stability. Reliability is the key issue which needs to be properly solved while designing such system. Because of the low performance and high price of the anti-radiation devices, nowadays, researchers often build parallel system with high-performance commercial off-the-shelf (COTS) chips, in order to greatly improve satellite system performance. The design requirements, which no longer consider reliability of the devices, make the fault-tolerant architecture and strategy increasingly important.

Currently, between two different traditional fault-tolerant methods, centralized scheme is widely used because it's easy to implement. However, it has the drawback of single point of failure (SPOF). The distributed method is complex, which also introduces high executing cost, so few relevant practical examples can be found.

In this paper, we propose the CSHFt scheme which effectively integrates and expands the ideas of centralized and distributed fault-tolerant methods. It constructs a composite and symmetrical system architecture consisting of isomorphic system nodes, which supports both these traditional fault-tolerant methods simultaneously. Based on it, functional module framework of system nodes and self-adaptable fault-tolerant strategy are designed.

CSHFt scheme achieves system fault tolerance in the sequence of 'first centralized then distributed'. System switches its fault-tolerant mode autonomously based on its fault tolerance history. In the centralized mode, system quickly selects master node according to the fixed priority, ensuring the real-time feature of system fault tolerance while there is only one master in the whole system. In the distributed mode, system intelligently chooses system nodes which have lower probabilities of being failed to be multi-masters by executing the sort electing algorithm (SEA). This hybrid fault-tolerant approach reduces the overall complexity and fault-tolerant overhead by integrating centralized fault-tolerant method. Meanwhile, it takes advantages of the distributed fault-tolerant method which has strong and sustainable fault-tolerant capability in the multi-master manner. CSHFt scheme reduces the probability of system failure to the probability of all system nodes being failed. It improves system reliability greatly.

Based on the prototype system, we verify the effectiveness of CSHFt scheme by software simulation. The performance of the proposed scheme is also analyzed and evaluated.

II. FAULT-TOLERANT SCHEMES FOR PARALLEL SATELLITE SYSTEM

Parallel structure provides great flexibility for the fault-tolerant design of satellite system. The manner of distributing fault-tolerant logic directly affects system fault-tolerant performance.

Centralized fault-tolerance scheme puts its fault-tolerant logic in a single module or node. Though it simplifies system fault tolerance, it brings in the SPOF problem at the same time, which weakens the sustainability of system fault tolerance [1-7]. SACFt scheme [7] is one of the centralized

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.35

fault-tolerant schemes. The character of it is that it is insensitive to scale change of the processing nodes. It constructs a Master-Slave 2-level fault-tolerant structure. By using anti-radiation devices and adding self fault-tolerant procedure, SACFt scheme reduces the probability of master node being failed. Independent and parallel slave nodes, which support online detection, form a flexible fault-tolerant structure supporting the combination of hot and cold standby. System can reset or replace failed node timely and adjust the parallelism of working nodes on demand. This scheme does not aim to solve specific problems, which makes it really general. But it still does not solve the SPOF problem fundamentally.

Distributed fault-tolerant scheme places its fault-tolerant logic in multiple modules or nodes, creating positive conditions for eliminating SPOF risks. Such method is relatively complex and introduces large overhead, so it is rarely applied in the current practical systems. ESDFt scheme [8] uses isomorphic nodes to compose a symmetric system framework, making it possible for each system node to switch between master and slave freely. In this way, it successfully reduces the probability of system breakdown to the probability of all system nodes being failed, guaranteeing system reliability greatly. By adopting Sort Elect Algorithm (SEA), ESDFt scheme chooses system nodes with lower probabilities of being failed to be masters intelligently and adjusts system fault-tolerant complexity and overhead at the same time, making up the shortcomings of traditional distributed schemes. ESDFt scheme is especially suitable for the fault-tolerant manner which has multi-masters. However, in the single master mode, as for the extra overhead and delay of operating master election, the real-time feature of ESDFt scheme is poor, and the electing cost can't be hidden effectively.

III. COMPOSITE FAULT-TOLERANT ARCHITECTURE

As a great fault-tolerant method worth learning from, ESDFt scheme fully uses of system resources and maximizes the sustainability of system fault tolerance. In order to make up its disadvantages while system only has one master node, we combine it with SACFt scheme and design a more suitable method of electing system master, so as to form a better fault-tolerant scheme.

A. CSHFt Architecture Framework

To combine SACFt scheme and ESDFt scheme effectively and complement each other, a general fault-tolerant architecture which supports both methods should be constructed first.

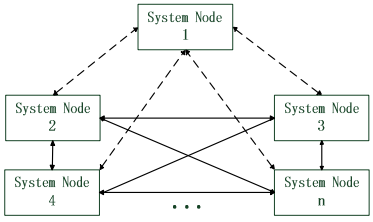


Figure 1. The topology of the CSHFt structure

Figure 1 shows the topology of the CSHFt structure. The broken lines are the fault-tolerant signal channels required when node 1 is the master node in the SACFt scheme. To realize ESDFt scheme, CSHFt structure is composed of isomorphic system nodes, and all pairs of system nodes are connected, making the whole framework as a complete graph topology. Because the signal channels of SACFt scheme only form a subset of the required channels of ESDFt scheme, the final framework of CSHFt architecture is as same as the topology of ESDFt structure.

In the CSHFt architecture, independent and parallel system nodes work as processing nodes or backup nodes. Meanwhile, they perform as master nodes or slave nodes. They can constitute a hot spare structure supporting re-computing, and form a cold standby structure which supports dynamic switching between redundant nodes and processing nodes as well.

B. Functional module framework of isomorphic node

Division and connection of the functional modules of each isomorphic node are shown as Figure 2. Being similar to ESDFt scheme, each system node consists of five parts: the voltage-convert modules, the power-isolate module, the fault-tolerant execute module, the fault-tolerant arbitrate module, CPU and other modules.

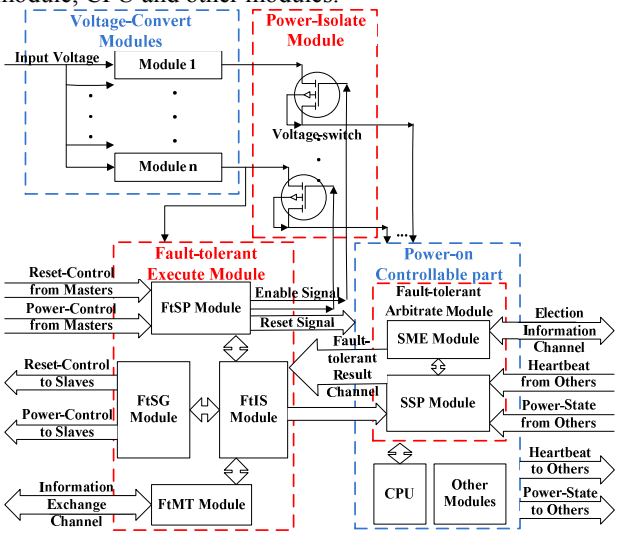


Figure 2. The functional module framework of each system node

Voltage-convert modules convert external input voltage to device supply voltages.

Power-isolate module controls the power states of the power-on controllable part, according to the enable signals from the fault-tolerant execute module.

The fault-tolerant execute module can be further divided into fault-tolerant information storage (FtIS) module, fault-tolerant signal generating (FtSG) module, fault-tolerant signal processing (FtSP) module and fault-tolerant message transceiver (FtMT) module. It generates power-control signals and reset-control signals to all the slave nodes when the system node itself is a master node. While as a slave node, this module responds to the fault-tolerant control from

system master and executes proper operations. Inside the module, there stores some important information related to system fault tolerance, such as fault-tolerant arbitration results, master election results, system node failure records and so on.

The fault-tolerant arbitrate module includes system status processing (SSP) module and system master electing (SME) module. This module monitors each node's current status all the time. While as a master node, it carries out fault-tolerant arbitration, making failed node reset and replacing broken node with backup node. If any master node goes wrong, it elects new master nodes collaborating with CPU.

CPU and other components deal with satellite applications. They also generate heartbeat signal and power-state signal which represent current operating status and power status of each system node.

IV. SELF-ADAPTABLE HIERARCHICAL FAULT-TOLERANT STRATEGY

Based on the fault-tolerant structure mentioned above, system performs autonomous fault-tolerant process shown as Figure 3.

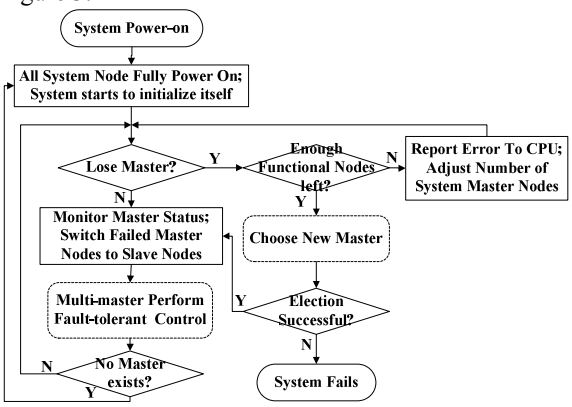


Figure 3. The overall process of the CSHFt strategy

When system executes fault-tolerant control and master election, it determines its current fault-tolerant mode based on its performing history.

As simple SACFt scheme can guarantee effective fault tolerance in most situations, system enters in the centralized fault-tolerant mode first after its initialization. If the master node goes wrong, system transfers its fault-tolerant authority according to the fixed priority which is set beforehand, avoiding the processing overhead introduced by complicated electing operations. In this way, system achieves quick switch between new master and failed master when there is only one master exists in the whole system.

If master node fails frequently in the centralized mode, it implies SACFt scheme can not guarantee system reliability adequately under current conditions. As soon as the times of master failure reaches a certain upper limit, system automatically changes to the distributed fault-tolerant mode, which has stronger fault-tolerant capacity. After that, if master node fails again, system will choose multi-masters according to the SEA. In order to limit the complexity and overhead of system fault tolerance and keep system stay in the multi-master fault-tolerant manner which ESDFt method works well [8], system sets its master number equal to 2 when it enters in the distributed fault-tolerant mode.

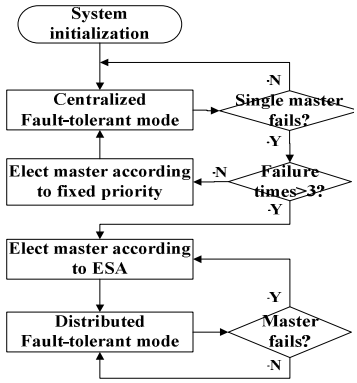


Figure 4. The switching process of system fault-tolerant mode

No matter what fault-tolerant mode system is in, slave nodes response to master's fault-tolerant control according to the Majority Principle Algorithm (MPA) [8]. Taking 'obey majority' as its basic judging principle, MPA makes system take fault-tolerant operations consistent with most system masters, avoiding some negative effects caused by few masters' arbitration mistakes and the noises in signals. It guarantees system fault tolerance be responded correctly.

By realizing the 'first centralized then distributed' hierarchical fault-tolerant process, CSHFt strategy lets system adjust its fault-tolerant mode actively according to the changes of external conditions. As negative influence caused by external factors becomes acute, system automatically switches into distributed mode from centralized mode. Even in the worst situation, CSHFt scheme still can achieve adequate fault-tolerant effect as same as ESDFt scheme. According to reference [8], if the total number of system nodes is larger than 4 and the probability of each node being permanently failed is less than 1%, CSHFt scheme can guarantee the probability of system breakdown lower than 10^{-8}.

V. EXPERIMENTAL SIMULATION

We validate the effectiveness of CSHFt scheme on a satellite system prototype which has 4 isomorphic system nodes. In this prototype, CPU on each node is Freescale's PowerPC MPC8379. Fault-tolerant related functional modules are fulfilled with Xilinx's CoolRunner-II CPLD XC2C256.

Because there are few system nodes in the prototype system, we execute software simulation with these following assumptions: system needs two working nodes while it performs normally; in the centralized fault-tolerant mode, the fixed priority of system nodes to be master is A>B>C>D; once master node fails twice in the centralized mode, system enters in the distributed fault-tolerant mode immediately.

To take a comprehensive test, we examine all these typical situations followed:

(1) Instantaneous failures occur on the slave node in the centralized fault-tolerant mode;

(2) Permanent failures occur on the master node in the centralized fault-tolerant mode;

(3) Permanent failures occur on the slave node in the distributed fault-tolerant mode;

(4) Permanent failures occur on the master node in the distributed fault-tolerant mode;

(5) Processing nodes exceed system demand;

(6) Processing nodes are less than system needs.

Using ModelSim 6.0, we simulate the verilog code of the logic inside the fault-tolerant related modules. Result (only necessary signals) is shown in Figure 5.

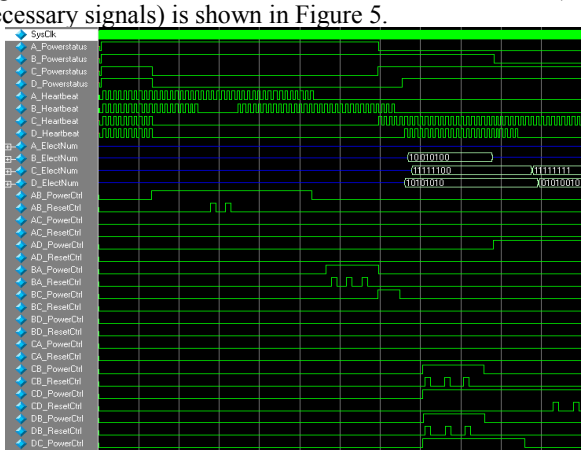


Figure 5. The simulation results

According to the waves shown in Figure 5, we can see that after system powers on, all system nodes turn on and system enters in the centralized fault-tolerant mode. As node A has the highest priority, it becomes the one and only master in the whole system. Because system only needs two working nodes, master A makes node C and D power off. Later, node B goes wrong. Master A resets it twice and turns it back to normal. While master A fails, it switches to slave automatically, and node B becomes the new master according to the fixed priority. Master B makes node A reset 3 times, but still can't fix it. So master B turns off node A and powers on node C at the same time. When node B fails again, master node fails the second time in the centralized mode. System switches to the distributed fault-tolerant mode promptly. Node D powers on and starts to elect system master together with node B and C. After executing SEA, Node C and D become the new masters. They make failed node B reset 3 times and turn it off eventually. Soon, master D goes wrong and switches to slave. System loses its master node and starts to elect master again. As there is only one normal system node left, system reduces the required number of system masters and chooses node C as its master. Master C reset node D, trying to recover it.

Based on the observation above, we can make sure that the design and implement of CSHFt scheme is correct and effective.

VI. ANALYSIS AND EVALUATION

A. System reliability

In a parallel satellite system which has m (m>1) system nodes, assuming the probability of each system node being permanently failed is n, the probability of system breakdown when system adopts CSHFt scheme is equal to the probability of all system nodes being failed, i.e.:

$$P_{System\_failure} = P_{All\_node\_failure} = n^m \quad (1)$$

It's as same as the probability of system breakdown when system adopts ESDFt scheme, which is much less than the failure probability $n+n^{m-1}(1-n)$ of SACFt scheme under the same conditions [8]. As the total number of system nodes increases, the gap between two failure probabilities will become wider (See Table I). Hence, CSHFt scheme has great fault-tolerant capacity, which can guarantee the reliability of parallel satellite system well.

TABLE I. THE PROBABILITIES OF DIFFERENT SCHEMES UNDER DIFFERENT CONDITIONS

m	n	SACFt	ESDFt/CSHFt
2	0.1%	$1.999*10^{-3}$	$1.000*10^{-6}$
	0.2%	$3.996*10^{-3}$	$4.000*10^{-6}$
3	0.1%	$1.001*10^{-3}$	$1.000*10^{-9}$
	0.2%	$2.004*10^{-3}$	$8.000*10^{-9}$
4	0.1%	$1.000*10^{-3}$	$1.000*10^{-12}$
	0.2%	$2.000*10^{-3}$	$1.600*10^{-11}$
5	0.1%	$1.000*10^{-3}$	$1.000*10^{-15}$
	0.2%	$2.000*10^{-3}$	$3.200*10^{-14}$

B. Fault-tolerant capability

CSHFt scheme not only combines SACFt scheme and ESDFt scheme, but also expands and completes them.

(1) By integrating SACFt scheme which is easy to implement and has little performing overhead, CSHFt scheme narrows down the complexity and executing cost of system fault tolerance.

(2) Selecting system master according to the fixed superiority will introduce little delay of switching failed master and new master, which guarantees the real-time feature of system fault tolerance better. CSHFt scheme eliminates the shortcomings of ESDFt scheme and ESA when there is only one master in the whole system.

(3) By assimilating ESDFt scheme, CSHFt scheme ensures persistent system fault tolerance, even under atrocious conditions. It also limits the complexity of system fault tolerance by reducing the number of parallel multi-masters when system performs in the distributed fault-tolerant mode.

(4) SEA intelligently selects system nodes which have lower probabilities of being failed to be system masters. The election overhead can be effectively hidden in the multi-master fault-tolerant manner.

(5) Transmitting control signals, status signals, election information and synchronization information through dedicated data channels avoids the collision and obstruction which may occur during communication. It improves the real-time feature of system fault tolerance in terms of information transmission.

VII. RELATED WORK

For present low degree of integration, parallel satellite systems usually only have two system nodes and adopt centralized fault-tolerant scheme. Reference [1] adopted dual redundant and warm spare fault-tolerant methods to build a satellite system, in which the master node processed while the slave node monitored. If any system fault existed, two nodes could switch their roles. Reference [2,3] designed the X2000 parallel system, adopted centralized fault-tolerant architecture and constructed multilevel fault-tolerant structure consisting of COTS device, dedicated circuit, hybrid multi-bus and redundant multi-module. Reference [4] constructed high-performance system using COTS components, adopted dual gigabit-bus and dual redundant backup, implemented centralized fault-tolerant control with anti-radiation devices and achieved fault-tolerant management in form of middleware at the software layer. Reference [5] combined hot spare and cold spare methods, detected system faults using strong-simultaneous comparison and achieved fault isolation and redundancy switching by independent power control, system reconfiguration and other mechanisms. In order to greatly improve system performance, reference [6] implemented a 4-board high-performance system, adopting centralized approach to achieve hot standby and supporting failure recovery based on dynamic reconfiguration. It achieved fault-tolerant control comprehending low-level drivers and operating system components, collaborating system hardware and software. Reference [7] presented the SACFt scheme, which was insensitive to scale change of the processing nodes. It ensured autonomous fault tolerance, and controlled processing node parallelism on demand flexibly at the same time.

Researches on distributed fault-tolerant scheme remain at the exploratory stage. A number of research achievements have been made in the fields of fault-tolerant theory and system management software. But useful information about the methods of designing or constructing such an actual system rarely can be found. Reference [8] introduced the ESDFt scheme. It used isomorphic nodes to compose a symmetric framework and utilized master election mechanism, which eliminated the risks of SPOF and made fault-tolerant complexity and processing overhead flexibly adjustable.

VIII. CONCLUSION

Nowadays, building parallel system with high-performance COTS chips becomes the main way to greatly improve satellite system performance, which makes fault-tolerant scheme extremely important. Though the centralized fault-tolerant scheme is simple to implement and has little processing overhead, it may lead to system failure easily because of its SPOF. Distributed fault-tolerant scheme hasn't been maturely used on actual systems because of its complexity and large overhead.

In order to combine these two traditional schemes and complement each other, we construct a general fault-tolerant architecture which supports both methods. It consists of isomorphic nodes and forms a symmetric framework. Each system node is independent and equal completely, which makes system fault tolerance more flexible. Based on the architecture, CSHFt strategy realizes self-adaptive hierarchical fault-tolerant process in the sequence of 'first centralized then distributed'. System switches its fault-tolerant mode actively according to its performing history, so as to accommodate itself to the changes of external conditions. By selecting system master according to the fixed superiority, CSHFt scheme improves system real-time feature in the centralized fault-tolerant mode and makes up the shortcomings of ESDFt scheme and ESA while there is only one master exists in the whole system. While in the distributed fault-tolerant mode of CSHFt scheme, system chooses multi-masters intelligently according to ESA, and the election overhead can be hidden effectively. MPA guarantees system fault-tolerant control could be responded correctly and corresponding fault-tolerant operations could be taken in time in both fault-tolerant modes.

As system fault-tolerant authority can be transferred between different system nodes, system failure only occurs when all the system nodes are broken. Compare to the other simplex fault-tolerant schemes, CSHFT scheme has stronger fault-tolerant capacity and suits for any application.

ACKNOWLEDGMENT

The research is supported by the National High Technology Research and Development Program of China (863 Program), item numbers of which are 2008AA01A201 and 2009AA01Z101.

REFERENCES

[1] Xiang Lin, Wu Xianghu, Liao Minghong, "Fault-Tolerant Mechanisms of House-keeping Computer System for The Small Satellite," Journal of Astronautics, vol. 26, Jul. 2005, pp. 8-11.

[2] Savio N. Chau, Leon Alkalai, Ann T. Tai, "Analysis of a Multi-Layer Fault-Tolerant COTS Architecture for Deep Space Missions," International ACM SIGACCESS Conference on Computers and Accessibility (ASSET'00), ACM, Oct. 2000, pp. 70-77.

[3] Savio N. Chau, Leon Alkalai, Ann T. Tai, "The Design of a Fault-Tolerant COTS-Based Bus Architecture for Space Applications," https://eprints.kfupm.edu.sa, Mar. 2009.

[4] J. Ramos, J. Samson, M. Patel, A, George, and R. Some, "High Performance, Dependable Multiprocessor," The 2006 IEEE Aerospace Conference. Big Sky, MT, Mar. 2006.

[5] Wang XinSheng, Sun Hanxu, Xu Guodong, Tong Zhihong, "Study on the on-Board Computer System Based on ARM Processor," Journal of Beijing University of Posts and Telecommunications, vol. 28, Aug. 2005, pp. 21-25.

[6] Jiang Jingfei, Tang Yuhua, Ning Hong, "Reconfigurable Parallel Embedded System," Computer Engineering, vol. 35, Jun. 2009, pp. 405-409.

[7] Hao Zhou, Yuhua Tang, Jingfei Jiang, "Design and Implementation of Scalable Autonomous Centralized Fault-Tolerant Scheme for Satellite System," The 3rd International Conference on Computer Design and Applications(ICCDA2011), May 27-29, 2011, In press.

[8] Hao Zhou, Jingfei Jiang, "Research on Equal Symmetric Distributed Fault-tolerant Architecture and Strategy for Parallel Satellite System," The 14th IEEE International Conferences on Computational Science and Engineering(CSE2011), Aug 24-26, 2011, In press.

Research and application of PV monitoring system based on ZigBee and GPRS

Hu Yujie
School of Internet of Things, Jiangnan University
Wuxi Jiangsu, China
e-mail:yujiehu1983@yahoo.com.cn

Zhang Xihuang
School of Internet of Things, Jiangnan University
Wuxi Jiangsu, China
e-mail:zxh163com@163.com

Abstract-Aimed at the current status of PV solar monitoring system as well as the PV solar monitoring network requirements, and application of Wireless Sensor Networks (WSN) to PV solar monitoring system is proposed. The Wireless Sensor Networks structure of PV solar monitoring system based on ZigBee and GPRS and topology of ZigBee Wireless Sensor Networks for PV solar monitoring were presented. Experiments show that the system meets the demand for PV monitoring system in terms of data acquisition and transmission, and has some advantages at installation and networking, it's the ideal solution of PV monitoring system.

Key words- wireless sensor networks; ZigBee; GPRS; PV monitor; Data acquisition and transmission.

Ⅰ.INTRODUCTION

Wireless sensor network, which integrates sensor, MEMS technology, wireless communication technology, embedded computing technology and distributed information management technology, has been under rapid development during recent years [4]. ZigBee is an open and global standard for WSN aiming at a low rate, low cost, low power consumption and self-forming wireless communication. This paper presents the current researches of ZigBee and GPRS wireless sensor network on PV monitoring system [4].

Solar photovoltaic power generation is divided into off-grid and grid operation, at present a broad-based installation and building installation primarily. A broad-based installation distributed in the open desert with light enough, Building installation distributed in the building roofs and facades with a large area, Installation position is remote, monitoring is difficult, etc. So both for off-grid PV system and grid PV system monitoring, operation monitoring is particularly important.

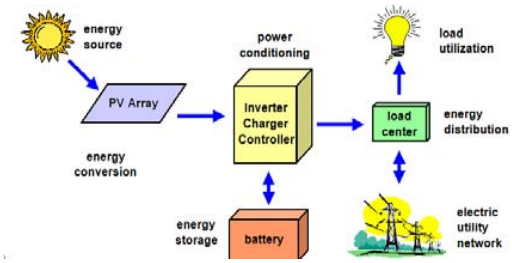


Fig.1 Principle diagram of PV system

PV system has solar cell module、controller or grid inverter etc, and it also includes data acquisition system、data exchange、operation shows and monitoring equipment、etc [7]. Communication is the key link ensures that monitoring system transmits messages and is efficient. And communication network system largely determines the topology of the monitoring system, therefore, it is very necessary that we thoroughly research and test the scheme of communication in monitoring system design. At present the monitoring technology of PV power station is still in development. In general speaking, single means of communication、the lack of network management、no high technical level 、low level of integration, difficult to carry the growing business needs, this topic research is how to use wireless network information technology to real-time monitor the status of PV system .

Ⅱ. ARCHITECTURE DESCRIPTION OF PV

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.74

MONITORING SYSTEM

Photovoltaic system monitoring software carries on the comprehensive monitoring for the condition of equipment operation and environment instrumentation. System will transmit real-time operating status information to the monitoring data management module through the data acquisition and transmission module. Monitoring data management undertake unity centralized management for real-time information、equipment management information and monitoring personnel information; Monitoring data applications support to provide real-time monitoring service、fault alarm service、data analysis service、information release service and so on, and combining monitoring applications to develop tools and statements/chart tools to implement real-time monitoring、intelligent alarm、data analysis、equipment management、system management、information release、etc.

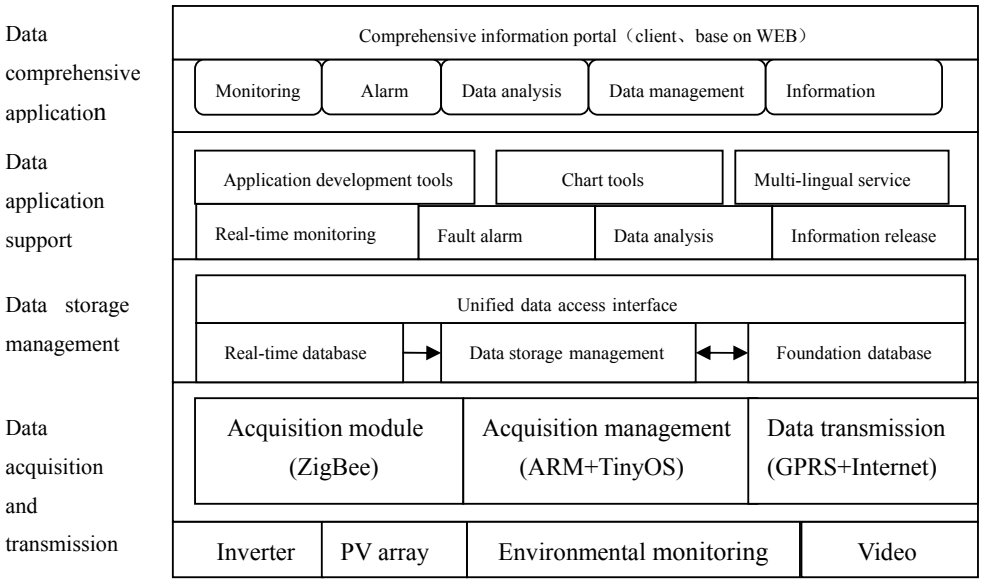


Fig.2 Logical structural diagram of PV monitoring system

Data monitoring includes: inverter and environment instrumentation such as the current state of the operation parameters、all derivatives parameters、equipment warning data、etc [7].

ZigBee wireless sensor network distributes across the monitoring area of PV subsystems, constituting a monitoring network by star or network topology structure, and each acquisition module all packed into the main module; The basic function of GPRS network sends packet data from ZigBee main module to internet network router, realizing the data long-distance transmission; Monitoring center system mainly process and storage PV monitoring data and analyze them.

ⅢI.KEY TECHNOLOGY AND REALIZATION OF DATA ACQUISITION AND TRANSMISSION

A. Wireless sensor node hardware design based on ZigBee

Wireless sensor networks have multiple sensor node structure [2], this system network node core board adopts JN5148 chip, Wireless sensor network is designed by star topology structure [6], among them main nodes use physical equipment for the FFD, which is said main equipment and has the role of convergence node [3]. It assumes the function of the network coordinator can communicate with network equipment of any kind. RFD can communicate with the main node. In the system quantities of sensors are few, the short distance communication, formation algorithm of a star network topology is relatively simple, largely

meet the demand of PV monitoring system, the internal circuit of RFD is simple than the FFD，so it favors the saving energy consumption.

At present, each sensor network node can provide four analog channels, can also connected four analog sensor, can also design more channels.4 simulation passageway of each sensor node can connect different types of sensor according to actual application. Gathering data are timing sent to the main equipment (gathering node) by ZigBee, and are sent to remote data management center through GPRS and internet. Main equipment and sensor nodes wireless communicate through protocol ZigBee, greatly improving mobility and flexibility of the system. Sensor connection mode as shown in figure:

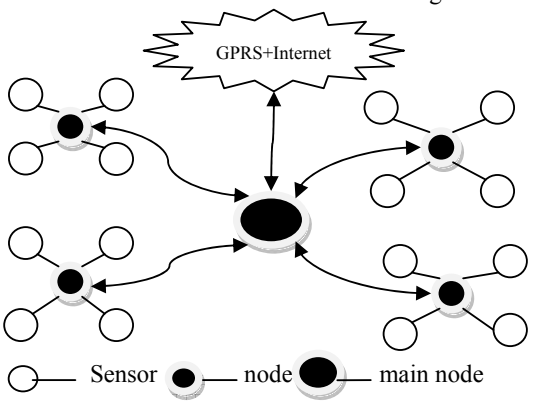


Fig.3 Structure of terminal data acquisition and transmission module

B. Design of embedded processor module

Hardware structure of embedded processor module is divided into two parts: the ARM processor and it's expand interface, and GPRS Modem.ARM controller and interface circuit is the core of this module ,and it is basic core circuit which guarantees this part normally work. Microprocessor controller is the core of the network coordinator, and coordinator is the control center of data acquisition and transmission. It influences the realization of the function of the whole monitoring system, comprehensive consideration of network management、data transceiver、memory、data throughput rate and processing capacity, choose 32bits embedded series microprocessor of TI company ARM7 kernel as microprocessor controller, using power supply. Wireless RF communication module adopts CC2430 chip [3]. GPRS communication module is mainly used for the wireless sensor network information remote transmission, and it's the remote communication interface of ZigBee network. GPRS communication module uses MC39i module produced by Siemens Company. This module integrates baseband processing unit and RF processing units, providing the 40pin ZIF connector. Through RS232 interface connector of the ZIF socket it sends commands and data to MC39i,can achieve module setting、system control、data input/output、and other functions.

This operation system chooses TinyOS [1] real-time operating system. TinyOS adapts to wireless sensor network operating system, adopting the modular programming language nesC、introducing lightweight thread、active message、event-driven mode、component-based programming、etc [1], which good uses of the limited resources of sensor nodes. TinyOS is event-driven operating system, the CPU will wake-up process when there is event-triggered, it can be sleeping in the rest time, and thus it can dramatically reduce the energy consumption of the system. This module's principle diagram as show in figure:

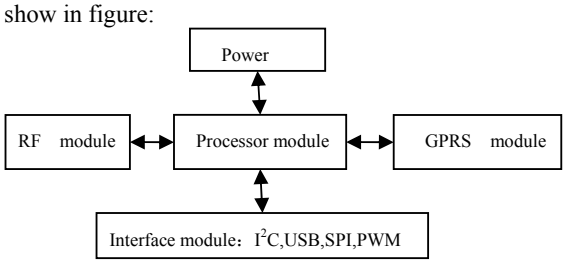


Fig.4 Diagram of controller module

C. Software design of sensor network management

Sensor network management software is divided into three parts: data collection and transmission module of RFD, data sending and receiving module of main node, control management module of ARM+TinyOS embedded system. Considering the development cycle and readability of program, communication between the RFD and the FFD with a dedicated suite of development tools jennic. ZigBee network platform can be easily constructed, effectively reduce the difficulty and cost of development, and also increase the stability of the system. System code developed by C language. After main equipment completes internet connection and joins in management, receiving the data from each node, and the data is sent to ARM+TinyOS embedded module through serial port, processed data is sent to remote control center. A two- step completion of data transmission through GPRS: dial-up internet and network communication based on socket, the essence of dial-up is to realize the PPP link(point to point protocol over Ethernet),It indicates that it can communication with the server via a socket after PPP connection is established. The PPP is expansion of TCP/IP, it can transmission TCP/IP packets through serial interface. Using a stream socket for network communication between different hosts is a typical client/server model (Server/Client), that the client sends a service request to the server, the server corresponds services according to the request. Program flow chart as shown in figure 5:

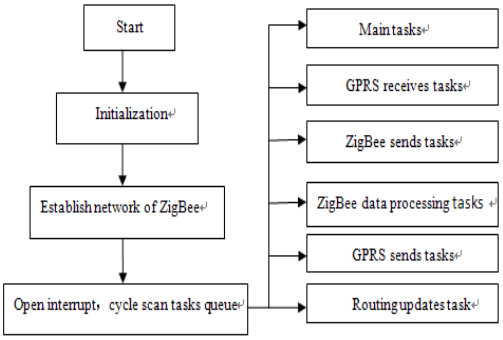


Fig.5 Flowchart diagram of host device control

Ⅳ. APPLICATION RESULTS AND ANALYSIS OF SYSTEM EXPERIMENT

In the field experiment of the system, setting up PV wireless network monitoring system, and test and apply it in the field of PV system, eventually achieve the design goal of the system. Comparing with PV wired monitoring system, wireless PV monitoring system have advantages with a simplified installation wiring、cost saving、the overall profile appearance、convenience、etc. This paper experiment with the help of installed PV system, the capacity of PV system is about 50KW,three sub-systems are located in the top floor of different building, Installation Angle is 22°, the geographical position is 32 degrees north latitude and 120 degrees east longitude. Nodes of ZigBee wireless sensor network distribute in the area of 1000m*1000m, and to realize the monitoring of PV system. Sensors acquire environment temperature、PV module temperature、solar radiation、daily electricity、etc [7], then upload acquisition data to internet by wireless sensor network, achieving to wireless remote data monitoring and analysis. Test results as shows in table 1(known by statistics).

Tab1Measuring parameter result of experiment condition

Monitoring parameters(March 2011)	Test data
Average environment temperature	8.3℃
Average module temperature	16℃
Wind speed	4m/s
Average irradiance	3.58 kwh/m2/day
Daily generating capacity	146kwh
Total capacity of March	4526kwh

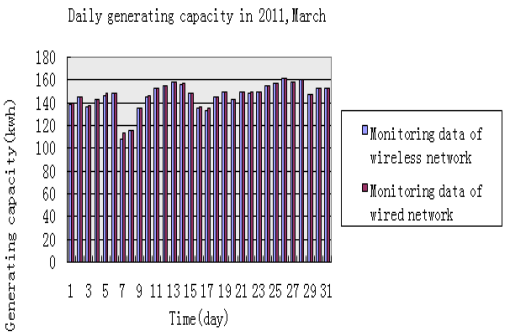


Fig.6 Data graph of acquisition and monitoring

Data in table 1 remote call through a database and data statistics by online chart, comparing monitored data with wired monitoring system, as suggested by the chart 6 is informed, they stay the same, which verifies the reliability and accuracy of the system.

Ⅴ. CONCLUSION

In this paper, the various sub-system of PV power system distributed in different regions, it's also difficult to layout cable between sub-systems, to achieve the uniform monitoring, namely that we proposed a remote monitoring method based on wireless sensor network. This paper introduces basic principle、system development process and analysis of experimental of remote monitoring system based on ZigBee and GPRS protocol. This paper uses wireless sensor network advantages with flexible networking、convenient、etc, to complete data acquisition and transmission function of PV system. In the experiment we found that the node equipment required low power consumption, transfer between nodes by data relay method, communication efficiency is very high; ZigBee also has the characteristics of self-organization and dynamic routing, which can be conveniently for network expansion and greatly improve communication reliability. ZigBee appropriates applying in all kinds of types of user-grid PV system, can effectively reduce the cost, improve reliability of the system and reduce working intensity of maintenance.

REFERENCES

[1] Sun Zhiyuan, Han Wei. Application of ZigBee technology in vibration [J]. Transducer and Microsystem Technologies,2009,28（10）. 114-116.

[2] Sheng Chaohua, Chen Zhaolong, Wireless sensor network and application [J], Microcomputer application, 2005, 21 (6). 10-13.

[3] Dai Yuan, He Dongjian, Zhang Jianfeng. Research and implementation of wireless transmission network of agricultural information based on ZigBee [J]. Transducer and Microsystem Technologies,2010, 29（7）. 14-16.

[4] Zhou Yiming, Yang Xianglong, Guo Xishan, Zhou Mingang, Wang Liren, A design of greenhouse monitoring &control system based on ZigBee wireless sensor network [C]. Wireless Communication, Networking and Mobile Computing, 2007. WiCom2007. International Conference. 2007. 2563-2567.

[5] Bhambri L P,Jindal C, Bathla S. Future wireless technology-ZigBee [C]. Proceedings of National Conference on Challenges &Opportunities in Information Technology, 2007. 154-156.

[6] Lee Jin-Shyan, Huang Yang-Chih ITRI ZBnode: A ZigBee/IEEE802.15.4 Platform for Wireless Sensor Networks [C]. Proc. of 2006 IEEE International Conference on System, Man and Cybernetics. Taipei, Taiwan, China: IEEE Press,2006.

[7] GB20513-2006T Performance monitoring、measurement、data exchange and analysis guideline of photovoltaic system.

Application of a modified PSO algorithm in PID controller optimization

Jiang Shi-cheng
IoT Engineering
Jiangnan University
Wuxi, Jiangsu, China
jiang.969606@163.com

Xu Wen-bo
IoT Engineering
Jiangnan University
Wuxi, Jiangsu, China
xwb@sytu.edu.cn

Abstract: A modified PSO for Optimum Design of PID Controller is stated. Particle Swarm Optimization is a new random global optimization algorithm. The algorithm feature is simple, ease to implement and powerful function. But it is easy to trap in the local pole, and search accuracy is low. PID controller principle is concise, physical meaning of its parameters are clear, and theoretical analysis system is integrity and the industrial sector are familiar with it, so in the industrial process control is still widely used. In this paper, the modified PSO algorithm has been used in PID controller to optimize parameters and compared with basic PSO algorithm. Emulation experiments demonstrated that the modified PSO algorithm increase the probability convergence to the global optimal solution.

Key words: PID control, Particle Swarm Optimization, parameter optimization

I. Introduction

Particle Swarm Optimization is an evolutionary computation method proposed by American social psychologist, James Kennedy and electrical engineer Russell Eberhart in 1995. Because of PSO algorithm concepts simple and easy to implement, in a few years, PSO algorithm has achieved great development[1-2]. Nowadays PSO algorithm has already used in function optimization, fuzzy system optimization and application field widely research. some researchers is applied to the optimization design of the PID controller, some researchers it in PID controller optimization[3-7].

PID controller principle is concise, physical meaning of its parameters are clear, and theoretical analysis system is integrity and the industrial sector are familiar with it, so in the industrial process control is still widely used (and over 90%).The control effect of PID control has an extremely important relationship with three parameters Kp、Ki、Kd. That is to say, the selection of Kp、Ki、Kd is directly related to the performance indicators and stability of the control system, so setting and optimization the PID parameters has always been interested by researchers. PSO algorithm is easy trapped in the local pole, and search accuracy is low, so in this paper, the modified PSO algorithm has been used in PID controller to optimize parameters. Emulation experiments demonstrated that the modified PSO algorithm is feasible and effective.

II. The basic principle of PID control

The PID controller transfer function is

$$u(t) = K_P[e(t) + \frac{1}{T_I}\int_0^t e(t)dt + T_D\frac{de(t)}{dt}] \quad (1)$$

Where

K_P	the proportional gain
T_I	integral time coefficient
T_D	derivative time coefficient
$e(t)$	the tracking error
$K_I = \frac{K_P}{T_I}$	the integral gain
$K_D = K_P T_D$	the derivative gain

A proportional controller K_P will have the effect of reducing the rise time and will reduce, but never eliminate, the steady-state error. An integral control K_I will have the effect of eliminating the steady-state error, but it may make the transient response worse. A derivative control K_D will

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.25

have the effect of increasing the stability of the system, reducing the overshoot, and improving the transient response. Adjusting the PID controller parameters reasonably is able to obtain satisfactory system performance. The so-called PID controller parameters optimization, is to seek a set of optimal PID control parameters. It allows the system to meet the fast response, small overshoot, short settling time the dynamic performance requirements.

III. PSO Algorithm

Population size is n. In M-dimensional search space, the position of the *i*th particle can be represented by a M-dimensional vector, $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})$.The velocity of the particle can be represented by another M-dimensional vector, $V_i = (V_{i1}, V_{i2}, \cdots, V_{im})$. The best position previously visited by the *i*th particle is denoted as $P_i = (p_{i1}, p_{i2}, \cdots, p_{im})$, also called $p_{best} \cdot g_{best}$ that is tracked by the global region of the particle swarm is the overall best value, and its location is obtained so far any particle in the group. The new position of the particle and its velocity will be determined according to the following two equations:

$$v_{ij} = w * v_{ij} + c1 * r * (p_{ij} - x_{ij}) + c2 * R * (p_{gj} - x_{ij}) \quad (2)$$

$$x_{ij} = x_{ij} + v_{ij} \quad (3)$$

Where inertia weight factor *w* is set according to (4). c1 is the acceleration weight coefficient of particle itself. c2 is global acceleration weight coefficient. r and R are two random functions in the range [0,1].

$$w = \frac{(w_{\max} - w_{\min}) * (\max iter\text{-t})}{w_{\max}} + w_{\min} \quad (4)$$

Where $w_{\max}$ is maximum value of inertia weight, $w_{\min}$ is the minimum value of inertia weight, $\max iter$ is the maximum of iteration in evolution process, t is current value of iteration. The speed of particles v_i is limited by a maximum speed $v_{\max}$. If the current acceleration of particles in a dimension causes it faster than the maximum speed of the dimension, the dimension of speed is limited to the dimensional maximum speed.

Particle swarm optimization can better complete the task, the key is on the amendment of the flight speed and position of each generation. Analysis of equation (1), it contains three parts, $w*v_{ij}$ made the particle maintain a original flight inertia. $c1*r*(p_{ij}-x_{ij})$ made the particles have a tendency toward its own best position. $c2*R*(p_{gj}-x_{ij})$ made the particles have a tendency toward current best position.

IV. PSO-PID control system design

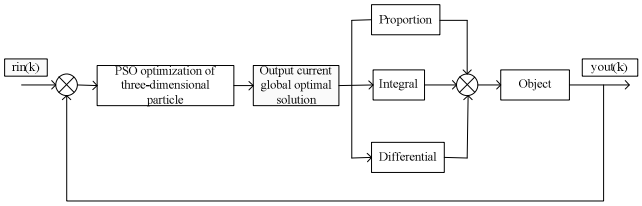


Fig. 1 The PID control system based on PSO

The idea of PSO algorithm optimize the PID controller parameters is: First determine the dimension of the algorithm(The optimization dimension for PID controller is 3 dimension. v_{1i} and x_{1i} represent the speed and position of k_p ,is the first dimension. v_{2i} and x_{2i} represent the speed and position of k_i ,is the second dimension. v_{3i} and x_{3i} represent the speed and position of k_d ,is the third dimension), Evolution algebra. Then, according to the requirements of the control performance, determine the optimal evaluation indicators, and code PID parameters as the following particles coding: $[k_p, k_i, k_d]_T$. The range of each variable in particles is estimated depending on the specific engineering background[8]. Seek the optimal solution through the PSO in the range of three parameters proportional, integral, differential. The specific algorithm process is as follows[9-10]:

Step 1: Initialize a population of particles with random positions and velocities on M-dimensions in the problem space, maximum number of iterations, parameters c1,c2, weight;

Step 2: For each particle, according to the selected fitness evaluation function, evaluate the fitness value f_x(i);

Step 3: For each particle, compare particle's current fitness evaluation with its f_pbest(i), if f_x(i) < f_pbest(i), the set f_pbest(i) equal to the current value, and pbest equals to the current position x(i). Find the smallest of all

the populations f_pbest, if f_pbest < g_best, then the set g_best(i) equal to the current value, and gbest equals to pbest;

Step 4: Update the velocity and position of the particle according to equation (2) and (3);

Step 5: Loop to step 2 until a criterion is met or the maximum number of iterations is reached. After optimization, output gbest, they are the results of the optimization of three parameters k_p、k_i and k_d.

V. Improve PSO algorithm

From the point of dynamic, in the convergence process, p is an attractor, the particle in PSO algorithm close to p as the speed of particle decreases continuously, and finally dropped to p. Therefore, in the process, actually p point exist some form of potential field which attracts the particles. That is why the entire population of particles maintains the clustering. However, in the classical PSO system, particles convergence is achieved in the form of orbit, and the particle velocity is always limited, the search space of the particles is a limited area, and can't cover the whole feasible space in the search process. General PSO algorithm can't guarantee convergence to the global optimal solution with probability 1, this is the biggest drawback of the general PSO algorithm.

To prevent this undesirable trend, put forward the modified PSO algorithm based on the general PSO algorithm. According to the location coordinates of the four particles generated randomly, construct two sets of random vectors. Replace $(p_{ij} - x_{ij})$ and $(p_{gj} - x_{ij})$ in equation (2) with the two random vectors with a certain probability *CR*. Then the random vectors are calculated by the following equation:

$$\delta_n = mu\_x_k - mu\_x_s \qquad (4)$$

Where mu_x_k and mu_x_s are two random particles generated in the problem space. One can see that as the random vector is introduced, the particles may escape from the current position and locate in a new search area.

VI. Simulation and Analysis

To comparing PSO with ZN and verify the efficiency of PSO, a third-order system are given as below:

$$G(s) = \frac{1}{s^3 + 6s^2 + 5s}$$

Population size=50. Acceleration constants are c1=c2=2, The number of iterations is G =50. According to third-order system, the initial range is $K_P \in [0,30]$, $K_I \in [0,1]$, $K_D \in [0,25]$.The evaluation function is:

$$J = 0.9 * ts + 0.1 * \delta \qquad (6)$$

Where ts is the settling time, δ is the overshoot.

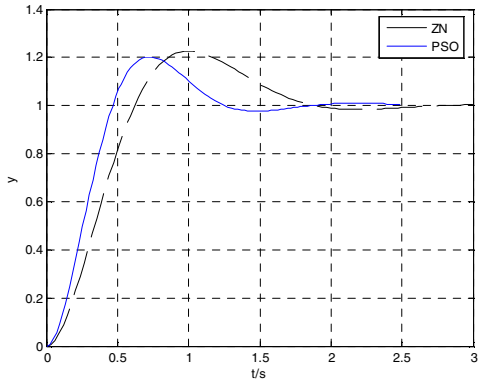


Fig. 2 Comparison of unit step responses

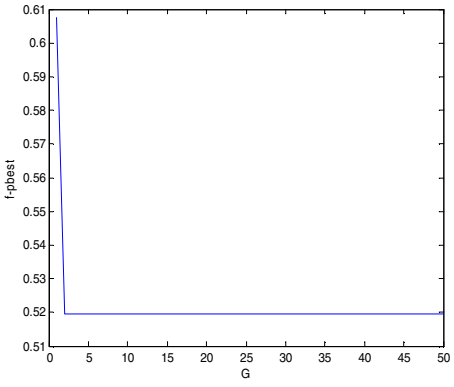


Fig. 3 The trend line of the value of fitness function

Table 1

SUMMARY OF TOW SIMULATION ON RESULTS

	k_p	k_i	k_d	J	ts	δ
ZN	25	0.1	15.0	-	1.5887	0.2219
PSO	17.204	0.34948	14.194	0.51949	1.0901	0.19355

Comparing with the ZN method, PSO algorithm obviously improved the system control performance, lowered the overshoot, reduce the settling time.

Use the basic PSO algorithm to optimize PID parameters, run 100 times, of which only 23 times is global convergence, the average fitness value f = 0.54210. Use the modified PSO algorithm to optimize PID parameters, run 100 times also, of which 42 times is global convergence, the average fitness value f = 0.50542. The modified PSO algorithm increased the probability of convergence obviously.

VII. Conclusions

In this paper，the basic PSO algorithm is applied to optimize the design of PID controller, and the modified PSO algorithm is used to improve the global convergence of the algorithm. The simulation results show that, PSO algorithm is not only simple to program, to avoid the complex process of encoding and decoding, but also will help to improve system quality, fast searching speed with the degree of fast, accurate, etc. comparing with the basic PSO algorithm, the modified PSO algorithm overcome the "premature" phenomenon and greatly reduce the possibility of premature convergence is has obvious advantages In global convergence.

[1] KENNEDY J, EBERHART R C . Particle swam optimization [C] // Proceedings of IEEE International Conference on Neural Networks, Piscataway , NJ: IEEE Press ,1995:1942-1948.

[2] R.Eberhart,Y.Shi. Particle Swarm optimization: development, application and resoutce [J].IEEE Int conf on evolutionary Computation, 2001 pp81-86.

[3] Ioan Cristian T relea. Particle Swarm Optimization Algorithm: Convergence Analysis and Parameter Selection [J]. Information Processing Letters, 2003, 85: 3172-3325.

[4] Zhang Kui, Qv Li-guo, Huang You-rui. Application of a modified PSO algorithm in PID parameters optimization [J]. ELECTRONICS OPTICS & CONTROL, 2008(2): 58-61.

[5] Sun Jun, Xu Wen-bo, Fang Wei. Solving multi-period financial planning problem via quantum-behaved particle swarm algorithm [C] // ICIC, 2006: 1158-1169.

[6] Xu Jing-bo. PID Tuning by Means of Particle Swarm Optimization [J]. JOURNAL OF DON GHUA UNIVERSITY(NATURAL SCIENCE), 2007(2): 135-137.

[7] Zeng Xiang-guang, Zhang Ling-ling. A Particle Swarm Optimization Approach for Optimum Design of PID Controller [J]. Machinery Design & Manufacture, 2007(4):81-82.

[8] Clerc M, Kennedy J. The Particle Swarm: Explosion, Stability, and Convergence in Multi-Dimension Complex Space [J]. IEEE Transactions on Evolutionary Computation, 2002, 16(1): 58-73.

[9] Zhao J, Li T P. Application of particle swarm optimization algorithm on robust PID controller tuning[C]//LNCS 3612, 2005: 948-957.

[10] Kazuaki I, Makoto I. Optimal design of robust vibration suppression controller using genetic algorithms[J]. IEEE Trans on Industrial Electronics, 2004, 51(5): 947-953.

Microprocessor Critical Design and Optimization

Jiangyi Shi ,Honghu Gong, Hongye Jia, Kang Li
School of Microelectronics, XiDian University
Xi'an, China
ghonghu@126.com

***Abstract*—In this paper a 32-bit multithreaded RISC microprocessor is designed and optimized to perform data moving and processing in the high-performance Network Processor which is flexible to a wide variety of networking, communications, and other data-intensive products. As a critical part of network processor, the microprocessor mainly takes in charge of Internet Protocol (IP) packets' transmitting. Forwarding logic is used to deal with the data hazard, and defer slots are used to deal with the control hazard, besides logic and physical optimizations are employed to solve timing of the critical path.**

Keywords- microprocessor; hazard; optimization; high-performance

I. INTRODUCTION

The microprocessor implements the RSIC instruction set. In order to implement complicated network packets processing and communications that require bit, byte, word and longword operations to forward data quickly and efficiently, some other instructions are employed. The microprocessor has four independent program counters, zero overhead context switching and hardware semaphores from other hardware units to ensure that each microprocessor can be fully utilized. A microprocessor's powerful ALU and shifter perform both ALU and shift operations in a single cycle. Each microprocessor contains a large amount of local memory and registers: 4 Kbytes organized as 1024 by 32 bits of high-speed RAM Control Store for program execution, 128 32-bit General Purpose Registers, and 128 32-bit transfer registers to service the SRAM and SDRAM Units. Fig.1 is a block diagram of microprocessor architecture with five-stage pipline which works as a custom processor for network packets processing.

- IF: instruction fetch stage. Instruction is fetched from IS_UNIT, and PC points address of next Instruction. There are four counters used in countering the pc for each thread, and a context event arbiter logic has arbitration of a round robin mechanism for the four threads. After a instruction is fetched, the local control state register (local CSR) that includes the next context state and active context state is updated.
- ID: instruction decode stage. Instruction is decoded to be control signal for selecting operands , type and address of operands to the next stage.
- RF: read register file. Reading operands from the corresponding address of General Purpose Registers, or transfer registers written by the external SRAM and SDRAM Units to the execution stage.
- EX: excute throughadd, sub and shift operand. The result is sent to the next stage and a branch condition code is generated in this stage which is used to excute Branch jump.
- WB: write the results back to the destination register.

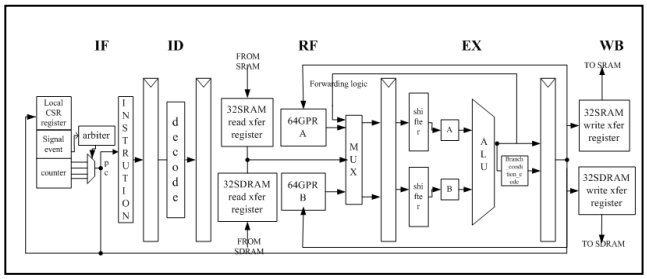


Figure 1. microprocessor architecture

II. ISTRUCTION SET

The microprocessor implements the RSIC instruction set. In order to implement complicated network packets processing, external reference instruction and ALU_SHF instruction，which is typically used in data accessing and calculating,are added.

- Add ALU_SHF instruction: Perform shift operation on one or two operands and then perform ALU operation all in one cycle.
- Get rid of LOAD\STORE instruction: As microprocessor has to deal with a large amount of data packets, many external memorys are inserted. At the same time, on-chip memory can't fullfill the requirement, so the LOAD\STORE instruction is deleted.
- Add external reference instruction：use a FIFO for sending 60-bit reference instruction，and accessing external SRAM and SDRAM.

III. TECHNICAL CHALLENGES

Pipeline hazards include structure, data and control hazards which reduce the efficiency of the pipeline. Because

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.60

of the simplicity of the microprocessor's pipeline, only data and control hazards exist. The WAW (Write- After-Write) data hazard causes the out-of order execution of instructions**[5]**. In our work, we mainly focus on RAW (Read-After-Write) data hazards, because we assume in-order execution of instructions and we use a write-before-read register file. RAW hazards arise from general purpose register hazard. And this problem can be solved with a simple hardware technique called forwarding (also called bypassing and sometimes short-circuiting)[3].

A. RAW data hazard

The RAW data hazard can be resolved by directly passing the value of ALU_out calculated in the EX stage, back to the MEM stage, using forwarding logic. The forwarding logic compare the instruction after it has been processed in the current stage with an instruction passed back from the ALU_out in EX stage. If there is a conflict (dependency), then the forwarding logic uses the value of ALU_out that is being passed back from a subsequent instruction, otherwise it uses the current value (the value in that stage) of that source register. it is shown in Figure 3, ALU_out is forwarded directly to RF stage, if the GPR_ read address in the current stage equal to ALU_out write GPR address generated in EX stage , then the MUX logic chooses the ALU_out instead of GPR_READ_out.

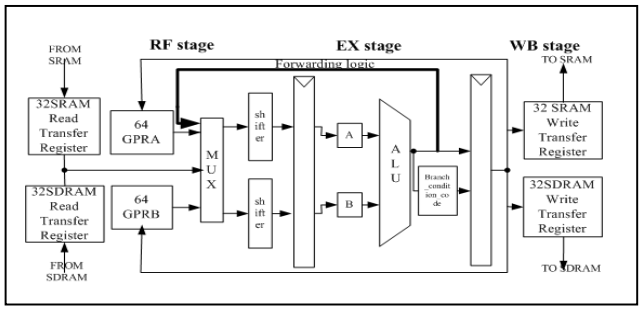


Figure 2. data forwarding

B. Control hazards

Control hazards can cause a greater performance loss than do data hazards. And control hazards arise from the delay of fetching instructions and deciding next PC operation. When a branch is executed, it may or may not change the PC to something other than its current value plus 4. Recall that if a branch changes the PC to its target address, it is a taken branch; if it falls through, it is not taken, or untaken[3].

The defer slots are used to deal with control hazards by filling branch delay slots. Branch is classified into type1, type2 and type3. Type1 Branch executes in ID stage with defer1, type2 Branch executes in RF stage with defer2, and type3 Branch executes in EX stage with defer3.Meanwhile, defer 1 means executing the instruction following this instruction before performing the branch operation. Defer2 means executing the two instructions following the current instruction before performing the branch operation. And defer3 means executing the three instructions following the current instruction before performing the branch operation.

But if there are more than one branch in different stages of pipeline, For example, when a type3 branch with defer 3 is in EX stage at the same time, a type1 branch with defer 1 is in ID stage, it is hard to deal with, so we give the type3 branch high priority, type 2 branch medium priority and type 1 branch low priority.

A branch with defer2 is shown below, where br is a branch to b1. I2 and I3 are both executed before the branch occurs.

TABLE I. BRANCH WITH DEFER2

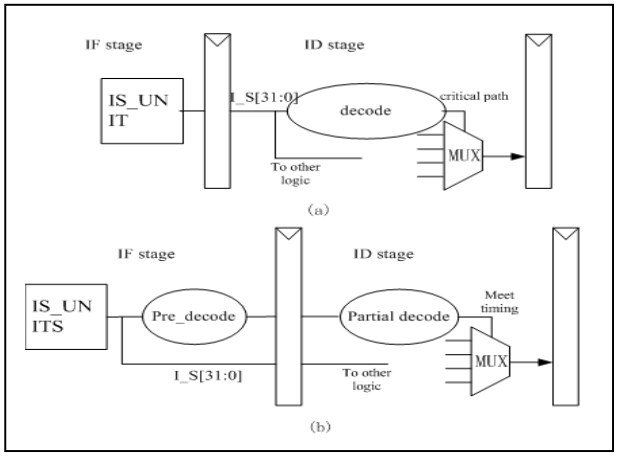
	1	2	3	4	5	6	7	8	9
I S	I1	br	I2	I3	b1	b2	b3	b4	b5
I D		I1	br	I2	I3	b1	b2	b3	b4
RF			I1	br	I2	I3	b1	b2	b3
EX				I1	br	I2	I3	b1	b2
WB					I1	br	I2	I3	b1

IV. OPTIMIZATION

We employed significant optimization to meet the frequency goal of microprocessor. Accustomed processor designs have increased frequency through excessive pipelining which reduces instructions per cycle (IPC) and exponentially increases area and power for little real performance gain[1]. In contrast, microprocessor maintains its pipeline depth, while enhancing the frequency by 50% with little area increasing.

A. Move logic across pipe stages

It helps in solving timing of critical path to move logic across pipe stages. Here we have a fetch instruction stage where instructions are accessed from instruction units (IS_UNIT) and a decode stage where instruction gets decoded to select input of MUX unit. As shown in Figure 3. As decoding and selecting input of MUX unit in a critical path, we have moved portion of the decoding into fetch stage thereby solving the critical path[2].It is called pre-decode which can be used when there is a critical path in the decode



stage with enough slack of timing in the fetch instruction stage.

Figure 3. Move logic across pipe stages : (a) original design. (b) optimal design.

Similarly, we move the shifter from EX stage to RF stage for achieving timing closure, as shown in Figure 2. As a result of our optimization in contrast with the original pipline, it is shown in table2.

TABLE II. SYNTHESIS RESULT OF MICROPROCESSOR PIPELINE

Module	Delay(pre_move logic) (ns)	Delay(post_move logic) (ns)
IS	2.08 ns	2.96 ns
ID	3.95 ns	3.12 ns
RF	1.79 ns	3.01ns
EX	4.62 ns	3.22 ns
WB	0.89 ns	0.89 ns
Total pipline	4.92ns	3.8ns

Note: SMIC 0.13 μm, 1.08 V, 125 °C

B. Logic and physical optimization

The physical synthesis approach opens up a wide range of optimizations where changing logic, electrical and physical characteristics of the design lead to significant benefits[6]. Logic and physical optimizations include net buffering, gate sizing, gate relocation and gate replication, rewiring and restructuring[4]. But Indiscriminate buffering may also create many gate overlaps, leading to potentially detrimental effects on circuit timing when overlaps are resolved [7], and uncontrollable logic replication may increase area and route length, making critical nets longer than expected during placement and routing [8]. A number of related works try to solve this problem. For example, Li et al. [7] proposed an incremental placement algorithm which maintains the stability of a placement for gate sizing and buffer insertion, while Luo et al. [9]and Brenner et al. [10] addressed this problem by designing legalizers that seek to preserve performance metrics.

Techniques based on a placement or routing with the goal to improve timing are often called physical synthesis. To achieve timing closure for high-performance circuits, it is now common to use physical synthesis — an approach that combines logic and physical optimization, potentially performing placement-aware buffer insertion, gate sizing, fanout optimization, etc.

As gate sizing, gate relocation and gate replication make sense in critical path with slight negative slack, so logic and physical optimization is used after our optimizations above, and most of the design meets goal of frequency, with small negative slack in some paths.

V. SUMMARY

This paper focuses on pipeline design and performance optimization for a 32-bit microprocessor with configurable extension instructions. The target operating frequency of microprocessor is 232 MHz in a 0.13-micron technology. The RISC architecture makes it ideal for computational intensive tasks and is especially preferable for networking applications of packet processing. The microprocessor being fully programmable processors, are able to examine packet contents at all levels of the networking stack. This makes it suitable not only for layer 2 and 3 switching/forwarding, but also for applications that require deeper inspection and manipulation of packet contents.

After our optimization work to achieve the speed specification and reduce the average clocks per instruction (CPI), we get an efficient pipeline with the critical path delay of 3.3ns in the worst case and an average CPI of 1.21. Microprocessor was integrated into a network processor SOC chip which has been taped out in SMIC 0.13-micron technology. Preliminary testing result showed the microprocessor works well as we expected.

ACKNOWLEDGMENT

Support for this work was provided by Natural Science Basic Research Plan in Shaanxi Province of China (2010JM8015).

REFERENCES

[1] Joshua Friedrich et al, “Design of the Power6TM Microprocessor,” IEEE International Solid-State Circuits Conference, 2007, pp. 96-97.

[2] C M.R.Thimmannagari, “CPU Design - Answers to Frequently Asked Questions,” Springer Publications, in press.

[3] David A. Patterson, John L. Hennessy, “Computer Architecture: A Quantitative Approach,” Fourth Edition, Ap professsional, in press.

[4] J. D. Warnock, J. M. Keaty,J. Petrovick, J. G. Clabes, C. J. Kircher, B. L. Krauter, P. J. Restle, B. A. Zoric and C. J. Anderson, “The circuit and physical design of the POWER4 microprocessor,” IBM J. RES. & DEV. VOL. 46 NO. 1 JANUARY 2002, pp, 27-51.

[5] XIAO Zhi-bin, LIU Peng, YAO Ying-biao, YAO Qing-dong, “Optimizing pipeline for a RISC processorwith multimedia extension ISA,” Zhejiang Univ SCIENCE A, 2006 7(2), pp. 269-274.

[6] YiuHing Chan et al, “Physical Synthesis Methodology for High Performance Microprocessors” DAC, June 2–6, Anaheim, California, USA. ,2003, pp. 696-701.

[7] C. Li, C-K. Koh and P. H. Madden, “Floorplan Management: Incremental Placement for Gate Sizing and Buffer Insertion,” ASPDAC, 2005, pp. 349-354.

[8]]M. Hrkic, J. Lillis and G. Beraudo, “An Approach to Placement-Coupled Logic Replication,” DAC, 2004, pp. 711-716.

[9] T. Luo, H. Ren, C. J. Alpert and D. Pan, “Computational Geometry Based Placement Migration”., ICCAD, 2005, pp. 41-47.

[10] U. Brenner, A. Pauli and J. Vygen, “Almost Optimum Placement Legalization by Minimum Cost Flow and Dynamic programming”., ISPD, 2004, pp. 2-9.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Research and Implementation of Evaluation System Model for Grid Investment Based on Improved Fuzzy-AHP Method

Kehe Wu
School of Control and Computer Engineering ,North China Electric Power University,
Beijing,China
Chinawukehe@ncepu.edu.cn

Zhiqi Xu
School of Control and Computer Engineering ,NCEPU,
Beijing,China
Harry_xu1985@yahoo.com.cn

Cheng Duan
College of Electrical and Electronic Engineering ,NCEPU,
Beijing,China
Chinaduancheng1985@hotmail.com

***Abstract*—The evaluation of grid investment projects is an important link which is necessary in Project Cycle. It is of great significance for Decision-makers to improve decision-making level of project and investment efficiency. Based on the characteristics of grid investment projects, this paper has built an index system of multi-level evaluation and uses a new AHP method which is improved by fuzzy adjustments to avoid the ambiguity of the subjective judgments given by the experts, making the evaluation more reasonable. It also makes a brief introduction on the implementation of this model.**

Keywords-grid investment; post-evaluation; improved Fuzzy-AHP; fuzzy-adjustment

I. INTRODUCTION

The evaluation of grid investment projects is a technology economic activity after the projects are delivered and has been running for a while. It's used to evaluate and analyze the decision-making, design, construction, completion acceptance and the whole process of the production and operation of the projects. Electric enterprises have their own characteristics, so attention must be paid to the post-evaluation of investment projects. To make the evaluation more scientific and reasonable we choose to build a system of indicator evaluation which can reflect the relationship among the investment, the benefits and the grid development as objectively as possible. And this system should also cover as many indicators which affected the evaluation results from different perspectives.

The evaluation system of grid investment projects is complex with multi-level and multi-factor, and its weights calculating should follow the principle of combining qualitative analysis with quantitative analysis. AHP (Analytic Hierarchy Process) is just a way of calculating weights like this. It divides the target into several levels which have different weights and include a number of indicators, and does the evaluation with these weight coefficients. The advantage of the way is that it takes the different importance of each layer of the index system into account, while its disadvantage is that it doesn't consider the degree of ambiguity in human's judgments while constructing the factor-evaluation matrix. In this paper, a fuzzy mathematical method is introduced to the construction of the factor-evaluation matrix to make this result more reasonable for evaluating the grid investment scientifically.

II. AN MODEL OF INDEX SYSTEM FOR GRID INVESTMENT EVALUATION

Generally speaking, the evaluation of the grid investment projects consists of several aspects such as the project-process evaluation and so on. In this paper the evaluation system model has been built based on a general one to adapt to AHP method. The model has 5 first-level indicators: safety and reliability of grid, social benefit of grid, comprehensive performance of gird, development-level of grid, economic benefit of grid. Each first-level consists of several secondary-levels with more indicators in it. Figure 1 shows us an example of the model by choosing some indicators.

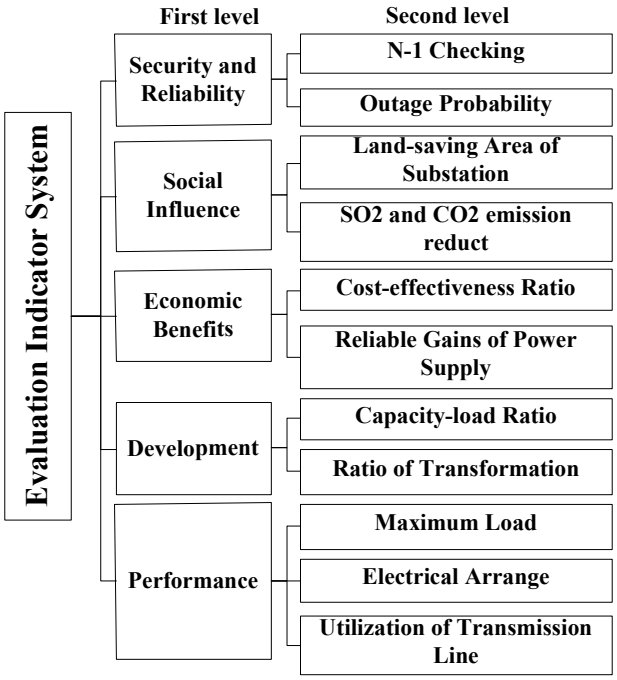


Figure.1an example of Evaluation System Model for Grid Investment

These secondary indicators in the figure above also can be divided into a series of third-level indexes through different voltage such as 1000kV, 750kV, 500(330) kV, 220kV, 110kV and so on.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.75

III. AN IMPROVED FUZZY-AHP METHOD OF CALCULATING WEIGHTS

AHP is a decision-making method with multi-criteria and multi-objective. By using this algorithm the evaluation target is subdivided into a multi-layer indicator system, and the information on the relation of each factor in the same layer is given by people who have made the judgment.

However, the socio-economic issues which experts will face to are usually much more complex and uncertain during deciding the correlation coefficients so that it's difficult to give some judgments clearly. Experts prefer using intervals than values to build the matrix. In this paper we import fuzzy-interval to AHP to calculate the weights.

The work is subdivided into 5 steps: 1) choose the factors needed and establish an objective function for post-evaluation; 2) establish the factor-evaluation matrix; 3) solve the matrix to get the weights using fuzzy interval; 4) calculate and evaluate by the normalized-values and the weights according to objective function.

A. Establishing an Objective Function for Post-evaluation

Based on the evaluation system model above, we build a multi-level hierarchical structure and express it into a linear objective function as follows:

$$M=A\times\omega_1+B\times\omega_2+C\times\omega_3+D\times\omega_4+E\times\omega_5 \quad (1)$$

In this formula, M means the evaluation score, A, B, C, D and E represent one of the five first-level indexes respectively, whose values depend on the indicator values and weights of the next layer like:

$$A=\alpha_1\times h_1+\alpha_2\times h_2+\alpha_3\times h_3+\alpha_4\times h_4+\alpha_5\times h_5$$

In the formula above, ω_i and h_i represent the weights of the first-level and second-level indicators respectively, and each of their sum is 1.

B. Establishing the Fuzzy Evaluation Matrix

The factor-evaluation matrix reflects the relevant ratio of the relative importance between the indicators in this level and in its parent level.

The first step, experts of the number k will establish an interval evaluation matrix by the measure of the relative importance in AHP respectively:

$$R_k=\begin{pmatrix} r_{11} & \cdots & r_{1m} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{mn} \end{pmatrix} \quad (2)$$

Here we have import fuzzy interval for pairwise comparisons between elements. In $r_{ij}=(l_{ij}, m_{ij}, u_{ij})$, l_{ij} and u_{ij} are the fuzzy intervals, and m_{ij} reflects the relative importance of index i than index j. And the higher the value of u-l is, the fuzzier it is; especially when its value is 0 the extent is not fuzzy. Compared with the common factor evaluation matrix, the fuzzy one has a fundamental change in giving each index a fuzzy interval $e_{ij} = u_{ij}$-l_{ij} which reflects the credibility of the evaluation results by experts and the higher it is, the less credibility it is.

The second step, we have an integration of all the factor evaluation matrixes to get WK as follows:

$$W_K=\left(w_{1k}, w_{2k}, \cdots, w_{nk}\right)^T, w_{ik}=\sum_{j=1}^{n} r_{ij} \quad (3)$$

The third step, we calculate the possibility extent of $r_{ij} \geq r_{kj}$ by Triangular Fuzzy Numbers:

$$p\left(r_{ij}\geq r_{kj}\right)=\frac{u_{ij}-l_{ij}}{\left(u_{ij}-l_{ij}\right)-\left(u_{ij}-l_{ij}\right)}\bigcap 1 \quad (4)$$

Then we would establish the matrix $P = (p_{ij})_{mn}$ by the possibility extent of $p(w_{ij} \geq w_{kj})$.

C. Adjusting the Evaluation Matrix

To avoid the errors result from subjective judgments of experts, an adjustment factor S is created to optimize the matrix as follows:

$$S=\begin{pmatrix} 1 & 1-\frac{u_{21}-l_{21}}{2m_{21}} & \cdots & 1-\frac{u_{1n}-l_{1n}}{2m_{1n}} \\ 1-\frac{u_{21}-l_{21}}{2m_{21}} & 1 & \cdots & 1-\frac{u_{2n}-l_{2n}}{2m_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ 1-\frac{u_{n1}-l_{n1}}{2m_{n1}} & 1-\frac{u_{21}-l_{21}}{2m_{21}} & \cdots & 1 \end{pmatrix} \quad (5)$$

Here $\frac{u_{ij}-l_{ij}}{2m_{ij}}$ reflects the fuzzy extent given by experts. The higher its value is, the greater the extent is, and the less credibility it is. The new matrix is as follows:

$$N=P\cdot S \quad (6)$$

Then we transfer it into a permutation the matrix with all its values in main diagonal is 1 as $M = (\mu_{ij})_{mn}$.

D. Calculating Weight Coefficient

At last, we calculate the weight of index i by square root method as follows:

$$W_i = \frac{\sqrt[n]{\prod_{j=1}^{n} \mu_{ij}}}{\sum_{i=1}^{n} \sqrt[n]{\prod_{j=1}^{n} \mu_{ij}}} \quad (7)$$

E. Calculation and Evaluation by Objective Function

We have got the normalized-values and the weight coefficients. By substituting them into the objective function (1) we can finally get the evaluation score M .

IV. The Implementation of the Algorithm

A. The Programming Framework

The grid investment evaluation system is based on the investment evaluation model. Its design is based on J2EE standards B / S centralized application model. The main functions include normalizing index values, calculating weights of indicators using AHP, calculating the evaluation scores through the objective function, and showing users the results by tables, graphs, bar charts and other forms of image in webpages. It also provides external interfaces to communicate with other relevant procedures and share information.

This program has 4 layers based on J2EE as follows:

1) Presentation Layer

This layer contains the user logic interface. It is responsible for providing users the operation interface.

In this system we choose Struts MVC framework. Struts MVC framework can make coupling with Spring framework, and drive its business logic to operate the Action. The Struts MVC framework consists of model, view and controller.

2) Business Logic Layer

Business layer is responsible for implementing all business logic rules covered in the model designing and doing the validation. It also transmits information as a bridge between presentation and persistence layer so that they do not have to communicate directly, which makes the system flexible and efficient. Business layer usually consists of several components, each of which exchanges information with system and communicates with each other through interfaces.

We choose Spring framework. The most two important parts in Spring framework is IoC (Inversion of Control) and DI (Dependence Injection) system, which means the dependencies between objects be transferred to the control of the outer container. And the container dynamically re-injected its dependencies into these components during operation.

3) Persistence Layer

Persistence layer is responsible for storing data to one or more data memories, and reading data from a set of classes and components. This layer must include a business entity model.

This layer uses iBATIS framework, in which developers could write SQL statements themselves. It is flexible to operate and easy to maintenance. iBATIS framework has abstracted and encapsulated all access to the data source to hide the implementation details through DAO (Data Access Object). When querying data, we only need to access the database via DAO.

4) Data System Layer

Data System Layer is in charge of the operation between the program and the database server, including establishing a database connection, maintaining the connection, accessing data and resource-sharing.

We have chosen Oracle as the database in this system.

The program infrastructure is like Figure.2:

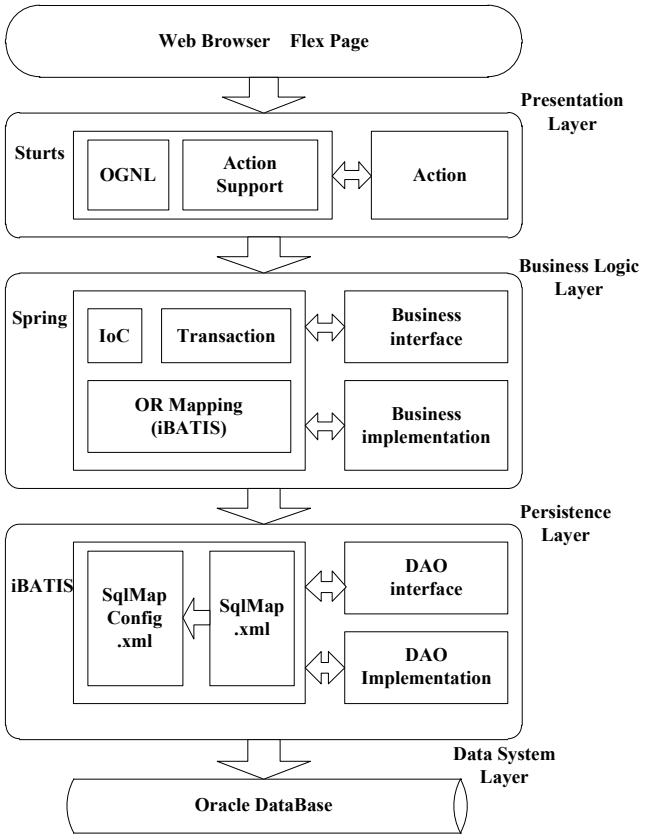


Figure.2 Program infrastructure

B. The Implementation for the Program Function

The program has mainly implemented three functions: the weights calculation, the normalization of the indicators and the comprehensive evaluation calculation. The algorithm process is in Figure.3.

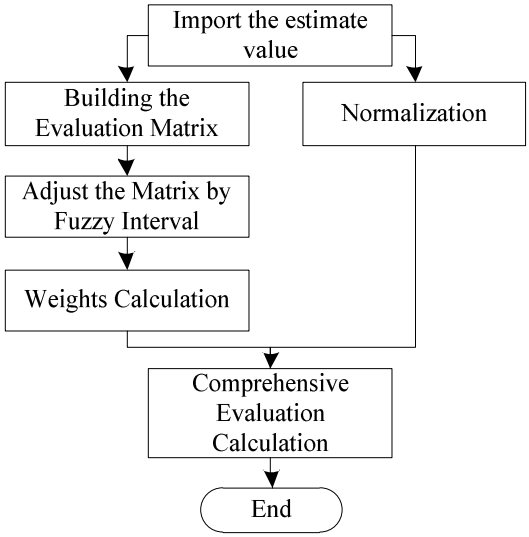


Figure.3 Algorithm process

The implementation of the business logic layer is as follows:

```
public interface AHPCompute {
   public double[][] Normalization(double[][] index){
   } //normalization of the indicators

   public double[] Weights(double[][]index, double[]
   weight, int N){
   } //calculate weights with AHP method, calling F-
   Matrix() and F-Adjust() method

   public double[][]F-Matrix(double[][]index, int N){
   }//build the evaluation matrix

   public double[][] F-Adjust(double[][]index, int N){
   }// adjust the matrix by fuzzy interval

   public double[] Assessment(double[] weights,
   double[][] index){
   }// comprehensive evaluation calculation
}
```

V. CONCLUSION

The evaluation of grid investment plays a vital role in project decision and management. Scientific evaluation of the effectiveness of investment projects and experiences in the construction of grid would give better references for grid enterprises with Limited funds. This paper imports a Fuzzy-AHP method based on fuzzy interval to reduce the subjectivity and overcome the subjective evaluation of the experts to make the results more reasonable.

Computer program is simple to operate to not reduce the workload of manual calculation, but also improve the calculation accuracy. The interfaces are flexible–designed to have better Expansibility

REFERENCES

[1] Li Yuhua , Hu Yunquan . A model of multilevel fuzzy comprehensive evaluation for investment risk of high and new technology project[C]. Proceedings of IEEE Conference on Machine Learning and Cybernetics, 2006.

[2] [Song Ping , Yang Qifeng , Feng Bin . Application of entropy coefficient optimization model of multi-criteria decision in post-evaluation of Urban Rail Transit[C] . Proceedings of IEEE Conference on Wireless Communications, Networking and Mobile Computing, 2007.

[3] Xiao Yi, Shao Dongguo, Wu Yuting. The methodology of post-evaluation of river basin harnessing project for sustainable development based on entropy theory[C] . Proceedings of IEEE Conference on Wireless Communications Networking and Mobile Computing, 2007.

[4] Wang Lin, Zeng Yurong, Zhang Dongfeng. A new approach to evaluating the criticality class of spare parts incorporating fuzzy comprehensive evaluation and grey relational analysis[C] . Proceedings of IEEE Conference on Computational Intelligence and Security, 2006.

[5] Zhang Chi , Cheng Haozhong , Xi Xun , et al . A study of distribution network feeding modes selection based on analytic hierarchy process and fuzzy comprehensive evaluation[J] . Power System Technology, 2002, 22(12) : 152-155(in Chinese).

[6] Ma Fanglan , Li Shangping , He Yulin , et al . Comprehensive evaluation of the cutting performance of sugarcane harvester based on fuzzy theory and neural network[C] . Proceedings of IEEE Conference on High-Performance Computing , 2005 .

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

QoS Modeling for Dependable and Distributed Cyber Physical Systems Using Aspect-Oriented Approach

Lichen Zhang
Faculty of Computer Science and Technology
Guangdong University of Technology
Guangzhou 510090, China

***Abstract*—In cyber-physical systems, the description, control, management, consultation and guarantee of QoS are very complex and challenging work, Quality of Service(QoS) is directly related to system's performance. In this paper, we present an aspect-oriented model for specifying Quality of Service (QoS) based on the combination of UML and RTL. We believe that the two types of notation, graphical (semi-formal) and, respectively, formal, can efficiently complement each other and provide the basis for an aspect-oriented specification approach that can be both rigorous and practical for QoS modeling. Two examples depict how aspect-oriented methods can be used during QoS analysis and design process.**

Keywords- *QoS, Cyber physical systems, Aspect-Oriented*

I. INTRODUCTION

A cyber-physical system (CPS)[1] is a system featuring a tight combination of, and coordination between, the system's computational and physical elements. Today, a pre-cursor generation of cyber-physical systems can be found in areas as diverse as aerospace, automotive, chemical processes, civil infrastructure, energy, healthcare, manufacturing, transportation, entertainment, and consumer appliances. This generation is often referred to as embedded systems. In embedded systems the emphasis tends to be more on the computational elements, and less on an intense link between the computational and physical elements.The dependability of the software [1]has become an international issue of universal concern, the impact of the recent software fault and failure is growing, such as the paralysis of the Beijing Olympics ticketing system and the recent plane crash of the President of Poland. Therefore, the importance and urgency of the digital computing system's dependability began arousing more and more attention. A digital computing system's dependability refers to the integrative competence of the system that can provide the comprehensive capacity services, mainly related to the reliability, availability, testability, maintainability and safety. With the increasing of the importance and urgency of the software in any domain, the dependability of the distributed real-time system should arouse more attention.[2]

Aspect-oriented programming (AOP) [3] is a new software development technique, which is based on the separation of concerns. Systems could be separated into different crosscutting concerns and designed independently by using AOP techniques. Every concern is called an "aspect". Before AOP, as applications became more sophisticated, important program design decisions were difficult to capture in actual code. The implementation of the design decisions were scattered throughout, resulting in tangled code that was hard to develop and maintain. But AOP techniques can solve the problem above well, and increase comprehensibility, adaptability, and reusability of the system. AOSD model separates systems into tow parts: the core component and aspects.

This paper introduce the aspect-oriented modeling method based on UML, mainly form an aspect-oriented profile by the UML meta-model extension[5], and model the crosscutting concerns by this profile. Then we make a in-depth study of UML extension of QoS, including general model, framework metamodel, profile and the QoS catalog, and highlights the dependability characteristic of QoS, the QoS aspect inherits the properties of the QoS model. Finally, we analyze the QoS aspect- oriented modeling via two examples, dividing the complex QoS into some sub-aspects, and the time, capacity and level aspect inherit the members and operations of the QoS abstract aspect.

II. ASPECTED-ORIENTED SPECIFICATION OF QOS

UML is acquainted to be the industry-standard modeling language for the software engineering community, and it is a general purpose modeling language to be usable in a wide range of application domains. So it is very significant to research aspect-oriented real-time system modeling method based on UML[6]. However they didn't make out how to model real-time systems, and express real-time feature as an aspect. In this section, we extend the UML, and present an aspect-oriented method that model the real-time system based on UML and Real-Time Logic (RTL). Real Time Logic is a first order predicate logic invented primarily for reasoning about timing properties of real-time systems. It provides a uniform way for the specification of both relative and absolute timing of events. The specification of the Object Constraint Language (OCL) [7] is a part of the UML specification, and it is not intended to replace existing formal languages, but to supplement the need to describe the additional constraints about the objects that cannot be easily represented in graphical diagrams, like the interactions between the components and the constraints between the components' communication . Since OCL is an expression language, it can be checked without an executable system. All these features turn out to be useful in representing QoS properties, which can be represented by

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.69

the combination of precondition, post-condition and invariant in OCL. The QoS attributes are represented by the member variables of the class, and the QoS actions are represented by the methods[8]. They are checked at run time, before and after the calls so that the change of the QoS parameters of the system is monitored in a timely basis.[9]

As the QoS concern needs to be considered inmost parts of the system, it is a cross-cutting concern.[10] Cross-cutting concerns are concerns that span multiple objects or components. Cross-cutting concerns need to be separated and modularized to enable the components to work in different configurations without having to rewrite the code. If the code for handling such a concern is included in a component, it can make the component tied to a specific configuration. This code will typically be scattered all over the component implementation and tangled with other code in the component. Modularizing it will make it more robust for change, and separating it totally from the component implementation will save the component programmers from having to implement it. Aspect oriented programming is a new method formodularizing cross-cutting concerns. By using AOP, concerns can be modularized in an aspect and later weaved into the code. Fig.1 shows aspect model of QoS framework.

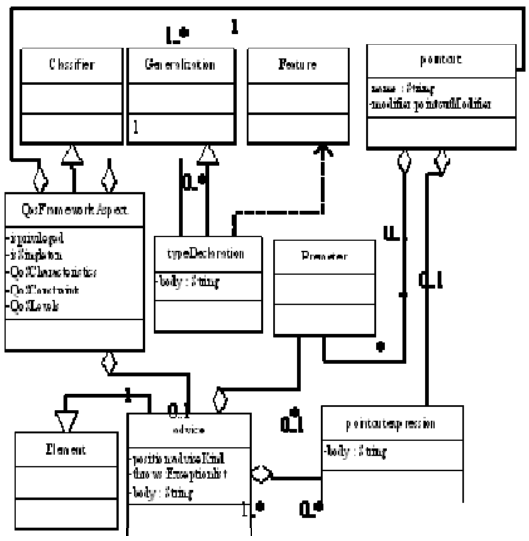


Figure 1. Aspect model of QoS framework

III. CASE STUDY: ASPECT ORIENTED SPECIFICATION OF QOS OFFIRE ALARMSYSTEM

An automatic fire alarm system is designed to detect the unwanted presence of fire by monitoring environmental changes associated with combustion. In general, a fire alarm system is either classified as automatically actuated, manually actuated, or both. Automatic fire alarm systems can be used to notify people to evacuate in the event of a fire or other emergency, to summon emergency services, and to prepare the structure and associated systems to control the spread of fire and smoke.

QoSConstraint Q1,Q2, Q5 of the fire alarm system is expressed as follows with formal technique RTL[11]:

[Q1]：∀i∃j@(↑data.collect,j)-
@(↓stop,i)≥COLLECT_MIN_TIME∧@(↓data.open,j) -@
(↓stop,i)≤COLLECT_MAX_TIME

[Q2]：∀i∃j@(↑data.process,j)-
@(↓stop,i)≥DATA_PROCESS_MIN_TIME∧ @(↓dat

a. process, j)-
@(↓stop,i)≤COLLECT_MAX_TIME

b.

[Q5]：∀i∃j@(↑alarm.process.,j)-
@(↓command.send,i)≥ALARM_PROCESS_ MINTI ME
∧@(↑alarm.process,j)-
@(↓command.send,i)≤ALARM_PROCESS_MAXTIME

QoSConstraint Q3 and Q4 of the fire alarm system is expressed as follows with XML[12]:

[Q3]： <QoS *type*=”Level”>
<Firelevel val = “FIRE_MAX_LEVEL”/>
</QoS>

[Q4]： <QoS *type*=”Contraint”>
<frame_rate val =
“FRAME_RATE_CONSTRAINT”/>
<audio_sample_rate val =
“FRAME_RATE_CONSTRAINT”/>
</QoS>

We separate QoS from real-time fire system as an aspect, the aspect-oriented model of QoS of real-time fire system is shown as Fig.2.

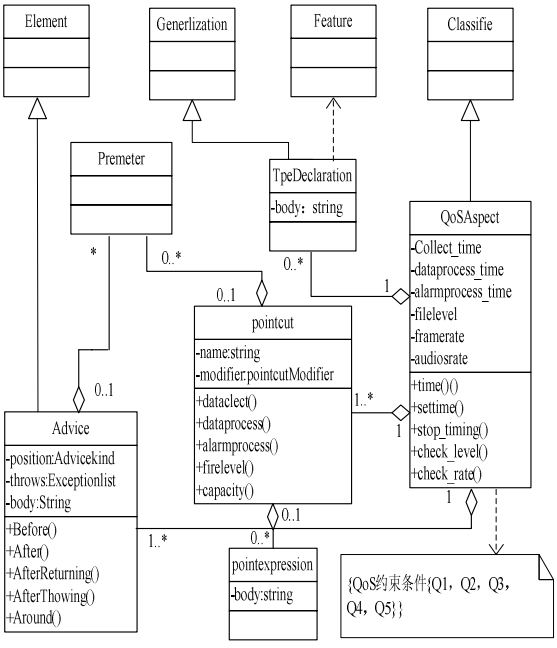


Figure 2. QoS aspect-oriented model of Fire real-time system

We use QoSAspect to express QoS of real-time fire alarm system. The class diagram of Fire Real-time System with the aspect-oriented extension is shown as Fig.3.

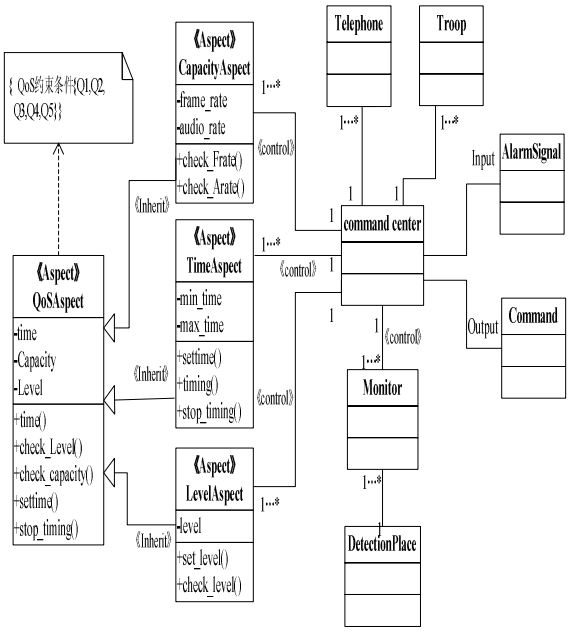


Figure 3. Class Diagram of Fire Real-time System

The QoS Aspect Weaving Diagram of real-time fire alarm system is shown as Fig.4.

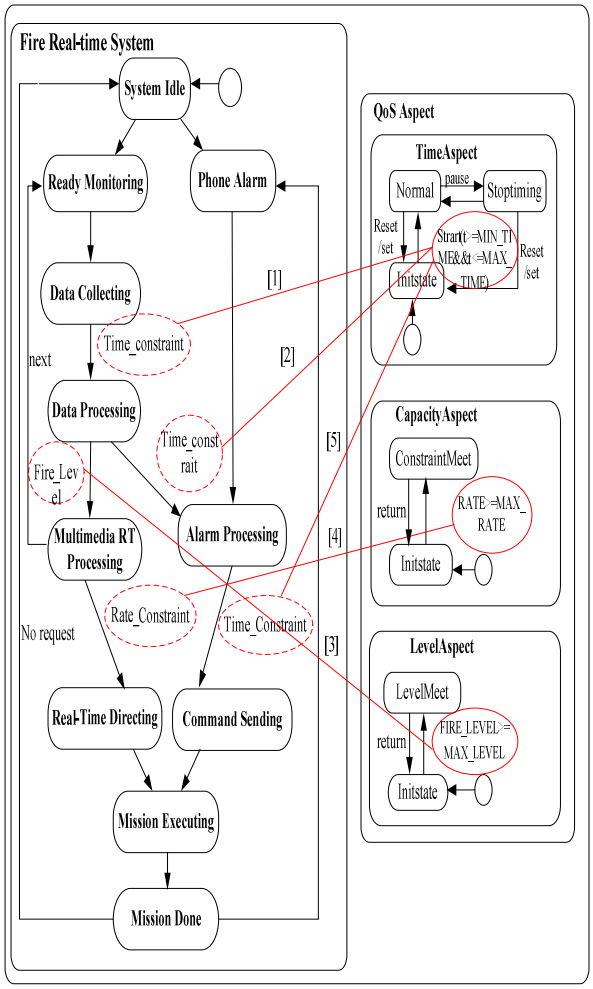


Figure 4. QoS Aspect Weaving Diagram of Fire Real-time System

IV. CASE STUDY: ASPECT ORIENTED SPECIFICATION OF ELEVATOR SYSTEM

An elevator control system illustrates the development process including static structure, dynamic behaviors and the weaving of time aspect by integration of Informal Specification and formal Specification with aspect-oriented approach. We consider that every floor has a pair of direction lamps indicating that the elevator is moving up or down. There is only one floor button and one direction lamp in the top floor and the bottom floor. Every floor has a sensor to monitor whether the elevator is arriving the floor.

We consider that the elevator is required to satisfy the following timing constraints[13]:

[T1] After the elevator has stopped at a particular floor, the elevator's door will open no sooner than OPEN_MIN_TIME and no later than OPEN_MAX_TIME.

[T2] After the elevator has stopped at a given floor the elevator's door will normally stay open for a

STAY_OPEN_NORMAL_TIME. However, if the CloseDoorButton on board of the elevator is pressed before this timeout expires, the door will close but no sooner than STAY_OPEN_MIN_TIME.

[T3] After the door is closed, the movement of the elevator can resume, but no sooner than CLOSE_MIN_TIME, and no later than CLOSE_MAX_TIME.

Separation of Concerns From Elevator System

Several concerns can be separated from the elevator control system, such as time aspect, control aspect, and concurrency aspect. The development process of the elevator control system is shown as Fig.5. However, in this paper we only simply consider the time aspect, and will complete other aspects in our future work.

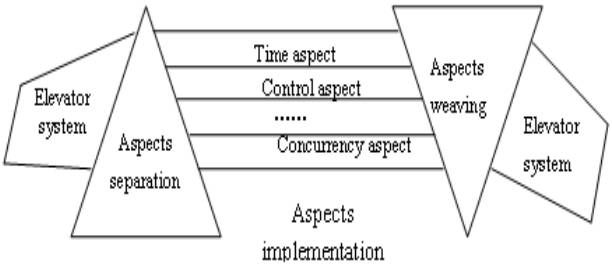


Figure 5. The Development Process of The Elevator Control System

Structural Description Using Class Diagrams

The real-time feature of real-time systems can be modeled using UML by extending stereotypes, tagged values, and constraints before. For example, timing constraints can be added on the class to express the time feature in class diagram. But the implementation of the time feature were still scattered throughout, resulting in tangled code that was hard to develop and maintain. So we describe the real-time feature as an independent aspect according to the AOP techniques, and design a time model to realize and manage the time aspect in order to make the system easier to design and develop and guarantee the time constraints.

We separate the real-time feature as a TimeAspect, which is an instance of <<aspect>> in the elevator control system. The TimeAspect crosscuts the core functional class by stereotype <<crosscut>> in class diagram. Also timing constraints can be attached to the TimeAspect explicitly. The elevator control system class diagram is shown in Fig.6.

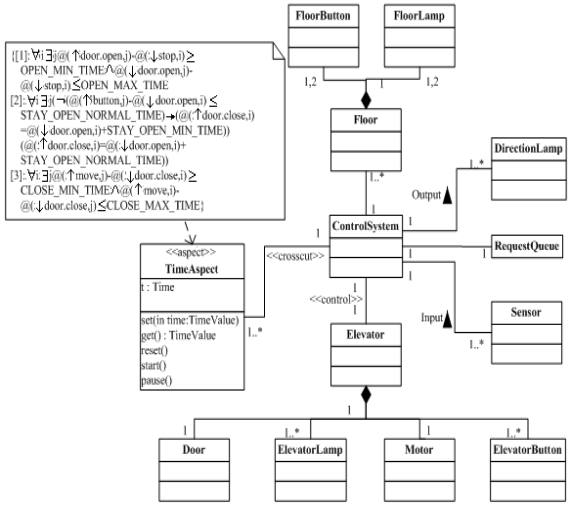


Figure.6 The Elevator Control System Class Diagram

Behavioral Description

UML has five behavioral diagrams to describe the dynamic aspects of a system as representing its changing parts. Use case diagram organizes the behaviors of the system. Sequence diagram focuses on the time ordering of messages. Collaboration diagram emphasizes on the structural organization of objects that send and receive messages. Statechart diagram focuses on the changing state of a system driven by events. Activity diagram focuses on the flow of control from activity to activity messages. Use case diagram, collaboration diagram, and sequence diagram belong to Inter-Object behavior diagrams. While statechart diagram belong to Intra-Object behavior diagrams.The time behavior is depicted by extending timing marks in the statechart traditionally. can model the intra-object aspectual behaviors well in the object-oriented programming paradigm. However, current specification of statecharts doesn't support aspect-oriented modeling. To support aspect-orientation within the context of statecharts, we need to provide a mechanism by which the modeler can express these aspects. Statecharts modeling aspects should consider the association between aspects and transitions instead of states. Orthogonal regions, which are shown as dashed lines in statecharts, combine multiple simultaneous descriptions of the same object. The aspects can be expressed as objects, which have their own sub-states. Interactions between regions occur typically through shared variables, awareness of state changes in other regions and message passing mechanisms such as broadcasting, and propagating events . The statecharts of the elevator control system is shown in Fig.7. Timing behaviors are described by the advanced features of statecharts, and the time concern is achieved implicit weaving with the core functionality of the system. Statecharts refine the model and aspects codes can be generated automatically by existing CASE tools.

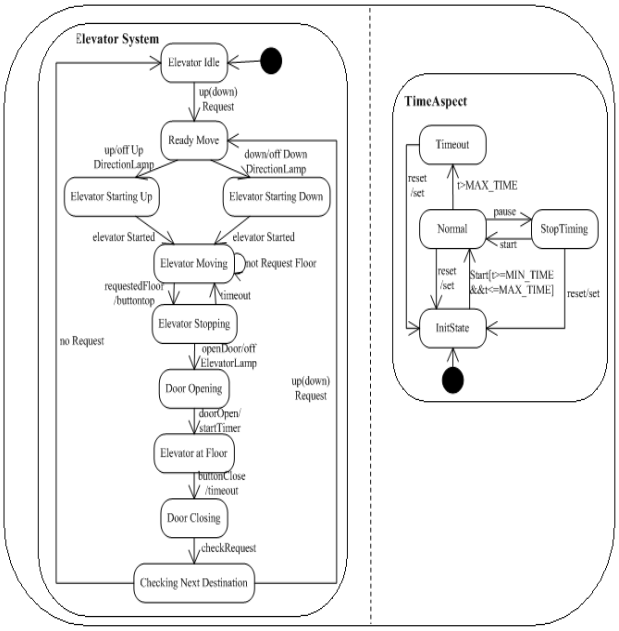


Figure 7. Statechart of the Elevator Control System

Weaving of Time-Aspect

The time aspect can be woven into the real time system by using the UML's statecharts as shown in Fig.8.

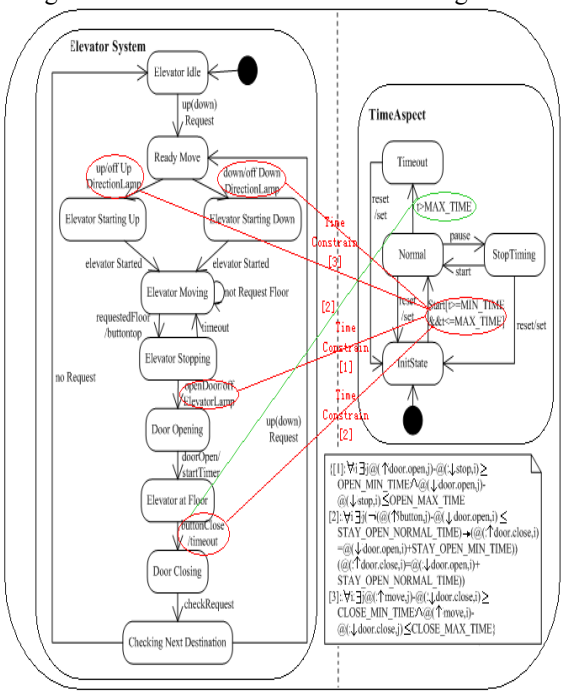


Figure 8. Weaving the Time Aspect

V. CONCLUSION

In this paper, we presented an aspect-oriented model for specifying Quality of Service (QoS) based on the combination of UML and RTL. Two types of notation, graphical (semi-formal) and, respectively, formal, can efficiently complement each other and provide the basis for an aspect-oriented specification approach that can be both rigorous and practical for QoS modeling. Two examples depicted how aspect-oriented methods can be used during QoS analysis and design process.

Future works will focus an automatic weaver for aspect oriented model of QoS.

ACKNOWLEDGMENT

This work is supported by the Major Program of National Natural Science Foundation of China under Grant No.90818008.

REFERENCES

[1] Lee E A.Cyber physical systems:Design challenges //Proc. of the 11th IEEE Int Symp on Object Oriented Real-Time Distributed Computing.Piscataway, NJ:IEEE,2008:363-369

[2] A. Svizienis, J.C. Laprie, and B. Randell. Dependability of computer systems: Fundamental concepts, terminology, and examples. Technical report, LAAS-CNRS, October 2000.

[3] Kiczales, G., et al. Aspect-Oriented Programming. Proceedings of the 11th European Conference on Object-Oriented Programming, June 1997.

[4] Svend Frolund and Jari Koistinen. Quality of Service Specification in Distributed Object Systems. IEE/BCS Distributed Systems Engineering Journal, 5:179–202, December 1998.

[5] Aldawud, T. Elrad, and A. Bader. A UML Profile for Aspect Oriented Modeling, Workshop on AOP, 2001.

[6] Clemente, P. J., Sánchez, F., and Perez, M. A. "Modelling with UML Component-based and Aspect Oriented Programming Systems",Seventh International Workshop on Component-Oriented Programming at European Conference on Object Oriented Programming (ECOOP). Málaga, Spain.,2002,pp.1-7

[7] Lavazza, G. Quaroni,M. Venturelli, "Combining UML and formal notations for modeling real-time systems", ACM SIGSOFT Software Engineering Notes, vol. 26, Sep. 2001, pp.196-206

[8] Object Management Group. UML Profile for Modeling Quality of Service and Fault Tolerance Characteristics and Mechanisms Joint Revised Submission, OMG Document realtime/03-05-02 edition, May 2003.

[9] OMG. UMLTM Profile for Schedulability, Performance, and Time Specification,http://www.omg.org/cgi-bin/doc?formal/2005-01-02,2005.

[10] Wehrmeister, M.A., Freitas, E.P., and Pereira, C.E., et al., "An Aspect-Oriented Approach for Dealing with Non-Functional Requirements in a Model-Driven Development of Distributed Embedded Real-Time Systems ", 10th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing, Santorini Island, Greece, May7-9, 2007, IEEE Computer Society, pp.428-432

[11] Farnam Jahanian, Aloysius K. Mok.Safety Analysis of Timing Properties in Real-Time Systems. IEEE Trans. Software Eng. 12(9): 890-904 (1986)

[12] S.Frolund, and J.Koistinen. QML: A Language for Quality of Service Specification, Technical Report HPL-98-10, Feb.1998

[13] Combining Semi-Formal and Formal Notations in Software Specification: An Approach to Modelling Time-constrained Systems"

The Optimization about PID controller's parameter based on QPSO Algorithm

Liu Jing-hui
School of Internet of Things Engineering
JiangNan University
Wuxi, Jiangsu, China
hui8609@126.com

Xu Wen-bo
School of Internet of Things Engineering
JiangNan University
Wuxi, Jiangsu, China
xwb@sytu.edu.cn

***Abstract*—PID controller for control system parameter optimization problem, we introduced Quantum-behaved particle swarm optimization (QPSO) algorithm, this method has the quantum behavior of particles, which can search the entire feasible region, the state only with particles position vector to describe, and they rely on only one parameter, the convergence speed and global convergence properties are superior to the traditional PSO. Simulation results verify the feasibility of quantum particle swarm optimization and effectiveness.**

Keywords- PID controller；Optimization about parameter; Quantum-behaved Particle Swarm Optimization

I. INTRODUCTION

PID control due to its simplicity and robust applications is one of the most widely used control strategy in industrial process control. The basic idea is to deviation through the proportional, integral and differential form linear combinations of the three parameters of the controller to control the controlled object. Advantages and Disadvantages of system control quality depend on the setting of these three parameters. Intelligent traditional classical tuning algorithm is mainly Z-N method, which algorithm is simple, easy to implement merit, but lack flexibility, and sometimes oscillate and large overshoot [1].To improve the effect of PID tuning parameters, many intelligent methods should be used by the PID parameter tuning up, such as particle swarm optimization (PSO)[2-3] and so on. Particle swarm search algorithm in the process of parameter tuning is easy early, vulnerable to the shortcomings of local optimum, selected Quantum-behaved Particle Swarm Optimization (QPSO) for parameter optimization, and compared with the particle swarm algorithm in this paper.

II. PID CONTROLLER

PID controller includes the proportion of links, Integral and differential link. Control system shown in figure 1:

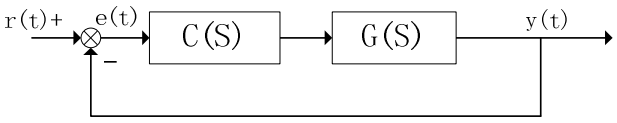


Figure 1. PID control system

C(S) is the controller transfer function; G(S) is the system transfer function； r, e, y denote the reference quantity, volume and output error.

$$C(S) = K_p + \frac{1}{T_I * S} + T_d * S$$

K_p、T_i、T_d determine the controller's proportional, integral and differential parameters. PID controller parameter tuning is to find the right K_p、T_i、T_d, a balance between the performance indicators in the system, such as the balance between the control system stability, and the balance between the Overshoot, steady-state error and

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.93

dynamic response rate.

III. DESCRIPTION ABOUT ALGORITHM

A. PSO algorithm

PSO algorithm is proposed by the American social psychologist JamesKennedy and Electronics Engineers RussellEberhart in 1995[4]. The basic idea is to simulate birds in group behavior, the use of a biological population model and "Group" and "evolution" concept, and based on the fitness value of individual particle to operate. PSO algorithm put each individual as a particle without the weight and volume in the N-dimensional search space, which flight in the search space to a certain speed. The Speed is adjusted by individual flying experience and groups of flying experience. Now most of the expansion and improvement about PSO algorithm examination is base on PSO algorithm with inertia weight[5], to this end the PSO algorithm with inertia weight is called the standard PSO algorithm.

Set $X_i = (X_{i1}, X_{i2}, ..., X_{in})$ as the current position of particle i; $V_i = (V_{i1}, V_{i2}, ..., V_{in})$ as current flight speed of particle i; $P_i = (P_{i1}, P_{i2}, ..., P_{in})$ as best position experienced of particle i, experienced by the particle i with the best fitness value of the position, Called the individual best position ($pbest$).Standard PSO algorithm evolution equation can be described as follows:

$$V_{ij}(t+1) = \omega V_{ij}(t) + c_1 r_{1j}(t)\left(P_{ij}(t) - X_{ij}(t)\right) + c_2 r_{2j}(t)(P_{gj}(t) - X_{ij}(t)) \quad (1)$$

$$X_{ij}(t+1) = X_{ij}(t) + V_{ij}(t+1) \quad (2)$$

Where subscript i denote particle i; subscript j denote the first n-dimensional; subscript t denote the first t generation. c_1, c_2 are Constant acceleration; ω is inertia weight; $r_{1j}(t) \sim \cup(0,1)$ $r_{2j}(t) \sim \cup(0,1)$ are two independent random number. Parameter ω can be fixed, also can be a linear change. c_1, c_2 are undetermined constants, and usually be assigned values in [0, 2]. According to the analysis of convergence behavior of particles, to ensure convergence of the algorithm, each particle P must converge to the respective point, which is determined by the follow of particles and the aggregation of particle group. In the standard PSO system, particles in the form of convergence is achieved orbit, and the particle velocity is always limited, so the particles in the search process the search space is a limited area, can't cover the whole feasible space. Therefore, the standard PSO algorithm does not guarantee convergence to global optimal solution, which is its greatest weakness.

B. Particle Swarm Optimization Algorithm Based on Quantum Behavior

In 2004, Sun and others from the perspective of quantum mechanics proposed PSO algorithm, a new model, this model is based on the DELTA potential well, and according to this model was proposed based on the quantum behavior of particles swarm optimization [6]. In the quantum space, particles in the whole feasible solution space to search, so QPSO global search performance is much better than the standard PSO algorithm. QPSO algorithm uses the wave function $\varphi(x,t)$ to describe the state of particles. And get the particles' probability density function, a point in space, by solving the Schrodinger equation, and then, by MonteCarlo random simulating, particle position equation is:

$$X(t) = P \pm \frac{L}{2} \ln \frac{1}{u} \quad (3)$$

Where u is a random number which distribute uniformly in [0, 1], the value is determined by:

$$L(t+1) = 2 * \beta * \left| mbest - X(t) \right| \quad (4)$$

Finally, the evolutionary algorithm by QPSO equation is:

$$mbset = \frac{1}{M}\sum_{i=1}^{M} P_i = \left(\frac{1}{M}\sum_{i=1}^{M} P_{i1}, \frac{1}{M}\sum_{i=1}^{M} P_{i2}, ..., \frac{1}{M}\sum_{i=1}^{M} P_{id} \right) \quad (5)$$

$$P_{id} = \varphi P_{id} + (1-\varphi) * P_{gd} \quad (6)$$

$$X_{id}(t+1) = P_{id} \pm \beta * \left| mbest - X_{id}(t) \right| * \ln \frac{1}{u} \quad (7)$$

Where L is the number of particles, D is dimension of the particle, u and φ are two random number in the average [0, 1], $mbest$ is the average best position point of all particles. And standards, like PSO, Pi is the best position of particle experienced; P_j is the best of all the particles position the population experienced. β be called the

contraction and expansion coefficient, is the only parameter of QPSO.

In QPSO algorithm, the state of particle only be described by position vector. The algorithm only have one control parameter, so this parameter choice and control is critical, it is related to the convergence speed. In each iteration, $\pm$ is decided by the random number generated randomly in [0, 1], when the generated random number is greater than 0.5, check - sign; otherwise take the + sign. Quantum particle swarm algorithm in the function test[7], filter design[8], multi-stage financial planning[9], neural network optimization[10] and other applications has demonstrated its superiority.

IV. QUANTUM PARTICE SWARM OPTIMIZATION PID CONTROLLER

Optimization of designing PID controller is a compromise balance between all performances of the system, so parameter setting can be seen as more than one objective optimization problem. Quantum particle swarm algorithm is a global search algorithm, with global convergence. Using it for optimization of PID controller parameters can overcome a number of other intelligent algorithm parameter tuning of the lack of access to satisfactory performance with PID controller. Specific implementation steps are as follows:

Step1. Initialize

Initialize a population of particles with random positions and velocities on D-dimensions in the problem space $X(0)=[X^1(0),X^2(0),\cdots,X^n(0)]$. $X^j(0)=[K_p^j(0),T_i^j(0),T_d^j(0)]$ is the coordinate component of the particle j in the solution space, as three parameters of PID controller. To take full advantage of the results of Z-N method and reduce the searching space, we take extending space, which center is the result of Z-N algorithm, as the search space of two others algorithms. The scope of the search space is decided by equation (8), equation (9) and equation (10).

$$(1-\lambda)K_p^{'} \le K_p \le (1+\lambda)K_p^{'} \tag{8}$$

$$(1-\lambda)T_i^{'} \le T_i \le (1+\lambda)T_i^{'} \tag{9}$$

$$(1-\lambda)T_d^{'} \le T_d \le (1+\lambda)T_d^{'} \tag{10}$$

K_p、T_i、T_d are the parameter values, which have been optimized by Z-N algorithm if PID controller.

Step2. Defining objective function

We take the weights, of output overshoot and steady-state error and adjust time, as performance function, to estimate the comprehensive performance of PID controller.

$$J(P^j)=\omega_1\sigma+\omega_2 t_S+\omega_3 e \tag{11}$$

$\omega_1,\omega_2,\omega_3$ are the weighting coefficients, and their size on the performance of the optimized PID controller has great influence on system performance. According to equation (11) calculated the population fitness value of each particle by the transfer function $f(P^j(0))=J(P^j(0))$ to evaluate the merits of the initial population of particles. Individual best position will be initialized to be:

$$P^j{}_{best}(0)=P^j(0), j=1,2,...,n \tag{12}$$

The best particle position of the fitness of the population assigned to $P_{gbest}(0)$.

Step3. Iteration

According to equation (5), (6), (7), update the current location of each particle generation.

Step4 . Updated

According to compare the fitness value of location information between the current generation and parent generation, get a new best position and best particle of group. The fitness value of current particle is $f(P^j(t))$; t is current generation; if $f(P^j(t)<f(P^j(t-1))$, then $P^j{}_{best}(t)=P^j(t)$, else $P^j{}_{lbest}(t)$ unchanged; if $f(P^j(t))<f(P_{gbest}(t-1))$, then $P^j{}_{gbest}(t)=P^j(t)$, else $P_{gbest}(t)=f(P_{gbest}(t-1))$.

Step5. Determine

Whether the terminating condition (the text is set to iterate 30 times), otherwise $t = t + 1$, go back to step3.

V. SIMULATION RESULTS AND ANALYSIS

Optimize PID parameter by Z-N, PSO and QPSO respectively.

$$G(S) = \frac{1}{S(S+2)(S+3)}$$

In the optimization process, the population size is 30; the number of iterations is 30, PSO algorithm acceleration factor coefficient $c_1 = c_2 = 2$. β in QPSO algorithm Obey the linear decline from 1.0 to 0.5. Weighting coefficient $\omega_1, \omega_2, \omega_3$ are 1.0, 0.1, 1.0. In order to study the stability of the algorithm to eliminate the chance, the two algorithms have conducted 10 rounds of experiments. Table 1 shows the results of their experiments:

10 optimal parameters and performance indicators						
Algorithm	K_p	T_i	T_d	σ	t_s	J
Z-N	7.5320	2.1452	1.0296	0.2610	2.3103	5.1437
PSO	5.6542	1.4561	1.7756	0.1572	1.6428	3.3658
QPSO	5.6235	1.4012	1.8035	0.1474	1.6246	3.1250

Table 1 Experimental Algorithm for optimal PID parameters

Table 1 shows the experimental data of QPSO algorithm faster than the PSO algorithm to find the optimal value of PID parameters. The average optimal value and the variance ratio of QPSO algorithm is than small QPSO algorithm in Table 1; QPSO can find the optimal value every time or very close to the optimal value in 10 experiments of the optimization process, but PSO algorithm relapse into a local optimal value in the 10 testes, indicating that the QPSO algorithm has better global convergence and stability than others.

Curve shows that QPSO algorithm has better rising time and overshoot than PSO algorithm in the step response.

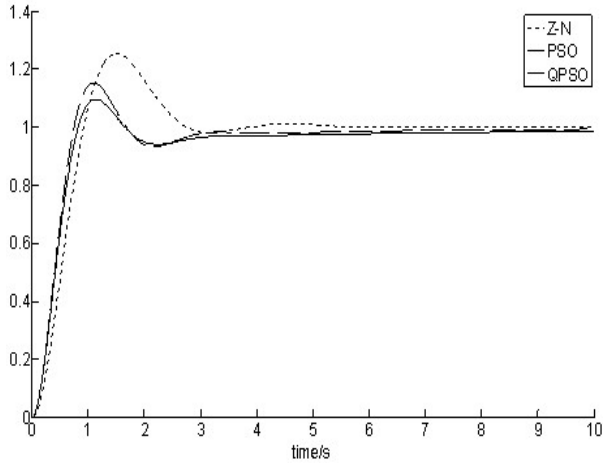


Figure 2. PID algorithm to optimize the three parameters

VI. Conclusions

For taking into account the performance indicators in control system, this paper propos the important gist that judge the system performance by Weighting and integrating the performance index. And for this multi-objective optimization problem, pointing out that some intelligent optimization algorithm to optimize the process parameters of the shortcomings. By analyzing the QPSO algorithm, proposed the use of quantum particle swarm algorithm optimization of PID controller parameters. By simulating three algorithms: Z-N, PSO and QPSO, it shows QPSO can find the optimal solution in the global scope effectively and quickly, for optimizing the parameters of PID controller. In fact, QPSO algorithm can also be applied to other control design.

VII. REFERENCES

[1] De Moura P B O,Cunha J B,Coelho J P.Design of PID controllersusing the particle swarm algorithm[C]//The IASTED Conference on Modelling,Identification,and Control,2002

[2] Liu Yi-jian,Zhang Jian-ming,Wang Shu-qing.Optimization designbased on PSO algorithm for PID controller[C]//Intelligent Controland Automation,WCICA,2004

[3] Kennedy J,Eberhart R.Particle swarm optimization[C]//Proceedings of IEEE International Conference on Neural Network,1995

[4] Shi Y,Eberhart R C.A modified particle swarm optimizer[C]//IEEE International Conference of Evolutionary Computation,Ancharage,Alaska,1998.

[5] Doug Stuns, eds Zhao lock translation. OCP： Ora-cle9iDBA basis II . Beijing： Electronic Industry Press, 2002

[6] SUN J, FENG B, XU W. Particle Swarm Op timization with Particles Having Quantum Behavior, IEEE Proc. of Congress on Evolutionary Computation, 2004

[7] FangWei, Sun Jun, XuWenbo, Design∏R Digital FiltersUsingQuan2tum2Behaved Particle Swarm Op timization. ICNC(2) 2006

[8] Sun Jun, Xu Wenbo, Fang wei. SolvingMulti2period Financial Plan2ning Problem Via Quantum2Behaved Particle Swarm Algorithm. ICIC(2) 2006

[9] Sun Jun, Xu Wenbo, Liu J ing. Training RBF Neural Network ViaQuantum2Behaved Particle Swarm Op timization. ICON IP (2) 2006：

[10] XIM L, Sun J, XuW B. Quantum2Behaved Particle Swarm Op timiza2tion for designing H infinity structured specified controllersDCABES

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

A Definition of the Water Information

Ping Ai, Ping Mu
College of Hydrology and Water Resources
Hohai University
Nanjing, China
E-mail: aip@hhu.edu.cn, muping@hhu.edu.cn

Ya-Li Chen
The Bureau of Hydrology
Yangtze River Water Resources Commission
Wuhan, China
E-mail: chenyl@cjh.com.cn

Abstract*—*In order to research water information essentially, a summary of information definitions or explanations is given. And then the paper puts forward an epistemological information definition to meet to the need of researches. On the basis of the general information definition and the distinct characteristics of water, the authors define water information. Water information is also classified according to current situation of the research in water resource. At last, the features of water information are discussed.

Keywords-definition;information;water information

I. INTRODUCTION

As a kind of material, water with natural laws circulates, motions, changes in the big system of the earth, as in [1]. In order to research the physical, chemical, and biological characteristics of water, different subjects about water, such as hydraulics and river dynamics, hydrology and water resources, water environmental and ecological hydrology, correspondingly emerge. Water information is all the information about water, and is the foundation of all the researches about water. These disciplines respectively study water information in their own fields. In essence, from different angles people subdivide water information, and make water information have different features, different sources, different patterns, different mediums, different time, different places, and different manifestations, as in [2]. However, the hydrologic cycle system is open and complicated, so we should study water information based on the Systems Theory. At present, there are rarely statements or explanations about water information, not to mention the authoritative definitions. The professor Ping Ai says in work, as in [1], that the essence of water information is numeral description of water affairs activities; Jing-jing Zhang thinks that water information is generally a kind of information which provides decision-making service for flood control and drought relief, the development, utilization, configuration, protection, management of water resource, as in [3]. Water information is the indispensable fundamental information in the domains of national economy, society, resources and environment, as in [4]. Just as the professor Yue-bo Xie says in his work water information technology. In my opinion, if we want to comprehend the essence of water information, first we should make clear the definition of information, or rather we should master the essence of information.

II. INFORMATION

We are all known to information, whether in our daily life or in scientific research, we often relate to information, yet for the concept of information, there is no unified expression. The authors search more than 100 kinds of definitions or explanations about information, and summarize the definition of information from four aspects, a philosophical sense of information, information as a kind of data, information being material attributes, information in the field of communication engineering, and then put forward the definition of information in the sense of "epistemology".

A. Information in Philosophical Category

This paper discusses the definition of information from the objective existence and subjective access. We define the objective existenting information as "Source Information", it independently existing regardless of whether there is a follower, whether they can receive and whether it can be recognized. This is in accordance with the "Ontological Information", which is put forward by the professor Yi-Xin Zhong. Under such circumstances, information is regarded as an objective existence, and it is self-expressions of the moving state of things and changes of moving states.

The information acquired through the objective endeavor can be defined as "Perceptible Information". Firstly, "sense" is required for information acquisition; it contains the perception and the extension of perception, the sensor. "Know" means understanding, comprehending, realizing. This saying corresponds to "Epistemological Information" of Prof. Yi-Xin Zhong, that is to say, it is the cognitive subject that perceives or represents the moving states of things and their changing patterns, including the form, meaning and utility, as in [5-6].

B. Information as a Kind of Data

From the view of mathematics, Vance thought that information is a kind of data organized according to certain form, that is to say, it could be stored, analyzed, displayed, and communicated through language, diagrams, or digits. After explanation, information can be deemed as a kind of information with the rich meaning in its structure, as in [7]. Through analyzing the characteristics of data, we can find its regularity, thus have it serve decision-making. The effective decision-making is based on the data and information. Information is the foundation of scientific decision. In this sense, information is valuable.

C. Information as a Kind of Basic Attributes of Material

Norbert Wiener said, information is information, neither matter, nor energy, as in [8]. Physicists Stonier Tom pushed forward the opinion that material, energy and information are three footstones supporting human society, and tried to make information be the basic attribute of the universe. He hoped that information as basic concept would appear in physics as with movement and force, as in [9].

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.1

Information is a kind of attribute of material. Quality, energy and information are three elements constituting every object in the physical world. Quality is the measurement of inertia, energy is motive power to maintain material itself and drive its movement, and information is a reflection of the material existing way. Quality, energy and information are complementary, none of the three complements can be dispensed with, and they construct different dimensional relationships togethe, as in [10].

D. Information in Communication Engineering

In 1948, the founder of Information Theory C. E. Shannon said, "The basic problem of communication is that the chosen news in the one side accurately or approximately recurs in the other side", as in [11]. Apparently, Shannon studied the information from the perspective of communication, and he defined information in the terms of metric. "Information is decrease or elimination of uncertain things people all known, also can be understood as the discrepancy between uncertainties", he also gave the method of calculating the amount of information.

In addition, we can also define or explain information according to the properties of information, such as uncertain information, fuzzy information, random information, grey information, pan-gray information, as in [12], etc.

E. Information in Epistemology

Some statements about information have been discussed in the above, they are the studies about information from a couple of different perspectives, in some sense they are reasonable to describe information. On the other side, they are one-sided, cann't express the information totally and objectively. The authors think that information not only includes the changes of object, but also the subject's reflection of changes, falls into the general category of epistemology. When the object changes, some subjects can feel these changes, namely they can get information through the changes of object. However, other subjects cannot feel, thus they cannot obtain information that object brings. Different subjects are different in feeling the changes of object, for the same changes, some subjects can get more information, however, some get less. Information is closely related to subject. If the information can not bring something new, it is not deemed useful for subjects, then it does not belong to our needed information, that is to say, information must meet the need of subject, and can help subject know the objective world.

Of course, the information we discussed above is "public information", there is no restriction on disciplines or domains. If we put the "public information" into some fields of disciplines or domains, it will be "professional information", just as water information, medical information, biological information, etc. Water information as a kind of "professional information", we can define it according to the definition or explanation of "public information".

III. WATER INFORMATION

Water, as one of the most active substance circulates all the time on the earth. The existing forms of water can be divided into three types: solid, liquid and gaseous water. Water motions from sea to land and then back to sea, in the process the existing form of it will change accordingly, from liquid to gas or from solid to gas, or to liquid to solid, namely the description of the information change in water circulation; In the circulation of water movement, because the river's slope and the boundary are different, water levels, velocity and flow, etc. are in constant changes. In the changing process from the laminar flow to turbulence flow or from the supercritical flow to the subcritical flow, the Froude number, Reynolds number change constantly.

A. The definition of Water Information

In a broad sense, water information is involved with all the water-related information. It comprises the existing forms, characteristics of motion, and patterns of change, as well as the surrounding environment of water. From a philosophical sense, this water information is objective, universal. The existence of water information is irrelevant to the subject, in other words, even if there is no human being, water information still exists; Water is widely distributed throughout the earth, it is the necessary condition for the survival of all living things, as long as where there is water, there is living beings, there is water information.

Narrowly, the water information is the one that is closely related to the living and production of human being. The narrow water information here iswith the addition of the human on the basis of the broad water information, in other words, we can understand the water information from the perspective of "epistemological". Human beings pay close attention to the water information that is closely related to their own living and development. The water information we have researched in the paper is the narrow one, its essential is the numerical description of the water affairs activities. The subject (manager or operator) of the water affairs activities is the human being, the object (being manipulated or managed) is water.

Water information is the basic premise of human survival and the development, and scientific research, whether in the daily life, or in the management application, such as hydrological forecasting, hydrological simulation, water resources scheduling, and water environmental management, etc. are inseparable from water information, which is the precondition to develop water conservancy.

B. Water information Classification

In order to better study water information, human subdivides water information into different categories according to their own research needs to be, such as surface water information, ground water information, water environmental information, water conservancy and other class information etc. . At currentt the classification for water information has no authoritative uniform norms, Prof. Ping Ai in his book, Introduction to Water Information Project refers that, for their own survival and development, human beings construct water conservancy facilities, and adjust the ways of utilization of water, change the boundary conditions of movement and change of water, and its quality factor, to serve for their own purposes. Analyzing the interaction of

human beings and the water, we can subdivide water information into the following four basic themes: the water information describing water cycle description, the water information describing hydraulic engineering (facilities), description of water management and the type of water science and technology description.

The information is the basic properties of substance. As water information is one of the information, we can subdivide water information into two categories in general: basic information, derived information, according to their properties. Basic and original information is that we can directly observe or can directly observe with the help of instrument and equipment, such as water level, rainfall, evaporation. The so-called derived information is that we cannot directly observe, but we can get it by calculating and reasoning according to the basic information, such as flow rate, sediment concentration.

At present, the reason why all kinds of models cannot meet the expected accuracy is largely owing to the selected water information. The basic, original information and derived information related to the water information will be the focus of research in the future.

C. Characteristics of Water Information

As is known to all, water is the source of life. Water is widely distributed on the earth, which determines water information being the feature of universality. Simultaneously, influenced by many factors, such as climate, geography and the development of economy, water information shows the characters of regionalism, non-repeatability and periodicity.

1) *Universality:* Water is among the most active substances, which is distributed widely in rivers, lakes, atmosphere, oceans and underground. It is no exaggeration to say that where there is life, there is water. Aimed at researching into water, water information also shows universality.
2) *Regionalism:* Different regions, different climate, different landscape makes the tremendous diversity of water information, mainly in the type and quantitative value of information. The ultimate aim of researches in water information is to improve human lives and society development. We shall choose the type of information according to the need of researches.
3) *Non-repeatability*: Water information is closely related to time. A significant property of water information is time dimension, which makes it different from test data. We can acquire accurate test data through repeated tests. Water information, however, cannot be surveyed repeatedly. If we cannot get the current water information, we will never get it.
4) *Periodicity*: Water information can reappear historically. Thanks to periodicity, we can make prediction and forecast. Different water information has different periods, like annual period, monthly period and daily period.

IV. CONCLUSION

Water information is indispensable in various fields, such as hydrologic forecast, water resource management, risk analysis, water conservancy construction and water disaster prevention. It is the foundation of water affairs activities. The success of such activities relies on the correct choice of water information. Now it is the critical stage of water informatization. It is of great significance to deliberate the substantive characteristics of water information, which is the main research object of water informatization. This paper defines broad water information and narrow one, and classifies the latter in order to offer reference for water information researches. The analysis of water information features is also helpful for the better use of water resource and better prevention of water disasters.

ACKNOWLEDGMENT

This research was supported by the National High-Tech Research and Development Plan (863) of China under Grant no. 2006AA01A126; the Key Project of Science and Technology Research of the Ministry of Education of China under Grant no. 107056;the Major Research Plan Training Project of the National Natural Science Foundation of China under Grant No. 90924027.

REFERENCES

[1] Ping Ai. Introduction to Water Information Engineering [M].Wuhan: Changjiang Press, 2010
（艾萍.水信息工程引论[M]，武汉：长江出版社,2010.）

[2] Xianjian Xiao, Tanghuai Fan, Xijun Yan, Huibin Wang, Lizhong. Xu Acquisition and Processing Techniques of Hydroinformation and Its Development [J]. Advances in Science and Technology of Water Resources, 2008（2）,28（1）:86-94.
(肖贤建,樊棠怀,严锡君,王慧斌,徐立中.水信息获取与处理技术及发展[J].水利水电科技进展,2008（2）,28（1）：86-94)

[3] Jingjing Zhang, Jinshui Chen. A WebGIS-based Component Technology of Water Resources Information Application[J]. Computer and Modernization,2006（5）,129:43-45.
(张晶晶,陈金水.水信息运用的组件式 WebGIS 技术[J].计算机与现代化.2006（5）,129:43-45.)

[4] Yuebo Xie.Water Information Technology[M]. China WaterPower Press,2009.
(谢悦波.水信息技术[M],北京：中国水利水电出版社,2009.)

[5] Yueming Zhu, Yishan Pan, Keming Sun. Discussion of information definition[J]. Journal of Liaoning Technical University(Social Science Edition), 2003（5）,5（3）: 4-6.
(朱月明,潘一山,孙可明.关于信息定义的讨论[J],辽宁工程技术大学学报（社会科学版）,2003（5）,5（3）：4-6)

[6] Yixin Zhong. Intelligence Oriented Comprehensive Information Theory[J]. Journal of Beijing University of Posts and Telecommunications, 1998（12）,21（4）:1-6.
(钟义信.面向智能研究的全信息理论[J],北京邮电大学学报,1998（12）,21（4）:1-6.)

[7] Ningning Jing, Junyu Cheng. Data,Information,Knowledge and Wisdom[J]. Information Science, 2005（12）,23（12）:1786-1790.
(荆宁宁,程俊瑜.数据、信息、知识与智慧[J].情报科学,2005（12）,23（12）：1786-1790.)

[8] Norbert Wiener. Cybernetics or contol and communication in the machine(2nd Edition)[M]. Translated by Hao Jiren.Beijing: Science Press, 1985.133.
(维纳.控制论——或关于在动物和机器中控制和通讯的科学（第2版）[M].郝季仁译.北京：科学出版社,1985.133.)

[9] Lu Jiang. Discussion on the Definition of Information[J]. Journal of Systemic Dialectics, 2004（4）,12（2）:28-30.
(姜璐.信息定义的探讨[J].系统辨证学学报.2004（4）,12（2）：28-30)

[10] Zhihong Sun. The Nature of Information and Information conservation problem[J].Journal of Hefei University of Technolog(Social Sciences), 1987(6):48-53
(孙志鸿.信息的本质和信息守恒问题[J].合肥工业大学学报（社科版）,1987(6):48-53.)

[11] C.E.Shannon. Mathematical Theory of Communication[J]. BSTJ.1948,27.

[12] Zhao Xiuheng, Chen Rui. The Expansion of Information Concept[A]. 2009 Fifth International Joint Conference on INC, IMS and IDC,1215-1218.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

The Application of Ultra-Efficient DEA Crossover Efficiency Model on Power Supply Enterprise Performance Evaluation

Sun Wei
School of Business Administration.North China Electric Power University
Baoding 071003, China
E-MAIL:jgsunwei123@163.com

Qin Huifang
School of Business Administration.North China Electric Power University
Baoding 071003, China
E-MAIL:whffzw@163.com

Du Qiushi
Electric Power Research Institute of Jilin Electric Power Co.Ltd
Changchun 130061, China
E-MAIL:ncepudqs@126.com

Abstract*---Data envelope analysis method has a good application to the enterprise relative efficiency evaluation. The paper aiming at current situation of Chinese power supply enterprise performance evaluation, established a comprehensive performance evaluation index system. Build an improved ultra-efficient DEA crossover efficiency evaluation model , and adopt a certain actual data of eight provinces to conduct empirical analysis. The calculated results show that ultra-efficient DEA crossover efficiency evaluation model has higher accuracy in application, and the sorting is more reasonable. The application of the model provide important reference for enterprises to improve operation and management, and further improve the competitive power of Chinese power supply enterprises.

Keywords*-ultra-efficient; DEA; crossover efficiency; performance appraisal

I.Introduction

Electric power industry is an important basic industries of national economy, and occupies an important position in people's life and social production. With the establishment of the market economy, optimization of social resource arrangement suddenly format the competitive pressure of power enterprise. Facing the challenge and pressure, the paper establishes performance evaluation index system based on the characteristics of power industry, which is suitable to electric power industry related system and environment. The effective evaluation of the enterprise operating performance has a vital significance to the enterprise development. Performance reflects the achievements or effect that people engaged in a certain activities. Enterprise performance evaluation usually means to possess, utilize, manage and configurate economic resources for the enterprise, namely: through evaluation to the enterprise operating results and operator performance evaluation, not only make the owner can decide enterprise strategy of further development, and inspect the condition of the contract fulfillment, business operators and other stakeholders can also make effective decisions, guide the enterprises to improve management and improve the economic benefit levels according to enterprise performance evaluation results.

Data Envelopment Analysis (DEA) is a kind of efficiency evaluation method developed from the relative efficiency concept, which put forward by the famous operational research scholars A. Charnes and W.W. Cooper .etc., The model especially can effectively handle the evaluation problem which contains a variety of input and output index. This paper uses ultra-efficient DEA crossover efficiency evaluation model, carries out the performance evaluation to the power supply enterprises of eight provinces, such as Yunnan province, and achieved satisfactory results in practical applications.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.89

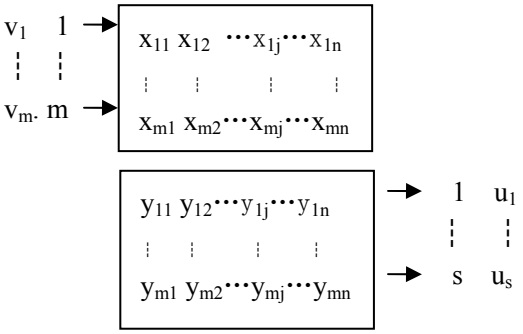
Eight provinces of yunnan power supply enterprises performance evaluation, and in practical applications achieved satisfactory results.

Ⅱ. Model introduction

A. CCR model

DEA is an evaluation method refers to the relative efficiency concept as basis, uses convex analysis and linear programming as the tool .It generalized the engineering efficiency concept which based on single input and single output to the validity evaluation which can deal with the problem related to the similar decision units which has multiple inputs and multiple outputs.

Assuming there are "n"individual data of the similar department by evaluation,which was called the decision unit DMU,every decision unit has "m" kinds of inputs and "s" kinds of outputs.Among them "x_{ij} "means the inputs which was put into no."i" by the DMU of no."j", $x_{ij}>0$; y_{rj} means the outputs which was put out to no."r" by the DMU of no."j".(the "r" means rights).i =1，....，m；j =1，....，n；r =1.....s. x_{ij},y_{rj} are the data which was known,v_i , u_r are variables.



2-1 DMU inputs and outputs

Do efficiency evaluation to the no."j_0" （1□j_0□0） decision unit, use weight coefficient "v" and "u" as variable vector, use the efficiency index of no."j_0" decision unit as the goal, take all the efficiency index of decision units (also included the first j_0 decision units) as constraint, namely:

hj ≦ 1, j = 1,... , n

After the model was constructed, using Charnes - Cooper change can make it converted into an equivalent linear programming problem.

$$(P_{C^2R})\begin{cases}\max h_{j_0}=\mu^{\mathrm{T}}y_0\\ s.t.\omega^{\mathrm{T}}x_j-\mu^{\mathrm{T}}y_j\geq 0\\ \omega^{\mathrm{T}}x_0=1\\ \omega\geq 0,\mu\geq 0\end{cases}\qquad (2\text{-}1)$$

Judge theorem:①If the optimal of linear programming （P_{C^2R}） is $h_{j_0}^{*}=1$,we define that the decision making unit j_0 is weak DEA effective.②If linear programming （（P_{C^2R}））exists optimal solution ω^{*},μ^{*}, satisfying $\omega^{*}>0,\mu^{*}>0$, and the optimal value $h_{j_0}^{*}=1$.We define the decision making unit j_0 is DEA effective.

B. DEA crossover efficiency model

In the above model, each decision unit choose the weights which can make its efficiency value reach the best, the advantage is objectivity, namely, it does not need any preference information; Its limitations are, when quote this method to evaluate, may also get the result that many other decisions effective, and will not sort the decision unit well. Meanwhile, since each decision unit based on its own ego thought evaluation, so every decision is not comparable between units. In order to solve this problem, the concept of efficiency was introduced by the professors such as Sexton in 1986. The so-called crossover efficiency evaluation, is to get "n" group optimal weights through solving "n" group linear programming, so as to evaluate the efficiency of each decision unit "n" times, like the below formula shows:

$$h_{kj} = \frac{\sum_{r=1}^{s} u_{rk} y_{rj}}{\sum_{i=1}^{m} v_{ik} x_{ij}} k, j = 1, ..., n,$$

Thus get cross efficiency matrix:

$$H = \begin{bmatrix} h_{11} & h_{12} & ... & h_{1n} \\ h_{21} & h_{22} & ... & h_{2n} \\ & ... & ... & \\ h_{n1} & h_{n2} & ... & h_{nn} \end{bmatrix}$$

Among them, the main diagonal elements h_{kk} $(k = 1, 2, ..., n)$ is self-evaluation, others which is not diagonal elements $h_{kj}(j \neq k)$ made for cross evaluation value.

Crossover efficiency evaluation reflects the theme that mix self-assessment and others-evaluation together: firstly, decision unit "k" evaluate itself stood in its own point of view, maximize its efficiency value, namely the self-evaluation; Secondly, other n-1 individual decision units stood in their view separately evaluate unit "k", namely others-evaluation decision .

Based on the crossover efficiency evaluation, the final evaluation makes units "j" expressed as: $\bar{\theta}_j = \frac{\sum_{i=1}^{n} h_{ij}}{n}$, $\bar{\theta}_j$ is rated to average crossover efficiency value, we can sort all the decision making units according to the size of $\bar{\theta}_j$ .Although crossover efficiency method considering the comprehensive efficiency of input-output among all of the decision making units. But there are some limitations:① The situation that several units of $\bar{\theta}_j = 1$ may appear, this makes the decision makers can't judge the quality relationship of these effective units, sorting is not clear.②During the crossover evaluation, it is likely to occur the situation that much group weight makes decision unit optimal, select different weights may make the decision unit appear inconsistent case during the sort.

C. DEA model based on ultra-efficiency evaluation model of crossover efficiency

According to the situation that there are more than one effective solution $\bar{\theta}_j = 1$ in the DEA model, and the paper use ultra-efficient model to make improvement when using DEA model to calculate the diagonal elements cross effectiveness matrix. Following the below formula shows:

$$\max h_{kk} = \frac{u^{\mathrm{T}} y_k}{v^{\mathrm{T}} x_k}$$

$$s.t. \begin{cases} h_{kj} = \frac{u^{\mathrm{T}} y_j}{v^{\mathrm{T}} x_j} \leq 1, j = 1, 2, ..., n; j \neq k \\ v^{\mathrm{T}} x_k = 1 \\ u \geq 0, v \geq 0 \end{cases}$$

Through the above formula, we can calculate and get each DMU ultra-efficiency evaluating value,and get the optimal weight calculated, and other elements in cross effectiveness matrix of value. In the formula, remove the $h_{kk} \leq 1$ restriction, thus efficiency evaluation value of DEA effective units are greater than 1, thus can reduce the situation $\bar{\theta}_j$ = 1, in the structure crossover matrix, and the other none effective decision units only take small changes.

Ⅲ. Empirical analysis

A.Establishment of index system

For using ultra-efficient DEA crossover efficiency evaluation method to make a comprehensive

investigation to the power supply enterprise performance evaluation under the market economy condition, the paper reference the 2006 China economic statistical yearbook of first national economic census data of the main data communique, quantitative index intercepted the related economic data of eight provinces power supply enterprise such as Yunnan province. The statistics of qualitative indexes by using Delphi method which was scoring by experts. Adopted the number of staff and workers, labor remuneration, total assets, main business income, total profit and power supply reliability, social satisfaction of power supply users indexes, as an evaluation index of power supply enterprise performance evaluation. These indicators covers three evaluation aspects such as the power supply reliability, the economic benefits of power supply enterprise and social responsibility of the enterprise, the evaluation system is comprehensive. Specific index and data are shown below:

TABLE 1. INDEX SYSTEM

The power supply enterprise Performance evaluation index	Economic benefits	Number of staff and workers
		Labor remuneration
		Total assets
		Main business income
		Total profit
	Power supply reliability	User power supply reliability
		The average power outage hours of users
	Corporate social responsibility	Power satisfaction of users

The provinces in proper sequence from top to bottom respectively is: Yunnan, Hebei, Xinjiang, Hubei, Jiangxi, Tianjin, Guangzhou, Anhui.

Input index: Number of staff and workers (x_1), Total assets (x_2) (billion), Labour remuneration (x_3)(billion).

Output index: Main business income (y_1)(billion), Total profit (y_2)(billion), Average power outage time (y_3)(user household/hours), Social satisfaction of power supply users (y_4), power supply reliability (y_5).

TABLE 2 THE NATIONAL EIGHT PROVINCES AND MUNICIPALITIES POWER SUPPLY COMPANY PERFORMANCE EVALUATION DATA

p	x_1	x_2	x_3	y_1	y_2	y_3	y_4	y_5
Y	68671	762.64	29.7	245.1	18.82	98.242	3	98.88
H	16590	1340.97	70.7	89	8.9	59.34	29.387	6	99.66
X	8209	48.4194	19.5	10.16	0.457	29.778	6	99.66
H	13940	2477.5	35.6	626.1	76.94	22.46	7	99.74
J	16210	2454.13	68.6	1241.	16	46.16	18.309	9	99.79
T	34600	559.12	24.9	288.8	16.78	19.68	8	99.77
G	26352	178.99	6.0168	58.47	0.57	39.443	5	99.551
A	22600	997	41.37	556.2	4.1	42.115	4	99.521

B.Empirical analysis results

The paper using EMS software which is exclusive to calculate and analysis DEA data, and calculate the power supply company performance evaluation index of the above eight provinces and municipalities respectively ,the calculated results as follows.

DMU	Crossover efficiency rank	Ultra-efficient crossover efficiency model evaluation value	Ultra-efficiency model crossover efficiency rankings
DMU1	7	0.4066	7
DMU2	6	0.573	6
DMU3	1	1.6967	1
DMU4	8	0.2774	8
DMU5	5	0.6859	5
DMU6	4	0.7773	4
DMU7	1	1.2486	2
DMU8	1	1.1177	3

C. Result analysis

From the calculation results we can see, the proposed ultra-efficient DEA model crossover efficiency evaluation method and super-efficiency evaluation method, the performance evaluation efficiency of power supply enterprise:DMU8 DMU3 and DMU7 is better. And the sort is more accurate and scientific, the proposed method is feasible. The proposed DEA evaluation method is considering self-evaluation and others-evaluation decision making units, and it is a DEA evaluation method which can evaluate the situation effectively and sort more accurate.

Ⅳ. Conclusions

The model proposed the evaluation thought which is nearer with social evaluation activity ,it effectively mixed self-evaluation and others-evaluation together. Using ultra-efficient DEA model sort all the decision making units, especially carry out the more precise sorting on the situation of evaluation efficiency that there are more than one effective value equals 1. And the model adopted the data has good tolerance, and can simultaneously make analysis on multiple input sand output of the qualitative and quantitative index for evaluating, and this method has better applicability to power supply enterprise performance evaluation problem .

REFERENCES

[1] Charnes A, Cooper W W, Rhodes E. Measuring the efficiency of decision making units [J] . Eurappear Journal of Operational Research, 1978, 2: 429～444.

[2] R H. Measuring efficiency: an assessment of data envelopment analysis [C] . SanFrancisco:Jossey-Bass, 2008, 32: 73～105.

[3] Anderson T R, etal. The fixed weighting nature of a cross-evaluation model[J]. Journal of Productivity Analysis, 2009, 18(1) : 249～255.

[4] Zhao Shenghan, DEA theory, method and application Beijing science press. Beijing. 2010.

[5] Quanling Wei, Evaluation relative effectiveness of DEA method Renmin university of China publishing house, Beijing.2009.

[6] Anderson T R, etal. The fixed weighting nature of a cross-evaluation model [J]. Journal of Productivity Analysis, 2002, 18(1) : 249～255.

[7] Hu Tongze，Huang Lijun．The DEA method based on super efficiency state-owned Industrial enterprises .Industry science and technology competitiveness measure [J]. Scientific progress and counter measures. 2007.24(5):56–58.

Constraint-aware Correctness Analyzing of Composite Web Service Based on Open Petri Net

Xin Gao
Department of Information and Computing Science
Anhui University of Science and Technology
Huainan, China
e-mail:gxnew@qq.com

Xianwen Fang
Department of Information and Computing Science
Anhui University of Science and Technology
Huainan, China
e-mail:xwenfang@tom.com

Zhicai Xu
Department of Mathematics
Chuzhou University
Chuzhou, China
e-mail:lxw7710@tom.com

***Abstract*—Under the environment constraint, the correctness analyzing of the composite Web service has been a research hot spot. Firstly, the paper proposes the modeling methods of Web services and Web environments based on open Petri net (OPN). Then, in order to analyze the correctness of the composite Web service under the environment constraint, several determining algorithms about correctness of composite web service are presented. Finally, a concrete example is given out. Theoretical analysis and experimental results indicate that the methods can analyze the correctness of constraint composite service effectively.**

***Keywords*--Open Petri net; Environment constraint; Composite Web service; Correctness**

I. INTRODUCTION

In practical application, the correctness of composite Web service is not only constrained by other Web services, but also constrained by the Web environment. Because the Web environment has special features, it is difficult to model and analyze by traditional methods and the research on the correctness of composite Web service under environment constraint demand new modeling tools and analysis methods.

In the literatures [1, 2] analyzed the influence on the mobile network service discovery by electromagnetism and bandwidth environment and proposed the perspective that the environment influences the correctness of services. In [3] the authors established the formalized model of the environment and composite Web service on the basis of the service tree theories, and presented a decision method about the model soundness.

The paper puts forward the concepts of Web service domain and Web environment domain, and uses OPN to model Web service domain and Web environment domain. Then, a sufficient condition for the composition of Web service domain and Web environment domain is given and some correctness decision algorithms about the composition are provided. Finally, the concrete example verifies the correctness of the algorithms.

II. PRELIMINARIES

At present, researchers define service composition in different perspectives. The Web service composition is a process that in order to add services value, some single services with simple function are composed. Usually, Web environment includes three parts: user environment, computing environment and physical environment.

With the development of Web service research and the actual application requirements, Web environment become a very important factor in Web service research field. Some researches paid attention to the relationship of different Web services, but ignored Web environment constraint. Web service environment does not provide true service, but the Web environment influences Web service composition, operation and correctness.

Definition 1: A Web service domain (SD) is a 4-tuple $SD = (S_I, S_o, S_C, C_I)$, where S_I is a service input, So is a service output, Sc is a service component and C_I is a communication interface.

Definition 2: A Web environment domain(ED) is a 4-tuple $ED = (E_I, E_o, E_C, C_I)$, where E_I is an environment input, Eo is a environment output, Ec is a environment component and C_I is a communication interface.

Definition 3[4, 5]: A Petri Net is a $3\text{-tuple} N = (S, T; F)$, where

(1) $S \cup T \neq \phi$

(2) $S \cap T = \phi$

(3) $F \subseteq ((S \times T) \cup (T \times S))$

(4) $dom(F) \cup cod(F) = S \cup T$

$dom(F) = \{x \in S \cup T \mid \exists y \in S \cup T : (x, y) \in F\}$

$cod(F) = \{x \in S \cup T \mid \exists y \in S \cup T : (y, x) \in F\}$

Definition 4: An Open Petri Net is a $7\text{-tuple} OPN = (S, T; F)$ [6], where

(1) $P \cup I \cup O, T, F)$ is a Petri net;

978-0-7695-4415-1 2011
U.S. Government Work Not Protected by U.S. Copyright
DOI 10.1109/DCABES.2011.30

(2) P is a set of internal places;
(3) T is a set of transitions;
(4) F is a set of flow relations;
(5) I is a set of input places, and $^{\bullet}I = \phi$;
(6) O is a set of output places, and $^{\bullet}O = \phi$;
(7) i is the initial marking;
(8). f is the final marking, and f is a deadlock.

We call the set $I \cup O$ is the interface places of the OPN. Two OPNs M and N are disjoint, if $P_N, P_M, I_N, I_M O_M O_M, T_N, T_M$ are pairwise disjoint.

Definition 5: Let $SD = (S_I, S_o, S_C, C_I)$ be a Web service domain, $ED = (E_I, E_o, E_C, C_I)$ be a Web environment domain and $M = (P, I, O, T, F, i, f)$ be an open Petri Net. If $S_I \subseteq \mathrm{i}$, $S_o \subseteq f$, $S_C \subseteq \mathrm{P}$ and $C_I \subseteq I \cup O$, SD should be mapped to an OPN and if $E_I \subseteq \mathrm{i}$, $E_o \subseteq f$, $E_C \subseteq \mathrm{P}$ and $E_I \subseteq I \cup O$,ED should be mapped to an OPN.

Through interface places, Web environment communicates with Web service, thus an interrelated whole is formed structurally and behaviorally. Because of the communication, Web environment inevitably constrains Web service behavior which is reflected by the correctness of composition.

III. Correctness Analysis of Composite Web Service under Environment Constraint Based on Open Petri Net

Correctness analysis of composite Web service under environment constraint mainly has two styles :(1). If a single Web environment and a single Web service satisfy soundness, correctness can be determined by the soundness of composition. (2). if a single Web environment and a single Web service do not satisfy soundness or we do not need to consider their soundness, correctness can be determined whether the composition has deadlock.

Whether Web environment and Web services can compose is the premise that Web environment constrains Web services, so we should give the condition they can compose.

Definition 6[7]: $\mathrm{A} \oplus \mathrm{B}$, Let A and B be two OPNs, their composition is an OPN $\mathrm{A} \oplus \mathrm{B} = (\mathrm{P}, \mathrm{I}, \mathrm{O}, \mathrm{T}, \mathrm{F}, \mathrm{i}, \mathrm{f})$ defined by:

(1).$P = P_A \cup P_B \cup (I_A \cap O_B) \cup (I_B \cap O_A)$
(2).$I = (I_A \setminus O_B) \cup (I_B \setminus O_A)$
(3).$O = (O_A \setminus I_B) \cup (O_B \setminus I_A)$
(4).$T = T_A \cup T_B$
(5).$F = F_A \cup F_B$
(6).$i = i_A \cup i_B$
(7).$f = f_A \cup f_B$

Two OPNs A and B are composable if and only if $(P_A \cup P_A \cup O_A \cup T_A) \cap (P_B \cup P_B \cup O_B \cup T_B) = (I_A \cap O_B) \cup (O_A \cap I_B)$

A. soundness decision algorithm for Web environment Composite Web service

Definition 7[8]: Let N be an OPN. The skeleton of N is defined as the Petri net $S(N) = (P_N, T_N, F)$ with $F = F_N \cap ((P_N \times T_N) \cup (T_N \times P_N))$ .We use $R(N)$ as a shorthand notation for $R(S(N), i_N)$. N is called sound if for any marking $\mathrm{m} \in R(N)$ we have $S(N) : \mathrm{m} \xrightarrow{*} f_N$.

In many cases, multiple Web service domains share a Web environment domain. As shown in Figure 1, we have three components A, B, and C. The composition of A with B is sound, as well as the composition of B with C. However, the composition of the three is not sound.

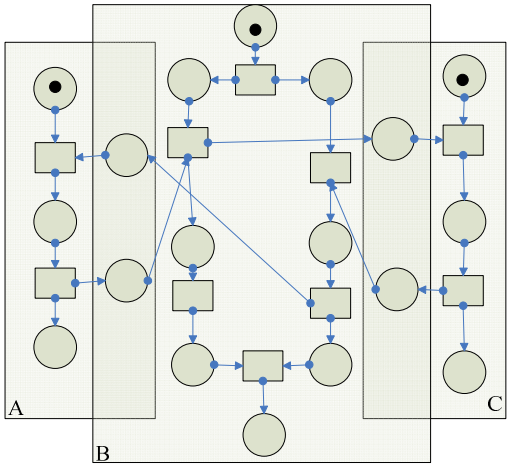


Figure 1. $\mathrm{A} \oplus \mathrm{B}$ and $\mathrm{B} \oplus \mathrm{C}$ satisfy soundness, $\mathrm{A} \oplus \mathrm{B} \oplus \mathrm{C}$ does not satisfy soundness

Definition 7[8]: $\Omega_{A,B}$ .A and B are two composable OPNs, $\Omega_{A,B}$ holds if and only if $\forall m \in R(A \oplus B), \sigma \in (T_A)^* : (A : m \mid P_A \xrightarrow{\sigma} f) \Rightarrow (\exists \tilde{\sigma} \in (T_A \cup T_B)^* : (A \oplus B : m \xrightarrow{\tilde{\sigma}} f_A + f_B) \wedge \tilde{\sigma} \mid T_A = \sigma$

A and B are two composable OPNs, if A satisfies soundness and $\Omega_{A,B}$ holds, $\mathrm{A} \oplus \mathrm{B}$ satisfies soundness. Let A, B, C are three pairwise composable OPNs such that A and C have disjiont interface places, A, B, C pairwise composable, if $\mathrm{A} \oplus \mathrm{B}$ satisfies soundness, and $\Omega_{B,C}$ holds, $\mathrm{A} \oplus \mathrm{B} \oplus \mathrm{C}$ also satisfies soundness.

Algorithm 1: decision for Three Web environment domain/Web service domain composition

Input: Web service domain $SD_1 = (S_I, S_o, S_C, C_I)$, Web environment domain $ED = (E_I, E_o, E_C, C_I)$, and Web service domain $SD_2 = (S_I, S_o, S_C, C_I)$.

Output: decision outcome

(1). By definition 5, put $SD_1 = (S_I, S_o, S_C, C_I)$, $ED = (E_I, E_o, E_C, C_I)$, and $SD_2 = (S_I, S_o, S_C, C_I)$ map into OPN: Web_{opn} A, $WebEnvir_{opn}$ B and Web_{opn} C, then execute step (2).

(2). By definition 6, if $(P_A \cup P_A \cup O_A \cup T_A) \cap (P_B \cup P_B \cup O_B \cup T_B) = (I_A \cap O_B) \cup (O_A \cap I_B)$, and $(P_B \cup P_B \cup O_B \cup T_B) \cap (P_C \cup P_C \cup O_C \cup T_C) = (I_B \cap O_C) \cup (O_B \cap I_C)$, A, B, C are composable and the composition of A,B,C is also an OPN: $A \oplus B \oplus C$, then execute step(3). Otherwise algorithm terminates and returns the result that they do not satisfy soundness.

(3). If $A \oplus B$ is sound, execute step (4). Otherwise algorithm terminates and returns the result that they do not satisfy soundness.

(4). By definition 9, if $\Omega_{B,C}$ holds, execute step (5). Otherwise algorithm terminates and returns the result that they do not satisfy soundness.

(5). If $A \oplus B \oplus C$ is sound, algorithm terminates and returns the result that they satisfy soundness.

In actual application, the composition of Web service domain and Web environment domain has large number and complex ways, therefore it is very necessary to determine soundness of multiple Web environment domains and Web service domains composed.

Algorithm 2: soundness decision algorithm for arbitrary Web environment domain/Web service domain composition

Input: D1,……Dn are Web environment domains/Web service domains

Output: decision outcome

(1). By definition 5,put D1,……Dn map into OPN: $A1, \ldots\ldots An$, then execute step (2).

(2). By definition 6,if $(P_i \cup P_i \cup O_i \cup T_i) \cap (P_{i(c)} \cup P_{i(c)} \cup O_{i(c)} \cup T_{i(c)}) = (I_i \cap O_i) \cup (O_{i(c)} \cap I_{i(c)})$ holds, $A1, \ldots\ldots An$ are composable, then execute step (3). Otherwise algorithm terminates and returns the result that they do not satisfy soundness.

(3). $A1, \ldots\ldots An$ is a domain tree with the parent A1, if and only if let $c: \{2, \ldots\ldots n\} \to \{1, \ldots\ldots n-1\}$ be such that:

$\forall i \in \{2, \ldots\ldots n\}: c(i) < i$

$\forall 1 \le i < j \le n : i = c(j) \Rightarrow I_{Ai} \cap O_{Aj} \neq \theta \vee O_{Ai} \cap I_{Aj} \neq \theta$

$\forall 1 \le i < j \le n : i = c(j) \Rightarrow$

$I_{Ai} \cap O_{Aj} \neq \theta \vee O_{Ai} \cap I_{Aj} \neq \theta$

, then execute step (4).

(4). If $A1, \ldots\ldots An$ is a domain tree with the parent A1, A1 is sound and $\Omega_{Ai, Ac(i)}$ holds , $A1 \oplus \ldots\ldots \oplus An$ is also sound, then execute step (5). Otherwise algorithm terminates and returns the result that they do not satisfy soundness.

(5). Algorithm terminates and returns the result that composition of $A1, \ldots\ldots An$ satisfies soundness.

B. Decision algorithm for whether arbitrary Web environment domains/Web service domains composition has deadlock

Above, we consider the case that a single Web environment and a single Web service are sound. However, when a single Web environment or a single Web service is not sound or they do not need soundness, the correctness can be determined by means of whether the composition has deadlock.

Algorithm 3**:** Algorithm for deciding whether Web environment domain/Web service domain composition has deadlock

Input: Web service domain $SD = (S_I, S_o, S_C, C_I)$, and Web environment domain $ED = (E_I, E_o, E_C, C_I)$.

Output: decision outcome

(1). By definition 5, put $SD = (S_I, S_o, S_C, C_I)$, and $ED = (E_I, E_o, E_C, C_I)$ map into OPNs: Web_{opn} A and $WebEnvir_{opn}$ B, then execute step (2).

(2). By definition 6, if $(P_A \cup P_A \cup O_A \cup T_A) \cap (P_B \cup P_B \cup O_B \cup T_B) = (I_A \cap O_B) \cup (O_A \cap I_B)$ holds, A, B are composable and the composition of A, B is also an OPN: $A \oplus B$, then execute step(3)；Otherwise algorithm terminates and returns the result that they satisfy correctness.

(3). Find out the incidence matrix of $A \oplus B$, $M_\circ$ is its initial status.

(4). M_X is a required reachable status and $M_X \in R(M\circ)$, if a nonnegative vector X exists such that $M_X = M_\circ + (A_{(A \oplus B)})^T X$ holds. Then execute step(5)；Otherwise algorithm terminates and returns the result that they satisfy correctness.

(5). If a transition sequence exists, such that M_X is

reachable from M_0, then execute step (6); otherwise algorithm terminates and returns the result that they satisfy correctness.

(6). If the place includes M_X after the composition of Web_{opn} A and $WebEnvir_{opn}$, it is not deadlock. Algorithm terminates, and returns the result that the composition of Web_{opn} A and $WebEnvir_{opn}$ satisfies correctness.

IV. EXAMPLE ANALYSIS

Above, we do research on some decision algorithms about the soundness and deadlock of Web service domain/ Web environment domain composition. In this chapter we use an example to analyze the algorithms.

Web provides HD VOD service which is composed of channel, item and part and HD VOD service needs to occupy abundant bandwidth resource, so in order to obtain HD VOD service with soundness, we want bandwidth environment to constrain HD VOD service. Meanwhile, Web also provides news voice live broadcast service which is composed of channel and language. Because live broadcast service has real-time property, we need time environment to constrain live broadcast service so that live broadcast service with soundness can be obtained.

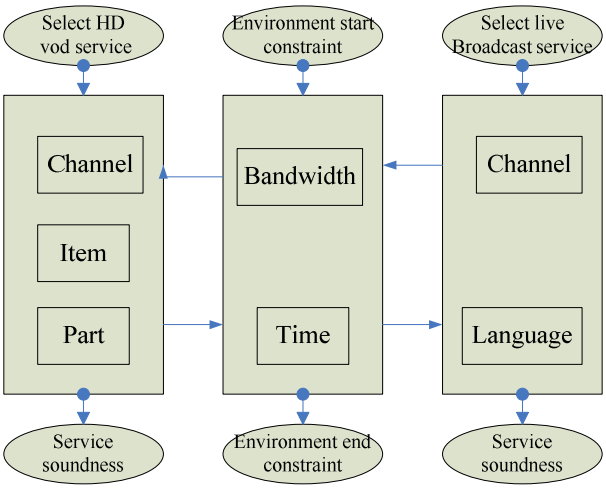


Figure 2. Composite Web service under environment constraint

First, we model the HD VOD service , voice live broadcast service and the environment domain by OPN, and map them into OPNs: N_1,N_2,N_3 where i_1,i_2 and i_3 are initial places, respectively including one initial marking; f_1,f_2 and f_3 are final places; p_1,p_2,p_3,p_4,p_5 and p_6 are internal places; o_1,o_2 and o_3 are output interfaces ; i_1,i_2 and i_3 are input interfaces. Note that o_1 equals i_2 in N_2, i_1 equals o_2 in N_2, o_2 equals i_3 in N_3, and i_2 equals o_3 in N_3.

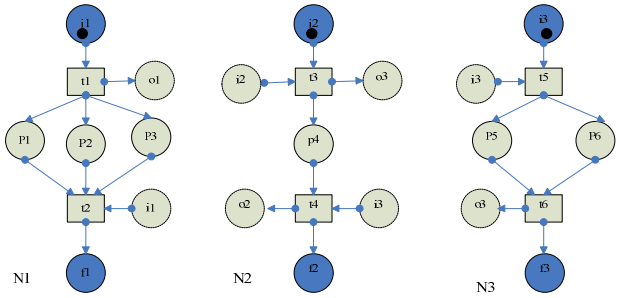


Figure 3. HD vod service domain N_1, Web environment domain N_2 and voice live broadcast service domain N_3

Then, we compose N_1 and N_2. $(P_A \cup P_A \cup O_A \cup T_A) \cap (P_B \cup P_B \cup O_B \cup T_B)$ $= (I_A \cap O_B) \cup (O_A \cap I_B)$ Holds, so N_1 and N_2 could compose. N_1 is sound and Ω_{N_1, N_2} holds, so $N_1 \oplus N_2$ is sound. Next, we compose N_3 and $N_1 \oplus N_2$, because $N_1 \oplus N_2$ is sound and Ω_{N_2, N_3} holds, $N_1 \oplus N_2 \oplus N_3$ is also sound by decision algorithm 1.

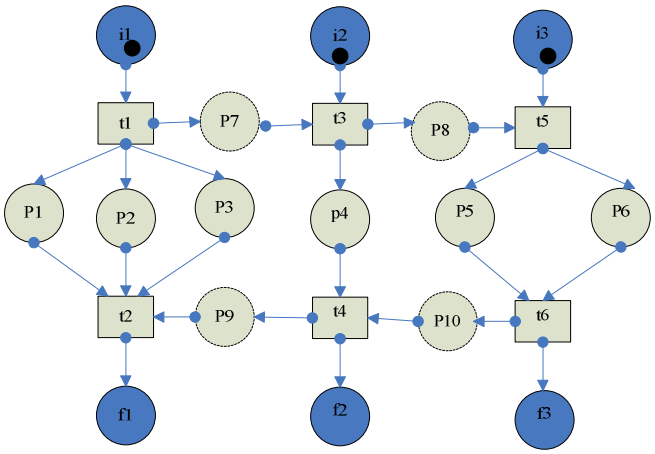


Figure 4. $N_1 \oplus N_2 \oplus N_3$

Finally, we confirm the soundness of $N_1 \oplus N_2 \oplus N_3$ by reachable marking:

$M:(i1,i2,i3,p1,p2,p3,p4,p5,p6,p7,p8,p9,p10,$ $f1,f2,f3)$

$M0:(1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0)$

$M1:(0,1,1,1,1,1,1,0,0,0,1,0,0,0,0,0)$

$M2:(0,0,1,1,1,1,1,1,0,0,0,1,0,0,0,0)$

$M3:(0,0,0,1,1,1,1,1,1,1,0,0,0,0,0,0)$

$M4:(0,0,0,1,1,1,1,1,0,0,0,0,0,1,0,0,1)$

$M5:(0,0,0,1,1,1,1,0,0,0,0,0,1,0,0,1,1)$

$M6:(0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1)$

As shown in the reachable sequences, the marking in the initial places i_1,i_2 and i_3 can reach the final places f_1,f_2 and f_3 , and the other places in $N_1 \oplus N_2 \oplus N_3$ have no markings, so $N_1 \oplus N_2 \oplus N_3$ is sound. Last, we can find out a nonnegative vector and reachable sequences exist, there is no deadlock in $N_1 \oplus N_2 \oplus N_3$. It proves that $N_1 \oplus N_2 \oplus N_3$ satisfied the two situations of correctness.

V. CONCLUSION

The paper models Web environment domain and Web service domain by OPN and analyze their dynamic operation. The correctness of Composite Web service under environment constraint include two situations :(1). Single Web environment/Web service are satisfy soundness, use the whole satisfy soundness analyze correctness ;(2). Single Web environment/Web service are not satisfy consider soundness, use the whole specific places not exist deadlock status analyze correctness. Finally, we validate the correctness of these algorithms through a concrete. The future research emphases on transplanting this decision algorithms to commercial web protocol and language.

ACKNOWLEDGMENT

We would like to thank the support of the National Natural Science Foundation of China under Grant No. 60873144, No.60973050 and No.61073102, the Natural Science Foundation of Educational Government of Anhui Province of China (KJ2011A086, KJ2010B310and KJ2009A50).

REFERENCES

[1] Ahlem Ben Hassine. A Constraint-Based Approach to Horizontal Composite Web service. ISWC 2006:LNCS 4273,2006.130–143.

[2] O.Ratsimor,V.Korolev, A.Joshi.Finin.Agents2Go:An Infrastructure for Location-Dependent Service Discovery in the Mobile Electronic Commerce Environment, ACM: ACM Mobile Commerce Workshop,2007.

[3] W.M.P.vander Aalst, K.M.van Hee, P.Massuthe, N.Sidorova, and J.M.E.M. vander Werf. Service Trees. Technical Report CSR: Technische Universiteit Eindhoven,2009.

[4] Wu Zhe-hui. Petri net theroy. BeiJing：China Machine Press, 2006.

[5] Yuan Chong-yi. Petri net principle and application. BeiJing : Electronic Industry Press, 2005.

[6] W.M.P.vander Aalst, N.Lohmann, P.Massuthe, C.Stahl, and K.Wolf. From Public Views to Private Views: Correctness-by-Design for Services. WS-FM 2007:volume 4937 of Lecture Notes in Computer Science,2007.139–153.

[7] WilM.P.vander Aalst, Kees M.van Hee, Peter Massuthe, Natalia Sidorova and Jan Martijn vander Werf. Compositional Service Trees.In WS-FM 2009: Volume 5606 of Lecture Notes in Computer Science, 2009.

[8] K.M.van Hee, H.M.W.Verbeek, C.Stahl, and N.Sidorova. A framework for linking and pricing no-cure-no-pay services. Transactions on Petri Nets and Other Models of Concurrency: Springer,2009.

Construction of Energy-Aware ZigBee Network

Xu Haiyan
School of Computer Science and Technology
Huaiyin Normal University
Huaian, 223300, Jiangsu, China
e-mail: catmouse99@163.com

***Abstract*—Generally, the Cluster-Tree algorithm and the AODVjr algorithm are used for routing strategies at ZigBee network layer. The former is simple, but its efficiency is low. The latter one can complement the cluster-tree routing's shortcomings. However, the energy consumption and network life are not considered adequately for energy-limited networks. This paper proposes an improved measure, which can deduce the node energy consumption and prolong the network life as far as possible, in order to avoid the premature segmentation of networks. Experiments show that this algorithm can efficiently improve the node's average life in energy limited networks, without reducing the network performance.**

Keywords-ZigBee; Routing; Energy-aware

I. INTRODUCTION

Generally speaking, the primary objective of wireless networks is to improve the quality of service (QOS) and fair efficient use of network bandwidth, but if the energy of ZigBee nodes is relatively sensitive, and if the planning is unreasonable, the premature failure of some key routing nodes leads to the emergence of network partition, with the result of seriously affecting the network life. Besides through optimizing design and timing dormancy, most energy-saving mechanisms are by improving the efficiency of routing and adopting the data-fusion technology to improve energy efficiency and postpone nodes' failure time. Thus the design goal of low cost, low power consumption and high reliability design is reached. ZigBee network in general adopts two routing protocols namely Cluster-Tree Routing (CTR) and Ad Hoc On-demand Distanceing (AODVjr) [1]. It combines the merits of both CTR algorithm and AODVjr algorithm. The formation of TR ZigBee network starts by choosing a router device that desires to be a coordinator. A CTR ZigBee network, with the characteristic of simple routing algorithm, is more suitable for static network and do not need to store the routing table. However, in most cases, the communication between the nodes requires parent nodes to transmit data. The upper nodes have more energy consumption. AODVjr is the improvement of AODV. It simplifies some characteristics of AODV and reduces the routing cost, energy consumption, and application complexity [2, 3].

For mitigating the drawback in CTR and AODVjr protocol we improved the original protocol. This protocol becomes more efficient than the original protocol by employing neighbor-table which is part of the existing ZigBee network specification. Our algorithm improves the original protocol by lightening the burden of the power-weak nodes by using the new neighbor-table, energy degree and the new routing algorithm.

II. ADDRESS ASSIGNMENT MECHANISM OF ZIGBEE NETWORK

ZigBee network generally adopts the cluster-tree hierarchy to organize nodes. The coordinator first creates ZigBee network. It will start broadcasting a beacon containing a PAN identifier to initialize a PAN coordinator. Other nodes will join the network and become the children nodes of these nodes.

ZigBee network has two address modes namely IEEE MAC address and the short address of 16 bits. The IEEE MAC address is a 64-bit address, allocated by IEEE organization, and used as the identifier of the equipment. The 16-bit address is used as network layer address. When a new node joins the ZigBee network, it will be allocated a

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.31

short address from its parent node, also as one of its neighboring nodes [1].

In addition to the coordinator, each node is divided into three types:RN+，RN-，RFD. RN+ and RN- nodes belong to fully functional nodes. RFD nodes only act as leaf nodes and forward data to parent nodes. RN- nodes only use static cluster-tree algorithm. RN+ can pose AODVjr to look for the best route to the destination node, and it can play a routing agent role to help the other nodes to find routes. It deliver data packets to its parent node of its own or forward them to children nodes.

Each node joined node will be assigned to a parent node and get a unique 16-bit address, which is used to communicate between nodes [3]. The parent node assigns an address to its first routing child node, which is the address of its parent node plus 1. The address of the second child routing node is as follows: Address of first routing child $+Cskip(d)$, and so on, where

$$Cskip(d)=\begin{cases}1+c_m\cdot(l_m-d-1), & \mathrm{R_m}=1\\ \dfrac{1+c_m-R_m-c_m\cdot R_m^{L_m-d-1}}{1-R_m}, & R_m\neq 1\end{cases} \quad (1)$$

Parent nodes allocate addresses to end nodes. By using (1), the address of *n*th end node is calculated by its parent node as follows.

$$A_n=A_{\text{parent node}}+Cskip(d)\cdot R_m+n \quad (2)$$

A. Strategy of ZigBee Routing

After the RN- node received a data packet, it can only use Cluster-Tree algorithm processing. The RN+ node has enough storage space and ability to execute AODVjr protocol. If needed, RN+ nodes will start the route discovery of AODVjr to find a shortest path leading to the destination node. If one of the links is interrupted, the RN+ node will launch the local repair process to repair the route [4].

The Cluster-Tree routing algorithm completely relies on the Cluster-Tree structure to forward the data. If the end node wants to send a packet to the other node in the network, it will directly forward the packet to its parent node, which will transmit the packet. If a router node wants to forward packets to the destination node, which is named D as its network address, and the network address and depth of this router node are known as A and D respectively, firstly the router node will judge whether the destination node is its descendant node according to the following expression: A <D<A + Cskip (D - 1).

If the destination node is its descendant nodes, the next hop address is described as follows:

$$N=\begin{cases}D & \text{If the destination node is its child node}\\ A+1+\left\lfloor\dfrac{D-(A+1)}{Cskip(d)}\right\rfloor\times Cskip(d) & \text{Otherwise}\end{cases} \quad (3)$$

Otherwise, the next hop is the parent of the node. Cluster-Tree routing limits addressing flexibility and the data transfer needs a long path, thus the network energy is wasted and the service delay is induced.

B. AODvjr routing algorithms

The AODvjr algorithm is based on the AODV algorithm, through the flooding RREQ packet looking for the best path. It abandons the target serial number and hop of the AODV. Because it is only for the goal node responding to the first arrived RREQ signal, so this strategy is also called the point-to-point strategy [5]. Meanwhile, sending HELLO messages is cancelled in the AODVjr, and the routing is maintained with the goal node regularly sending the KEEP-ALIVE connection information to the source node. Although the AODVjr algorithm is able to find a shortest path, the flooding addressing messages and routing maintenance all need lots of extra expenses. And the shortest path choice often causes to a rapid energy consumption of the node in special position, which will cause to the imbalance of the network energy.

In the initialization stage, the coordinator assigns a unique address to each node according to network address distribution mechanism. In addition to RFD nodes, each node starts with the initialization neighboring-table, which records all the relevant information of neighboring nodes. It will also decide its power degree by comparing the remaining energy with the definition levels, by which the suitable routing strategy is chose when acting as a router. Nodes detect the level of its energy degree periodically, and update its neighboring-table so as to determine whether to act as the role of transmitting RREQ [6, 7].

While transmitting packets, routing and nodes status

must be updated and maintained in real time. If the energy degree of nodes has been turned into Critical level, it will send a message to the source node to restart another routing search.

III. IMPROVEMENTS OF ROUTING STRATEGIES

A. The Consideration of Data Structure

In this paper, the node energy is divided into three levels: enough, low and critical. According to the node types, nodes with enough energy can choose the Cluster-Tree algorithm or AODVjr algorithm to find routing. The low- power node chooses the agent mechanism to send AODVjr flood packets, and the critical node only responds to the node whose destination address is itself [8]. The neighboring-table concludes several items of information such as network address, power type, routing proxy status, the parent node, child node, PAN identifier, MAC address, and link of children. ZigBee devices can confine the size of the neighbor table by choosing the maximum number of neighbor entries in the neighboring-table.

B. The Improvement of Routing Algorithms

Analysis of the AODVjr routing algorithm and Cluster-Tree algorithm, we find that the RN+ node consumes most energy in the network transmission. The RN+ node, is not only in charge of data collection, but also responsible for routing discovery and data forwarding task. It consumes more energy to be a failure node, so the network partition occurs. Some key nodes become failure, bring heavier burden to the surplus nodes, and accelerate the death time of the network [9]. So the burden of RN+ nodes should be reduced, and when the nodes' energy is reduced to be a lower level, the routing is no longer found to balance the energy consumption, and increase the reliability of the network.

C. Route Discovery of RN+ Nodes

When a RN+ node starts a route discovery, and intermediate nodes receive the RREQ packet, the proposed algorithm encompasses the following steps:

Step 1: If it is the destination node, it will give a RREP answer back to the source node.

Step 2: Update power degree. If the power degree reaches the level of Critical, and if it is not the sink node, it will send a RERR message. If the power degree reaches the level of low, it can't act as the routing proxy for other nodes except its parent and descendant nodes.

Step 3: If the destination node is included in its neighbor-table and the node has enough power, it will answer a RREP to the source node.

Step 4: Judge whether the destination node is the parent node or descendant node of the current node. If so, forward the RREQ packet to the representative father node or child node, act as the agent node, thereby reducing the RREQ flooding. Otherwise enter the next step. If it is an RN- node, the cluster of trees routing is used to communicate, and the RN+ node of the parent node will start to look for the routing.

IV. THE SIMULATION ANALYSIS

The OMNeT++ 4 and MF frame node are used in our simulation experiment. The network coverage area is 80m x 80m, the data transmission rate is 250KB, and the packet length is 128bit. Judging from the experimental results, the average death of the node is prolonged by using our algorithm, so the new algorithm proposed in this paper is better than the original AODVjr algorithm and the Cluster-Tree algorithm.

TABLE 1. ENERGY CONSUMPTION OF ORIGINAL AND IMPROVED ALGORITHMS

Time (Sec.)	1	5	10	15
Original algorithm(Joules)	2000	8000	14000	20000
Improved algorithm(Joules)	1900	4000	8000	11000

TABLE 2. FAILURE NODE PERCENTS OF ORIGINAL AND IMPROVED ALGORITHMS

Time(Sec.)	1	5	10	15
Original algorithm(Sec.)	0	5	23	26
Improved algorithm(Sec.)	0	2	8	10

TABLE 3. AVERAGE HOPS OF ORIGINAL AND IMPROVED ALGORITHMS

Number of Sources	5	10	15	20
Average Hops of Original Algorithm	15	12	9	7
Average Hops of Improved algorithm	12	9	7	6

TABLE 4. AVERAGE END-TO-END DELAY OF ORIGINAL AND IMPROVED ALGORITHMS

Number of Sources	5	10	15	20
Average End-to- End delay of Original Algorithm (mSec.)	60	85	93	95
Average End-to- End delay of Improved algorithm (mSec.)	40	55	65	70

V. CONCLUSION

By analyzing the ZigBee routing algorithm in the lack of energy control, an improved measure of the AODVjr and Cluster-Tree algorithms is proposed in order to save energy in this paper. In this new algorithm, efficient use of neighbor list, node energy classification mechanism and new RREQ forwarding mechanism, make the network be able to balance energy consumptions in the whole network. Simulation results show that this algorithm can optimize the energy consumption status in the whole ZigBee network, prolong the network's total life, and also make nodes have a better performance of the failure rate.

REFERENCES

[1] Electronic Publication: ZigBee Alliance, ZigBee Specification, December 1, 2006, Available: www.zigbee.org.

[2] Perkins CE and Royer EM, "The Ad Hoc On-Demand Distance Vector Routing," Proc of the 2nd IEEE Workshop on Mobile Computing Systems and Application, 1999, pp.90-100.

[3] Ran Peng, Sun Maoheng, and Zou Youmin, "ZigBee routing selection strategy based on data services and energy-balanced zigbee routing", Pr oceedings of the 2006 IEEE Asia2 Pacific Conference on Services Computing , Washington DC, IEEE Computer Society, 2006, pp. 400 - 404.

[4] Charles E. Perkins and Elizabeth M. Royer, "The Ad hoc On-Demand Distance Vector Protocol," Charles E. Perkins, pp.173–219, 2000[Ad hoc Networking, Addison-Wesley].

[5] Xiaohui, L., Kangling, F., Jinguang, G., and Liang, Z., "An improved ZigBee routing strategy for monitoring system," In First international conference on intelligent networks and intelligent systems, pp. 255–258, ICINIS,2008.

[6] Fariborzi, H., and Moghavvemi, M., "Energy aware multi-tree routing for wireless sensor networks," IET Communications, 3, pp.733–739,2009.

[7] LEE KK, KIM SH, and PARK HS, "An effective broadcast strategy for route discovery in the ZigBee network," International Conference on Advanced Communication Technology, Gangwon-Do, South Korea, pp.1187-1191, 2008.

[8] J. Wieselthier, G. Nguyen and A. Ephremides, "Energy efficient broadcast and multicast trees in wireless networks," ACM Mobile Networks and Applications 7, pp.481–492, 2002.

[9] P.-J.Wan, G.Calinescu, X.-Y.Li, O. Frieder, and Erratum, "Minimum-Energy Broadcast in Static Ad Hoc Wireless Networks," Wireless Networks 11(4): pp.531—533, 2005.

Preprocessing of samples in modeling of fetal macrosomia with counter propagation neural network

Xu Zhipeng
School of Physics Science and Information Engineering, Liaocheng University
Liaocheng, China
xuzhipeng@lcu.edu.cn

Shen Aifang
Department of Gynaecology and Obstetrics
Liaocheng Brain Hospital
Liaocheng, China
aifangshen@yahoo.com.cn

Abstract— **In modeling of fetal macrosomia, some inconsistent data are mixed in samples. Because the procedure of childbearing has been finished, the sample data can not be validated by visiting previous pregnant woman. In order to eliminate the inaccurate samples, a counter propagation neural network is established. Some traditional methods are also used to verify the discarded samples. The new training set shows better classification than original data set.**

Keywords- fetal macrosomia; preprocessing; counter propagation artificial neural network(CPANN)

I. INTRODUCTION

When the fetal weight is greater than 4000g, then the fetal is called macrosomia. During delivery large fetus suffers a great risk of shoulder dystocia, fetal hypoxia. Fetal macrosomia has great influence on selection of mode of delivery. Prediction of fetal macrosomia has important clinical significance.

In order to predict the fetal weight, usually several parameters are gathered from the mother and the fetus, they are height (H), weight(W), fundus height(FUH), abdominal circumference(AC) of the mother and biparietal diameter (BPD), femur length(FL), amniotic fluid index (AFL) from the ultrasonic measurement of the fetal. All the parameters should be measured just before the delivery. For each pregnant woman there is a sample data, the above seven variables form an input, and the actual fetal weight. Many samples from pregnant woman make up an original data set. Then a classification method should be used with these samples. In this paper a counter propagation artificial neural network(CPANN) is established to predict fetal macrosomia based on sample data. Due to the different professional skill of operation staff during examination, some sample's parameter may be inconsistent. In modeling some inconsistent parameters are found in sample data. But the delivery is already finished; the parameters can not be measured once again. In this paper with the assistance of CPANN and other three traditional methods of estimating fetal weight, some inconsistent samples are removed.

II. COUNTER PROPAGATION ARTIFICIAL NEURAL NETWORK

Counter propagation artificial neural network (CPANN) is developed from kohonen artificial neural network(K-ANN). K-ANN is proposed by Professor Teuvo Kohonen in 1982[1]. K-ANN is also called self organizing maps(SOM). K-ANN can map the similarity of multi-dimensional data to a 2-dimensional space. A K-ANN has two layers, the input layer and the output layer. In this paper the seven parameters of each sample, which are represented by $X_s(x_1, x_2, ...x_7)$, compose the input layer. The output layer are composed by many neurons which have the same number of dimensions as sample data, represented by $Y_t(y_1, y_2, ...y_7)$. Usually the output layer forms a top map. For each sample a distance is calculated between the neurons in the output layer and the input layer. The neuron which has smallest distance between the sample and every neuron is called the winner in top map. The winner and neighboring neurons are then updated according to a certain rule.

$$\Delta Y_{c,i} = k\left(X_{Si} - Y_{c,i}\right)$$

The $Y_{c,i}$ means the winning neuron. *K* is the factor that is decided by distance of neighboring neuron and winning neuron. For winning neuron, k=1; for neurons far from winning neuron, k = 0.

After all the samples are given to the K-ANN one by one, it is called an epoch. The K-ANN usually is trained with many epochs. When training is over, the neurons in the top map represent a special mapping of samples. Neighboring neurons in top map have similar structure than distant neurons. On the other hand, similar samples corresponds close neurons in top map. So K-ANN usually has widely application in unsupervised clustering.

CP-ANN is very similar to K-ANN. CP-ANN has an extra layer on top of the output layer of K-ANN. The extra layer has exactly the same number of neurons in output layer of K-ANN. But it has different number of dimensions. In fact the neurons in extra layer have the same number of dimensions of target. CP-ANN can label out every classes of samples in top map and makes classification easier. A Matlab toolbox of CP-ANN developed by Davide Ballabio and Roberto Todeschini is used in this paper [2][3].

III. CLUSTERING OF SAMPLES BASED ON CPANN

A. clustering of 2 class

A CPANN model is established based on 391 samples. Each sample has 7 parameters including height (H), weight(W), fundus height(FUH), abdominal

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.67

circumference(AC) of the mother and biparietal diameter (BPD), femur length(FL), amniotic fluid index (AFL) of the fetal. These parameters are the inputs of CPANN model. The targets of each sample are assigned a value according to the actual fetal weight. When the fetal weight is below 4000g, the value is 1, otherwise the value is 2. So the target has two classes. The output layer has 11*11 neurons. After 2000 epochs of training, the top map of model is shown in figure 1. Table 1 is the result of classification. From table 1, one can see that 26 samples whose actual fetal weight is above 4000g are assigned to class 1, at the same time 4 samples whose actual fetal weight is below 4000g are assigned to class 2. Totally there are 30 incorrect classifications. In figure 1, several neurons include different samples from 2 classes.

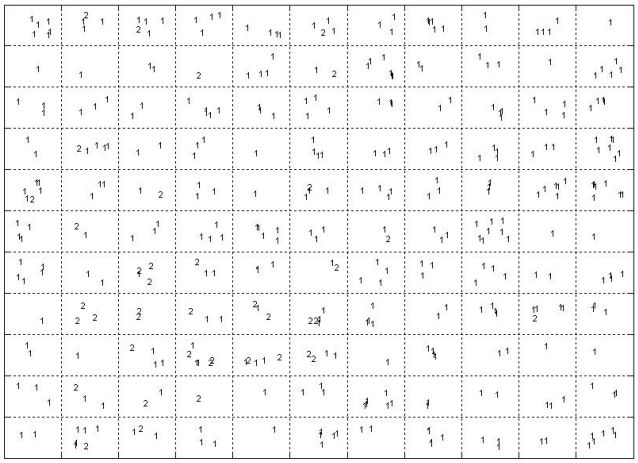


Figure 1 topmap of CPANN model with 2 classes

TABLE I. RESULT OF CLASSIFICATION

real/predicted	class 1	class 2
class 1	347	4
class 2	26	14

These inconsistent neurons scattered on the top map. According to the SOM theory, similar inputs will have same winning neuron, or they will have neighboring winning neurons. But the distribution of inconsistent neurons shows that the sample set has some inconsistent data.

B. clustering of 10 class

In order to analyze the inconsistent data more clearly, another CPANN model is established. The target of each sample is assigned to 1,2,3,… ,10 based on the actual fetal weight. Suppose yi is the output of sample i, then $y_i = floor((f_i - 2000)/300) + 1$, f_i is the actual weight. The fetal weight which is greater than 2000g and less than 5000g are divided into 10 intervals. The output layer also has 11*11 neurons. After the model is trained 2000 epochs, the top map is shown in figure 2.

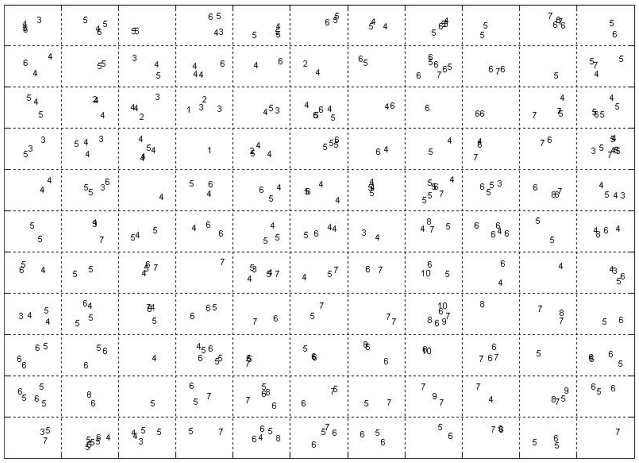


Figure 2 topmap of CPANN model with 10 classes

In figure 2, one can see several neurons correspond to clearly different inputs. Suppose the coordinate (*x,y*) represents a neuron whose position is the *x*th row of topmap from up to down and *y*th column from left to right. Then the neuron (4,11) corresponds to 8 samples whose fetal weights range from class 3 to class 7. The actual weights range from 2900g to 4400g. Such samples should not activate the same neuron. In other words, the sample data correspond to neuron (4,11) has inconsistency. Besides neuron (4,11), there are also several other neurons that have same circumstances.

C. 391 samples decrease to 366 samples

A detailed analysis of neurons on the top map is done in sequence. The neurons whose class difference is greater than 3 are selected; accordingly the difference of actual fetal weight is greater than 1200g. Totally 26 neurons are selected from top map of 121 neurons. For each neuron, the corresponding parameters and actual fetal weight are extracted from sample data and form a new group. In every group classes scattered at the both ends will have the most probability to eliminate. For instance the group of neuron (4,11) has classes 3,4,4,5,5,5,5,7. The samples correspond to classes 3 and 7 are most likely to eliminate. Certainly the doubtful samples should not be eliminated curtly. In fact, several conventional methods are used to calculate the fetal weight of entire group. These methods are proposed by Luo Laimin[4],Ling Luoda[5],Yuan Dongsheng[6]. The group of neuron (4,11) is shown in table 2. The eliminated sample should satisfy the following two conditions:

1) The eliminated sample has to be scattered at the both ends.

2) The eliminated sample has to meet at least two maximum of error within three methods.

The sample with ID 1 gets maximum in error of Yuan (678g)and Ling(794g), also it's corresponding class is 7. So the sample with ID 1 should be eliminated according to above two conditions. Totally there are 25 inconsistent samples being eliminated from original 391 samples. So the remaining number of samples is 366.

IV. RESULT

In top map the neighboring neurons have more similar structure than distant neurons. Every neuron in top map corresponds to several samples, which have similar parameters according to SOM theory.

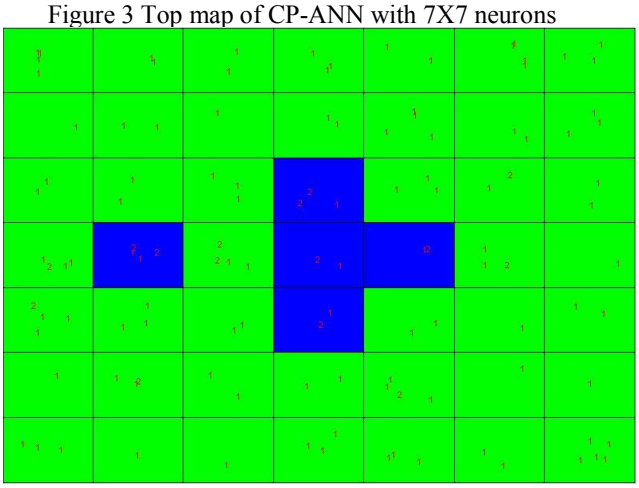
The training set should represent the various characteristic of whole sample data. In each group of neuron only one sample is selected to represent the entire group because of similarity. The selected sample should have the minimal distance to the average of the group [7]. Table Ⅲ shows groups corresponding to partial neurons of top map. Table Ⅳ shows the selected sample indices from each group. There are 121 samples in table Ⅳ, which compose the training set.

A CP-ANN is established with 7X7 neurons. The top map is shown in figure 3. In the figure, the green neurons represent the normal fetal whose weight is below 4000g, while the blue neurons represent fetal macrosomia. The blue neurons focus on the center of the top map, while the green neurons surround the blue neurons. The two classes are classified clearly. Neuron (4,3) has a normal weight but its position is close to fetal macrosomia. It can be explained that to Neuron (4,3), two fetal macrosomia and two normal weight make up the group, the neuron should be inclined to fetal macrosomia.

V. CONCLUSION

In this paper, a K-ANN method is used to eliminate the inconsistent samples. Several traditional methods are used to verify the discarded samples. The result shows new training set has more clear classification than original sample data.

Figure 3 Top map of CP-ANN with 7X7 neurons



REFERENCES

[1] T.Kohonen, "Self-Organizing Formation of Topologically Correct Feature Maps", Biological Cybernetics, 1982, Volume 43, Number 1, Pages 59-69.

[2] Davide Ballabio and Roberto Todeschini, Kohonen and CPANN toolbox (for MATLAB), http://michem.disat.unimib.it/chm/download/kohoneninfo.htm.

[3] Ballabio D, Consonni V, Todeschini R. "The Kohonen and CP-ANN toolbox: a collection of MATLAB modules for Self Organizing Maps and Counterpropagation Artificial Neural Networks",Chemometrics and Intelligent Laboratory Systems, 2009, 115-122.

[4] Luo Laimin, Dai Zhongying,Tong Jianqian, "estimation of weight of fetal macrosomia", Chinese Journal of Obstetrics and Gynecology,Vol 30, June,1995,pp 375.

[5] Ling Luoda, Gu Meili, dystocia of vertex presentation, Chongqing Publishing House,1992.

[6] Yuan Dongsheng,Gao Huizhen, "Discussion of estimating fetal weight", intermediate medical journal, Vol.4,April,1984,pp 18.

[7] Jure Zupan, Marjana Novic a, Itziar Ruisinchez, "Kohonen and counterpropagation artificial neural networks in analytical chemistry", Chemometrics and Intelligent Laboratory Systems, Vol.38, 1997,pp1-23.

TABLE II. SAMPLE DATA OF GROUP CORRESPONDING TO NEURON(4,11) AND FETAL WEIGHT ESTIMATED BY YUAN, LUO AND LING

ID	H (cm)	W (kg)	FUH (cm)	AC (cm)	BPD (cm)	FL (cm)	AFL	Actual Fetal Weight (g)	class	Yuan (g)	Error Of Yuan (g)	Luo (g)	Error Of Luo (g)	Ling (g)	Error Of Ling (g)
1	163	73	32	96	9.3	7.2	15.2	3950	7	3272	678	3821.6	128	3156	794
2	162	72	31	96	9.3	7	14	2800	3	3176	376	3792.8	993	3033	233
3	162	67	35	95	9.2	7.3	15.2	3300	5	3525	225	3897.5	598	3505	205
4	165	70	34	98	9.1	7.35	14.8	3300	5	3532	232	3899.6	600	3442	142
5	164	66	32	93	9.5	7.5	13.9	3100	4	3176	76	3792.8	693	3096	4
6	162	72	33	97	9.3	7.1	15.5	3500	5	3401	99	3860.3	360	3299	201
7	165	69	34	98	9.4	7.1	12.6	3000	4	3532	532	3899.6	900	3442	442
8	164	68	32	97	9.3	7.3	13.2	3300	5	3304	4	3831.2	531	3176	124

TABLE III. SAMPLE INDICES OF NEURONS IN TOP MAP(PART)

[59;108;113;184]	[124;206;256;271]	[19;273]	[127;196;235;270]
[48;95;374]	[244;336]	[243;294;360]	[101;185;325;335]
[263;326;382]	[15;85;259]	[16;58;110;154]	[155;159;225;358]
[94;207;224]	[29;39;121;391]	[41;54;67;169;353]	148
[149;362]	[53;89;233;381]	198	[57;191;386]

TABLE IV. SELECTED SAMPLE INDICES FROM TABLE III(PART)

113	256	19	270	253	23	5	309	300	76

Real-time Automatic Evaluation Technology in Vehicle Road Test System Based on Neural Network

Yefu, Wu
College of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
wuyefu@whut.edu.cn

Li Bo
College of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
cstlibo@whut.edu.cn

Chaozhong, Wu
Intelligent Traffic System Center
Wuhan University of Technology
Wuhan, China
wucz@whut.edu.cn

Shigang, Pan
College of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
panda_198711@yahoo.com.cn

Shen Hui
College of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
Huiwei870620@163.com

Abstract*—With the development of computer, sensors, artificial intelligence technology, network, advanced automotive technology, the vehicle road test by a automatic evaluation system from the traditional manual evaluation has become an irresistible trend. Currently, the research of intelligent test system for vehicle road test in China has just started. Automatic evaluation module is core component of the system. In this study, according to the theory and technology of artificial neural network and the characteristics and requirements of the vehicle road test system, we introduce the one-way transmission of multi-layer neural network structure to the automatic evaluation module, discuss the principle, structure, transfer function and specific implementation of the application.Finally the proposed model was applicated in the vehicle road test system.

Keywords: Vehicle Road Test System, Neural Network, Automatic Evaluation, Artificial Intelligence

I. INTRODUCTION*

In the driver test, improving the overall quality of motor vehicle drivers, consolidating drivers' safety, law-abiding, civilized driving awareness, developing their good driving habits are essential for the prevention of traffic accidents and the formation of traffic civilization. To ensure the quality and efficiency of the driver test, intelligent road test system based on computer, artificial intelligence technology etc has been applied to more and more institutions in the driving test.

Presently, the existing examination mode of vehicle road test is basically a manual evaluation. Only a few areas used intelligent test system for the evaluation of vehicle road test, but these systems are designed based on closed and simulated road. They don't meet the requirements of Document 111(Provisions of Driving License's Apply and Use) which is issued by Ministry of Public Security. For the above, in support of National Nature Science Foundation of China (60974094) and National Key Projects in the Science & Technology of China (2009BAG13A05), we developed the vehicle road test system which is based on actual road. The system can acquire data on candidates' driving behavior in real time, then process the data through artificial intelligence methods, and finally implement the real-time automatic evaluation.

Since 1943, American psychologist McCulloch and mathematician Pitts proposed the abstract mathematical model of neurons form (MP model), a lot of research achievement about artificial neural network theory and technology has been made, and they have been widely used in expert systems, intelligent control, robotics, pattern recognition and so on.

II. FORWARD NEURAL NETWORK

According to the interconnection structure between neurons, neural network can be divided into the following types: (1) forward neural network without feedback; (2) a neural network with feedback from the output layer to input layer; (3) a type like this: there are neurons in combination in the same layer; (4) combination network.

According to the actual characteristics of the vehicle road test system, in the automatic evaluation process, we can use a multilayer forward neural network whose structure is shown in Figure 2.1. In the network, besides the input layer and

*This study is supported by National Nature Science Foundation of China (60974094),National Key Projects in the Science & Technology of China (2009BAG13A05)

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.73

output layer, there are one or more hidden layers. There are no links between nodes in the same layer; each node is a single neuron. According to the actual situation, the transfer function of neurons can be chosen as Step-type function, Sigmoid function or Gaussian type function (following equation).

$$f(x)=\begin{cases}1 & (x\geq 0)\\ 0 & (x<0)\end{cases} \quad \text{(1) Common Step-type function}$$

$$f(x)=th(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad \text{(2)Common Sigmoid function}$$

$$y_i=f(x)=\exp\left(-\frac{1}{2\sigma_i^2}\sum_j(x_j-W_{ji})^2\right) \quad \text{(3) Common Gaussian function}$$

(σ_i^2 is standardized parameters).

In the multilayer forward neural networks, there is no coupling between the neurons in the same layer. In each layer, neurons only accept the input of neurons in the layer before and their outputs only affect the outputs of the neurons in the next layer. In the above neural network structure, input data X=(x1, x2,⋯,xn) from the input layer is processed by the transfer function fi(X)of the hidden layers to reach the output layer, resulting in the output data Y= (y1,y2, ⋯ ,ym).This is very consistent with the actual evaluation process of the vehicle road test system, so it can be well applied to the automatic evaluation process of the vehicle road test system.

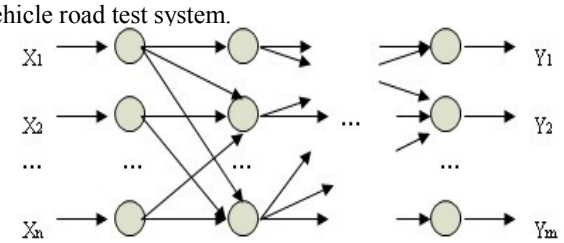


Figure 2.1 Structure of multilayer forward neural network

III. AUTOMATIC EVALUATION MODEL BASED ON NEURAL WORK

A. Design of Vehicle Road Test System

According to the actual test process of the Vehicle Road test and the hardware and software conditions of the system, the design of the vehicle road test system which is consistent with real-time, dynamic and stability is shown in Figure 3.1.

The major steps of data acquisition module includes: (1) Access the test process data from the test vehicle; (2) Preprocessing of these data; (3) Generate input layer data of the automatic evaluation module. All the steps are real-time and dynamic. Data acquisition module constantly get the test process data from related equipment and interfaces, then process the data and transfer to the automatic evaluation module rapidly. It's processing speed in milliseconds. Data acquisition module processes the test data into the state space, as follows:

$$D=\begin{pmatrix} D_{11} & \cdots & D_{1n} \\ \vdots & \ddots & \vdots \\ D_{41} & \cdots & D_{4n} \end{pmatrix} \quad \text{(4) Test process data state space}$$

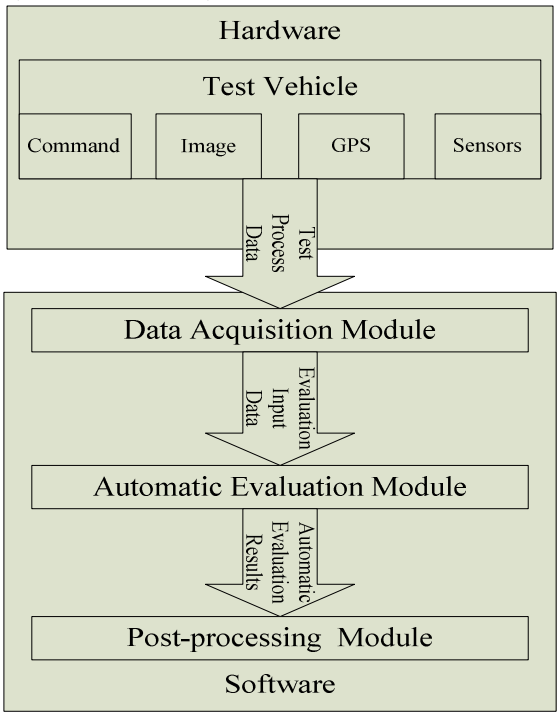


Figure 3.1 Structure of vehicle road test system

In the above formula, D_{1n} said the sensor data, such as steering wheel angle (0-720), foot brake (0, 1), the handbrake (0, 1), door (0, 1), belts (0, 1), manual gear (0-6), speed (0-200); D_{2n} said GPS data, such as longitude (double type), latitude (double type), lane (int type), over the line (0,1); D_{3n} said the image state data of head (int type), such as look straight (0), left swing (2), right swing (3), look down (1); D_{4n} said test command data, such as start (00), turn left (02), turn right (03), overtake (04), pull up (09).

Automatic evaluation module is core module of vehicle road test systems, which is mainly for test data processing such as knowledge acquisition, knowledge reasoning, and knowledge interpretation. In the automatic evaluation module, for a specific test data input mode, we can get related output mode which is computed by the transfer function of the neural network. Then interpreter explain the output mode (usually numeric vector) into a high-level logical concepts such as an automatic evaluation item points, so as to automatically judge the candidate's driving behavior. The structure of automatic evaluation module is shown in Figure 3.2.

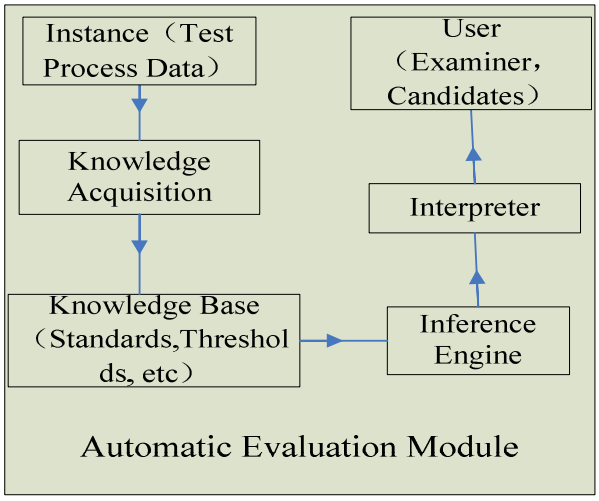


Figure 3.2 Structure of automatic evaluation module

During the process of knowledge acquisition, automatic evaluation module gets learning instance and their corresponding expectations solution (evaluation item standards). Until the network is stable, it generates the knowledge base (rules and thresholds for evaluation) for reasoning. According to the specific instance (test process input data), inference engine can get the corresponding output mode (vector mode) which is to be interpreted by the interpreter into some high-level logical results. Then automatic evaluation module presents the results to the user. Knowledge in expert system based on neural network is implicit, non-local. Unlike the interpreter in a rule-based expert system, the interpreter in a neural network expert system can't record the chair of rules in the reasoning process, so it is difficult for the interpreter to provide users the interpretation for "why" and "how", it can just provide users the evaluation results.

Post-processing module is mainly to process the evaluation results, such as save the evaluation results into the database, upload them to the test center, provide the view function of evaluation results and test process. It has an important role in enhancing the system's automation and intelligence, improving the function and efficiency of the system.

B. Application of Automatic Evaluation Based on Neural Network

In the automatic evaluation module of vehicle road test system, we set three hidden layer to process the data from input layer. Transfer function fn (n=1, 2, 3) in hidden layer are responsible for knowledge acquisition, knowledge base update and knowledge reasoning.

According to the rules of vehicle road test, we can use rules vector Ri(x1, x2, ⋯ ,xn) to stand for automatic evaluation items, where Ri indicates the corresponding rules, xi said the test process data which is associated with the evaluation item. For example, the rule "Start without putting down the handbrake" can be expressed as R (start (1), speed (10), handbrake (1)). According to predefined automatic evaluation items, we can get the corresponding rule vector Ri to generate the initial knowledge base. In the process of updating the knowledge base, stable rule vector can be expressed as $R_i() = \overline{R}_i = 1/n \sum_{j=1}^{n} R_{ij} = R_i(\overline{x}_1, \overline{x}_2 \cdots \overline{x}_n)$, where $\overline{x}_n$ indicates stable threshold of test process data in corresponding evaluation rule.

In order to process data collected by the data acquisition module, the transfer function of hidden layer 1 f1 is defined as follows:

$$f_1(D_{mn}) = \begin{cases} 1 & (D_{mn} = true \,||\, D_{mn} \ \ Exceeds \ \ the \ \ threshold) \\ 0 & (D_{mn} = false \,||\, D_{mn} \ \ Not \ \ exceed \ \ the \ \ threshold) \end{cases} \quad (5)$$

We can see that the transfer function of hidden layer 1 f1 is a kind of step-type function. Function f1 can streamline the data from the state space of test process data, generate the binary form of input data which can facilitate knowledge acquisition.

According to the predefined rules vector of automatic evaluation items, the transfer function of hidden layer 1 f2 links the test process data which is associated in the same evaluation item to form the input vector which is effective for the neural network. This kind of input vector is the directional input for the automatic evaluation module's reasoning and evaluation, so it can improve the speed of automatic evaluation's reasoning. The transfer function of hidden layer 2 f2 is defined as follows:

$$f_2(f1(D)) = V_{Ri}(f_1(x_1), f_1(x_2), \cdots f_1(x_n)) \quad (6)$$

（x1,x2,⋯,xn said the associated test process data in Ri, coming from state space D, scilicet xn=Dmn）。

For example, suppose the threshold of start speed is 5, we have the test process data shown in Table 3.1.

Table 3.1 Init Test Process Data

Command: start	Sensor: speed	Sensor: handbrake	GPS: lane	GPS: borderline	Image: head swing
true	10	true	1	false	1

Processed by f1, we can get the binary input data shown in Table 3.2.

Table 3.2 Binary input data

Command: start	Sensor: speed	Sensor: handbrake	GPS: lane	GPS: borderline	Image: head swing
1	1	1	1	0	1

Suppose R0 stands for the automatic evaluation item "Start without putting down the handbrake", then R0 can be expressed as R0 (start, speed, handbrake) .After processed by f2, we can get the input vector VR0(1,1,1) which is corresponding to the rule R0.

The transfer function of hidden layer 3 f3 matches VRi to the rule vector with the stable threshold in the knowledge base, then resulting in an output mode. In the matching process, we can set the credibility factor for each evaluation

item based on the item's accuracy. For the credible match, the output mode of corresponding automatic evaluation item which means points penalized will be given. For the other match that is not credible, no treatment will be done which means the automatic evaluation item is not penalized. The transfer function of hidden layer 3 f3 can be defined as follows:

$$f_3(V_{Ri}) = \begin{cases} 1 & (V_{Ri}(\bar{x}_1, \bar{x}_2 \cdots \bar{x}_n) \geq V(R_i(\bar{x}_1, \bar{x}_2 \cdots \bar{x}_n)) * \mu(R_i)) \\ 0 & (V_{Ri}(\bar{x}_1, \bar{x}_2 \cdots \bar{x}_n) < V(R_i(\bar{x}_1, \bar{x}_2 \cdots \bar{x}_n)) * \mu(R_i)) \end{cases} \quad (7)$$

According to hardware and software of the system and the credibility factor of the evaluation items, we can set the appropriate $\mu(R_i)$ .Generally, for the more accurate automatic evaluation item; $\mu(R_i)$ can be set to (0.9-1.0).

Interpreter transfers the output mode given by f3 to the evaluation results, then present the results to examiner and candidate. According to the definition of f3, we can design the structure of interpreter as shown in Figure 3.3.

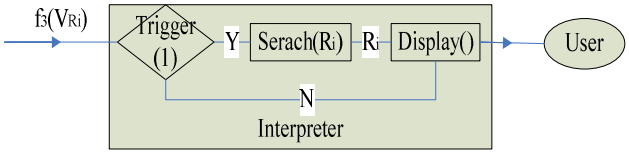


Figure 3.3 Structure of interpreter

When $f_3(V_{Ri})$ =1, interpreter searches corresponding content of Ri (such as content, point, etc), and then presents them to the user. Some code of the transfer function in the automatic evaluation module is given as follow:

```
f1(D[m][n]){
for(i=0;i<m;i++){
for(j=0;j<n;j++)
if((D[i][j]==true)||(|D[i][j]|>D[i][j]_0) //Binary input data
D[i][j]=1;
else
D[i][j]=0;}
return D[m][n];}
//Form the input vector
f2(D[m][n])
{
for(i=0,R=R0;Ri;i++)
{vector VRi;
D[i][j]=select D from Ri; /Search the associate data of Ri
VRi.pushback(D[i][j]); //Generate the input vector of Ri
} }
//Give output mode
f3(VRi){
for(i=0,R=R0;Ri;i++)
{VRi_0=Select V from V(Ri);
μ(Ri) = Select μ from V(Ri);
if(VRi>VRi_0* μ(Ri) {Flag[i]=1;}
else{ Flag[i]=0;}
if(Flag[i]){Display(Select R[i] from R);} //Give results
}}
```

IV. INSTANCE

In the system, the automatic evaluation of test item "Start without putting down the handbrake " is shown in Figure 4.1, at the same time, we can see that the system also give the evaluation results of the items "Start with which the door is open " and "Driving with zero gear" etc. Thus, the automatic evaluation based on neural networks is parallel and global.

Figure 4.1 Instance of automatic evaluation

In Table 4.1, we can see the comparison of some main technical parameters between automatic evaluation and manual evaluation in the test of subject three.

Table 4.1 Comparison of automatic evaluation and manual evaluation

	Automatic evaluation	**Manual evaluation**
Process time	100ms	>=3s
Effective evaluation item in one process	68	1-5，Average 3
Accuracy	90%-100%	60%-100%, Average 80%
Real-time	High	General
Overall efficiency（Accurate evaluation times per second）	68/0.1*95%=646	3/3*0.8=0.8

In the table, we can see that real-time automatic evaluation technology based on neural network, which is very practical and applied, has greatly improved the efficiency of vehicle road test, reduced the burden on the examiner, and improved the accuracy of the evaluation.

V. CONCLUSION

Artificial intelligence and neural network have been widely used in many areas. In this paper, we mainly study the artificial intelligence and neural network's application in the area of transportation. Meanwhile, we design and implement the vehicle road test system which is based on the automatic evaluation technology and neural network.

ACKNOWLEDGMENT

This study is supported by the National Nature Science Foundation of China (60974094), National Key Projects in the Science & Technology of China (2009BAG13A05). We would like to thank Chunliang Cheng, Liangli Zhang, Gao Song, and Zondy Cyber Group for their helpful comments.

REFERENCES

[1] Haykin, 2009 S. Haykin, Neural networks and learning machines (3rd ed.), Pearson Education Inc., Upper Saddle River, New Jersey (2009).J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Chio,A., Nareshkumar, N. "Simultaneous verses successive learning in neural networks" in: Systems, Man and Cybernetics, 2009.SMC 2009.IEEE International Conference on, 2009, pp4224-4231.

[3] Paliwal and Kumar, 2009 M. Paliwal and U.A. Kumar, Neural networks and statistical techniques: A review of applications, Expert Systems with Applications **36** (2009), pp. 2–17

[4] J. Velagic, N. Osmic, B. Lacevic "Neural Network Controller for Mobile Robot Motion Control", in: World Academy of Science, Engineering and Technology, vol. 47, 2008.

[5] Wang X W, Tan D J. Application and Developing Trends of Neural Network [J].Computer Engineering and Application, 2003, (3).

[6] Ministry of Public Security P.R.C,《Provisions of Driving License's Apply and Use》,Document 111,2010.

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Detecting and correcting the index of DAE by the combinatorial relaxation algorithm

Yongxin Yan, Xiaolei Zhang
International School
Beijing University of Posts and Telecommunications
Beijing, China
e-mail: ee08b044@bupt.edu.cn;
ee08b226@bupt.edu.cn

Xuesong Wu*, Jianwen Cao
Laboratory of Parallel Software and Computational Science,
the Institute of Software,
Chinese Academy of Sciences
Beijing, China
e-mail: xuesongwu@msn.cn
caojianwen@gmail.com

***Abstract*—The mathematical modeling and simulation based on the Modelica language usually gets a high-index differential-algebraic equation (DAE) system. The structural index reduction algorithms can serve as a fast method to reduce such high-indexed issues. In order to solve the failure of the structural index reduction algorithms in some cases, the combinatorial [1]relaxation algorithm is analyzed and studied. Finally, the result of an example shows the combinatorial relaxation algorithm is an effective way to improve the stability of the index reduction algorithms based on the structural index of DAE.**

Keywords-Modelica; differential-algebraic equation; index reduction; combinatorial relaxation algorithm

I. INTRODUCTION

Mathematical modeling and numerical simulation of product design enables the study of complex systems that would be too expensive or dangerous, or even impossible, to study by direct experimentation. However, in recent years, many products are integrated by machinery, electronics, gas, control, heat and the other in different fields. Their design is very complicated. Traditional single-field simulation technology can not adapt the task of simulating complex systems. Multi-domain modeling language has developed for modeling and simulation technology of complex systems.

Modelica[1-3] is an object-oriented modeling language to model large, complex and heterogeneous physical systems. The language is designed for convenient, component-oriented modeling of physical multi-domain systems. Modeling and simulation based on the Modelica language describe the physical laws and phenomena in different areas by mathematical equation, and solve the differential-algebraic equations (DAE) for numerical simulation. Therefore, in the electrical, mechanical, thermal dynamics, chemical, biological, economic, and aerospace vehicle design and other fields, Modelica has a wide range of applications. The simulation softwares based on the Modelica language often used now are Dymola[4] , OpenModelica[5], MathModelica, SCICOS, MapleSim and SimulationX etc.

However, the modeling software based on the Modelica language probably produces the high-index DAE that is difficult to solve directly. In order to solve the high-index DAE system, it is necessary to convert it into a low-index DAE system by index reduction method. But the index reduction method will fail in some cases. In this paper, we first analyze the reasons why the failure occurred. Then we introduce the combinatorial relaxation algorithm used to detect and correct the problem, and discuss the details of the implementation of the combinatorial relaxation algorithm. Finally, we give an example of the combinatorial relaxation algorithm.

This paper is organized as follows: Section 2 contains an overview of index reduction method in DAE; in Section 3, we discuss the details of the implementation of the combinatorial relaxation algorithm; we show an example of the implementation of the algorithm in Section 4; Section 5 contains some conclusions.

II. DAE INDEX REDUCION

To support flexible and safe reuse of model components, the modeling software based on the Modelica language probably produces high-index DAE. It is well known that the high-index DAE is difficult to solve directly. The existing DAE solvers generally only solve the ODE and DAE system of index less than or equal to 3. For example, DASSL, RADAU5 and RADPK etc. To solve the high-index DAE system, a common approach is to transform the DAE system into ODE or DAE system of index-1 by index reduction method, and then solve the DAE system of index less than or equal to 1 by direct solvers.

Index reduction for DAE system is very important. There are two types of index reduction method: one is the traditional numerical methods based on the differential index, such as gear algorithm [6]. Gear algorithm reduces the differential index of the DAE system by differentiating some equations. It search algebraic equations in the DAE system by the symbolic manipulation, and then differentiates the algebraic equations. The implementation of this process will repeat until the DAE system is transformed into the ODE

*Corresponding author

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.37

system. But it often fails for the DAE system that includes some nonlinear equations, because some nonlinear equations do not permit an explicit solution.

Index reduction algorithm based on the structural index includes pantelides algorithm [7] and dummy derivatives algorithm [8] algorithm etc. It is based on matrix and graph theory, through determining the index of the structure of equations, tests a subset of the matrix of DAE system repeatedly, and then differentiates the equation subset, and then modifies the corresponding bipartite graph of the DAE system, finally gets the low index DAE system. These methods are based on algorithms of graph theory and symbol processing of correlation matrix. For comparison with the traditional numerical methods based on the differential index, it has smaller quantity of symbolic processing, lower computational complexity, so it has been widely used (for example, in OpenModelica). But for some the DAE problems, index reduction method based on structural index will fails. It is because the method based on structural index only considers the structural features of the equation; it may lead to incomplete reduction or excessive reduction when the differential index of the DAE system is incompatible with structural index. So in a few cases, it will cause the inconsistency between the differential index and the structural index, and lead to failure of reduction. In some specific cases, the reduced DAE system is not an ODE system, which leads to error when calling the solver.

Fig.1 shows an example of failed index reduction.

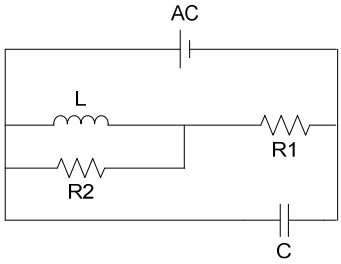


Fig.1 Electronic circuit model

By modeling with the Modelica language:

```
model Circuit
   Modelica.Electrical.Analog.Sources.ConstantVoltage
AC;
   Modelica.Electrical.Analog.Basic.Resistor R1 ;
   Modelica.Electrical.Analog.Basic.Inductor L ;
   Modelica.Electrical.Analog.Basic.Resistor R2 ;
   Modelica.Electrical.Analog.Basic.Capacitor C;
equation
   connect(AC.n,R1.n);
   connect(AC.p,L.p);
   connect(L.n,R1.p);
   connect(C.p,L.p);
   connect(R1.n, C.n);
   connect(R1.p,R2.n);
   connect(R2.p,L.p);
end Circuit;
```

and BLT sorting, a DAE subsystem is given by (2.1):

$$\begin{cases} der(I_4)*L.L-U_4=0 \\ -I_5+C.C*der(U_5)=0 \\ -U_1-U_2-U_4=0 \\ U_2+U_3-U_5=0 \\ -U_3+U_4=0 \\ I_2*R1.R-U_2=0 \\ I_3*R2.R-U_3=0 \\ I_1-I_2-I_5=0 \\ -I_1+I_3+I_4+I_5=0 \end{cases} \tag{2.1}$$

By calculation, the structural index of the DAE system (2.1) is 1, but the differential index is 2. It will fail by using index reduction method based on structural index. And the solver will fail to solve the reduced DAE system. Therefore, the detection and correction for structural index of the DAE system is very important for the index reduction method based on the structural index.

III. COMBINATORIAL REDUCTION ALGORITHM

Kazuo Murota, Satoru IWATA and Izumi SAKUTA proposed the combinatorial relaxation algorithm [9-11], which combines combination-optimization theory and structural index reduction methods and gives a good solution to detect the failure and correct the problem of structural index reduction methods. The theory detect whether the structural index is compatible with the differential index by four rank constraint rules. Only when all four rules are satisfied, the two types of index are compatible; contrary to any rule, index reduction algorithm based on the structural index is invalid and the original matrix of DAE must be transformed. In response, the combinatorial relaxation algorithm gives an effective transformation strategy, and finally solves the problem. The outline of the combinatorial relaxation algorithm to compute and correct the structural index of the DAE system is as follows:

Phase 1: Compute $\hat{\delta}_k(A)$ by solving the weighted bipartite matching problem.

Phase 2: Test whether $\hat{\delta}_k(A)=\delta_k(A)$ or not. If so, then return $\hat{\delta}_k(A)$ and stop.

Phase 3: Modify $A(s)$ to another matrix pencil $\widetilde{A}(s)$ that satisfies $\delta_k(\widetilde{A}) = \delta_k(A)$ and $\hat{\delta}_k(\widetilde{A}) < \hat{\delta}_k(A)$. Put $A(s) := \widetilde{A}(s)$ and go to Phase 1.

Where the definitions of $A(s)$, $\delta_k(A)$, $\hat{\delta}_k(A)$ are the following:

1. Let $A(s)$ be an $m\times n$ matrix pencil of rank r with the row set R and the column set C. [11] presents an algorithm for computing the maximum degree $\hat{\delta}_k(A)$ of

minors of order k for a given integer k. Namely, the author intend to compute

$$\delta_k(A) = \max\{ \deg \det A[I, J] \mid |I| = |J| = k \},$$

where $A[I, J]$ designates the submatrix of $A(s)$ with the row set I and the column set J.

2. Construct a bipartite graph $G(A) = (R, C; E)$ with the vertex sets R and C. The edge set E corresponds to the nonzero entries, i.e., $E = \{(i, j) \mid i \in R, j \in C, A_{ij}(s) \neq 0\}$. Each edge $e = (i, j) \in E$ has its weight $w_e = \deg A_{ij}(s))$ and the set $\partial e = \{i, j\}$ of its end-vertices. A subset M of E is called a matching if $\partial M = 2|M|$ holds, where $\partial M = \bigcup \{\partial e \mid e \in M\}$. The termrank of A, denoted by t-rank A, is the maximum size of a matching in $G(A)$. Let $\hat{\delta}_k(A)$ denote the maximum weight of a matching in $G(A)$, where the weight of a matching M refers to $w(M) = \sum \{w_e \mid e \in M\}$ [11].

In this process, the combinatorial relaxation algorithm first gives a mapping between the DAE system and a bipartite graph, then uses the combination algorithm to get the maximum weight matching problem of the bipartite graph, and finally calculates the value of the variable which decides the size of the structural index. The structural index can be given by

$$\nu_{str}(A) = \hat{\delta}_{k-1}(A) - \hat{\delta}_k(A) + 1;$$

the differential index can be given by

$$\nu(A) = \delta_{k-1}(A) - \delta_k(A) + 1$$

for a regular matrix pencil $A(s)$ in the context of DAE.

Then through the rank theory of matrix, the combinatorial relaxation algorithm detect whether the structural index is compatible with the differential index by four rank constraint rules of relative matrices.

(r1) Rank $A^*[R, C] \geqslant k$;
(r2) Rank $A^*[I^*, C] = |I^*|$;
(r3) Rank $A^*[R, J^*] = |J^*|$;
(r4) Rank $A^*[I^*, J^*] \geqslant |I^*| + |J^*| - k$.

Where the definitions of A^*, I^*, J^* are the following [11]:
Consider the linear programming formulation $PLP_k(A)$ of the weighted matching problem:

$$\begin{aligned} &\text{Maximize } \sum_{e \in E} w_e \xi_e \\ &\text{sub. to } \sum_{\partial e \in i} \xi_e \leq 1 \quad (i \in R), \\ &\qquad \sum_{\partial e \in j} \xi_e \leq 1 \quad (j \in C), \\ &\qquad \sum_{e \in E} \xi_e = k, \\ &\qquad \xi_e \geq 0 \quad (e \in E). \end{aligned}$$

The dual linear programming problem $DLP_k(A)$ is as follows:

$$\begin{aligned} &\text{Minimize } \sum_{i \in R} p_i + \sum_{j \in C} q_j - kt \\ &\text{sub. to } p_i + q_j - t \geq w_e \quad (e = (i, j) \in E), \\ &\qquad p_i \geq 0 \quad (i \in R), \\ &\qquad q_j \geq 0 \quad (j \in C). \end{aligned}$$

Denote by $\pi_k(p, q, t)$ the dual objective value. Since the coefficient matrix of the constraints in $PLP_k(A)$ is totally unimodular, both $PLP_k(A)$ and $DLP_k(A)$ have integral optimal solutions.

With reference to an integral optimal solution (p, q, t), we put

$$R^* = \{i \mid i \in R, p_i > 0\}$$

and

$$C^* = \{j \mid j \in C, q_j > 0\}.$$

The tight part of A, denoted by A^*, is a constant matrix with A^*_{ij} being the coefficient of $s^{p_i + q_j - t}$ in $A_{ij}(s)$.

Finally we construct the transformation matrix so that the difference between the structural index and the differential index gradually will be reduced. When the structural index is not equal to the differential index, we need to construct the transformation matrix and transform the original DAE, gradually reduce the difference between the structural index and the differential index. Finally we make the two types of index equal, and transform the original DAE problem into a new DAE problem that the index reduction algorithm based on the structural index can reduce the new equations correctly.

IV. IMPLEMENTATATION

Consider the problem (2.1), and assume R1.R=1.0, R2.R=1.0, L.L=1.0, C.C=4.0, AC.V=5.

The Matrix $A(x)$ is:

$$A(x) = \begin{bmatrix} 0 & 0 & 0 & x & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 4x \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The bipartite graph is:

$$G(A)=\begin{bmatrix} -\infty & -\infty & -\infty & 1 & -\infty & -\infty & -\infty & 0 & -\infty \\ -\infty & -\infty & -\infty & -\infty & 0 & -\infty & -\infty & -\infty & 1 \\ -\infty & -\infty & -\infty & -\infty & -\infty & 0 & -\infty & 0 & -\infty \\ -\infty & -\infty & -\infty & -\infty & -\infty & 0 & 0 & -\infty & 0 \\ -\infty & -\infty & -\infty & -\infty & -\infty & -\infty & 0 & 0 & -\infty \\ -\infty & 0 & -\infty & -\infty & -\infty & 0 & -\infty & -\infty & -\infty \\ -\infty & -\infty & 0 & -\infty & -\infty & -\infty & 0 & -\infty & -\infty \\ 0 & 0 & -\infty & -\infty & 0 & -\infty & -\infty & -\infty & -\infty \\ 0 & -\infty & 0 & 0 & 0 & -\infty & -\infty & -\infty & -\infty \end{bmatrix}$$

Compute the maximum weight of a matching in $G(A)$, get I={1,1,0,0,0,0,0,0,0}, J={0,0,0,0,0,0,0,0,0}, and q=0. So the I*={1,2}, J*=∅.

Compute the rank of A* when k = 9:

(1). Rank A*[R, C] = 8， k = 9;
(2). Rank A*[I*, C] = 2， |I*| = 2;
(3). Rank A*[R, J*] = 0，|J*| = 0;
(4). Rank A*[I*, J*] = 0，|I*| + |J*|-k = -7.

The result is contrary to the condition (r1). The construct a matrix $U(x)$:

$$U(x)=\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.25x^{-1} & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$V(x)=I$, then

$$A'(x)=U(x)A(x)V(x)=\begin{bmatrix} 0 & 0 & 0 & x & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 4x \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & -0.25x^{-1} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Then compute rank (k=9) again:

(1). Rank A*[R, C] = 9, k = 9;
(2). Rank A*[I*, C] = 8, |I*| = 8;
(3). Rank A*[R, J*] = 0, |J*| = 0;
(4). Rank A*[I*, J*] = 0, |I*| + |J*| - k = -1.

The results are in accord with the conditions (r1)-(r4). Compute the rank of A* when k = 8:

(1). Rank A*[R, C] = 9, k = 8;
(2). Rank A*[I*, C] = 8, |I*| = 8;
(3). Rank A*[R, J*] = 0, |J*| = 0;
(4). Rank A*[I*, J*] = 0, |I*| + |J*| - k= 0.

The results are in accord with the conditions (r1)-(r4) too. So the structural index is equal to differential index.

V. CONCLUSION

The mathematical modeling and simulation based on the Modelica language often gets a high-index DAE system. It is necessary to translate it into a low-index DAE system that can be directly solved by index reduction method. However, the structural index reduction algorithms will fail in some cases. It is because that the structural index of the DAE system is not equal to the differential index. The combinatorial relaxation algorithm is an effective way to detect and correct the failure of the structural index reduction algorithms. The process usually consists of the following stages:

Phase 1: Compute $\hat{\delta}_k(A)$ by solving the weighted bipartite matching problem.

Phase 2: Test whether $\hat{\delta}_k(A)=\delta_k(A)$ or not. If so, then return $\hat{\delta}_k(A)$ and stop.

Phase 3: Modify $A(s)$ to another matrix pencil $\widetilde{A}(s)$ that satisfies $\delta_k(\widetilde{A})=\delta_k(A)$ and $\hat{\delta}_k(\widetilde{A})<\hat{\delta}_k(A)$. Put $A(s)$:= $\widetilde{A}(s)$ and go to Phase 1.

We implement the combinatorial relaxation algorithm by the combinatorial relaxation theory, matrix algorithms, matroid algorithms and graph theory algorithms etc. The algorithm can improve the stability of the index reduction algorithm based on the structural index in DAE.

REFERENCES

[1] Modelica. http://www.modelica.org.

[2] Fritzson P.Principles of object-oriented modeling and simulation with modelica 2.1. New York:IEEE Press,2003.

[3] Peter Fritzson.Modelica meta-programming and symbolic transformat ions. http://www.ida.liu.se/~adrpo/omc/omdev/mingw/MetaModelica -UsersGuide.pdf,2007.

[4] Dymola. http://www.dymola.com.

[5] OpenModelica. http://www.openmodelica.org.

[6] Gear WC. Differential-algebraic equations index transformations. SIAM Journal on scientific and statistical computing, 1988, 9(1):39~47.

[7] Pantelides C. The consistent initialization of differential-algebraic systems. SIAM Journal on scientific and statistical computing, 1988, 9(2):213~231.

[8] Mattsson SE, Soderlind G. Index reduction in differential-algrbraic equations using dummy derivatives. SIAM Journal on scientific and statistical computing, 1993, 14(3):677~692.

[9] Kazuo Murota.Matrices and matroids for systems analysis(algorithms and combinatorics 20).Berlin:Springer,2010.

[10] Kazuo Murota, Combinatorial Relaxation Algorithm for the Maximum Degree of Subdeterminants: Computing Smith-McMillan Form at Infinity and Structural Indices in Kronecker Form. Applicable Algebra in Engineering, Communication and Computing, 1995,6:251-273.

[11] Satoru Iwata. Computing the Maximum Degree of Minors in Matrix Pencils via Combinatorial Relaxation. Algorithms, 1999, 476–483.

An Improved Uncertainty Measure Method of Rough Sets

Yu Hua-jiao
School of Internet of Things Engineering
Jiangnan University
Wuxi, Jiangsu, China
jiao8709@126.com

Leng Wen-hao
School of Internet of Things Engineering
Jiangnan University
Wuxi, Jiangsu, China
lwh@sytu.edu.cn

***Abstract*—The rough entropy always reduce with the subdivision of knowledge particles when it is used to measure the uncertainty of rough sets, no matter what the knowledge particles come from positive domain, negative domain or boundary domain. so the method rough entropy can not reflect the relationship between uncertainty and knowledge particles well. To solve this problem, an improved method based on rough entropy is proposed. This method is proved that it satisfies the basic standards and extended standards of uncertainty measure of rough sets. Then come to a conclusion that the uncertainty of rough sets will reduce only when knowledge particles which belong to boundary domain are subdivided. And this improved uncertainty method is applied into the fuzzy-rough algorithm based on fuzzy-rough theory.**

Key words-rough entropy; rough sets; boundary domain; uncertainty; roughness

I. INTRODUCTION

The problem of uncertainty mainly be used to study how to manage uncertain information and data effectively in order to find their internal knowledge and rules. The theory of rough sets raised by professor Pawlak in Poland Warsaw university of science and engineering in 1982 is a theory method that study incomplete、uncertain knowledge and the expression、learning、induction of data.

The uncertainty measure method is an important subject in the theory research of rough sets, currently there are several main methods: roughness、rough entropy、fuzzy degrees and fuzzy entropy. However with the subdivision of knowledge particles, the variation of roughness、rough entropy、fuzzy degrees and fuzzy entropy showed defects in their respective, for example, when the knowledge particles come from positive domain or negative domain are subdivided, the roughness remain unchanged, when the knowledge particles come from boundary domain are subdivided ,the roughness may also remain unchanged nevertheless, the subdivision of knowledge particles in boundary domain should make the uncertainty of rough sets reduce; Another example is rough entropy will reduce with the subdivision of knowledge particles in the negative domain, but the negative domain particles should not influence the uncertainty of rough sets.

To overcome the defects of rough entropy, this paper proposed an improved roughness definition and a entropy definition, and on this basis, produced a improved uncertainty measure method of rough sets based on entropy, which highlight the boundary domain's influence to the uncertainty of rough sets. This improved rough entropy will Drab diminishing with the boundary particles are subdivided, and remain unchanged when the knowledge particles in the positive domain or negative domain are subdivided, this result is Accord with cognitive laws, consequently this improved method is a more intuitive and reasonable way to measure the uncertainty of rough sets.

II. BASIC CONCEPT OF ROUGH SETS THEORY

Set $K=(U,R)$ is a approximation space, $U=\{x_1,x_2,\cdots,x_n\}$ is a non-empty finite set ,which called universe of discourse ; R is a equivalence relation or unrecognized relationship on U ,$U/R=\{R_1,R_2,\cdots,R_n\}$ is a partition of U .An information system is quaternion $S=(U,A,V,f)$, $A=\{a_1,a_2,\cdots,a_m\}$ is attribute set , $V=\bigcup_{a\in A}V_a$, V_a is range of attribute a , $f_a:U\to V_a$ is information function. Arbitrary $B\subseteq A$ correspond the unrecognized relationship:
$IND(B)=\{(x,y)\in U\times U\mid \forall_{a\in B}(f_a(x)=f_b(x))\}$,generated partition called U/R. Quaternion $S=(U,A,V,f)$ can

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.24

be described by the decision table below:

TABLE I. THE DECISION OF QUATERNION

A / U	C			D
	C_1	C_2	C_3	D_1
x_1	1	1	2	1
x_2	2	1	3	1
x_3	1	1	3	1
x_4	0	0	2	0

$A = C \cup D$ is the collection of all attribute, C is conditional attribute set, means characteristics of objects , D is decision attribute set, means classes of objects, and $C \cap D = \phi$.

The partition of a rang constitute a approximation space of rough sets, every block in the partition is called a knowledge particle. The change of knowledge particle will lead to the uncertainty of rough set make changes.

Set (U,R) as Pawlak approximation space， for arbitrary set $X \in U$， which is called a concept in U ,the following definition：

$$\overline{R}(X) = \{x \in U \mid [x]_R \cap X \neq \phi\} = \cup\{[x]_R \in U/R \mid [x]_R \cap X \neq \phi\} \quad (1)$$

$$\underline{R}(X) = \{x \in U \mid [x]_R \subseteq X\} = \cup\{[x]_R \in U/R \mid [x]_R \subseteq X\} \quad (2)$$

$$BN(X) = \overline{R}(X) - \underline{R}(X) \quad (3)$$

$\overline{R}(X)$ is called upper approximation of X in (U,R), $\underline{R}(X)$ is called lower approximation of X in (U,R), $BN(X)$ is called boundary domain of X in (U,R). When $\overline{R}(X) = \underline{R}(X)$， that is $BN(X) = \phi$， then says that X is certain in (U,R), otherwise says that X is rough in (U,R), that is said there are uncertain factors in the set.

The domain of discourse is divided into three parts by upper approximation and lower approximation:

Positive domain $P(X) = \underline{R}(X)$, all the objects in this part are must belong to set X ,thus these objects are certain.

Negative domain $N(X) = U - \overline{R}(X)$, all the objects in this part are must not belong to set X, thus these objects are also certain.

Boundary domain $BN(X)$, the object in the boundary domain could be part of set X, or perhaps be not belong to set, therefore these objects are uncertain. Domain of discourse $U = P(X) \cup N(X) \cup BN(X)$.

Pawlak defined the rough precision of set X as:

$$\alpha_B(X) = \frac{|\underline{R}(X)|}{|\overline{R}(X)|} \quad (4)$$

the roughness of X:

$$\rho_B(X) = 1 - \alpha_B(X) = 1 - \frac{|\underline{R}(X)|}{|\overline{R}(X)|} \quad (5)$$

Set $U/P = \{P_1, P_2, \cdots, P_m\}$ and $U/Q = \{Q_1, Q_2, \cdots, Q_n\}$ are two partitions on U ,if $\forall P_i \in P$ ， $\exists Q_j \in Q$,makes $P_i \subseteq Q_j$,then says that U/Q is a fine differentiate space of U/P ,write down for $U/Q \preceq U/P$； if $U/Q \preceq U/P$,and $\forall P_i \in P, \exists Q_j \in Q$,makes $P_i \subset Q_j$,then says that U/Q is a strict fine differentiate space of U/P ,write down for $U/Q \prec U/P$.[3]

Set $U = \{x_1, x_2, \cdots, x_n\}$, $U/B = \{X_1, X_2, \cdots, X_m\}$ is the partition of attribute set $B(B \subseteq A)$ on domain of discourse， $X \subseteq U$ ， then the entropy of attribute set B is:

$$E(B) = -\sum_{i=1}^{m} \frac{|X_i|}{|U|} \log_2 \frac{1}{|X_i|} \quad (6)$$

the rough entropy of X on the partition U/B is defined as：

$$E_B(X) = \rho_B(X) E(B) \quad (7)$$

III. THE DESCRIPTION ABOUT UNCERTAINTY MEASURE METHOD

A. Uncertainty Measure Standards

The uncertainty measure method of rough set reflects the uncertainty of rough sets, the standards of uncertainty measure of rough sets are used to evaluate the measure methods

Set (U,R) as a Pawlak approximation space, X is a subset of U , if mapping function $\varepsilon : P(U) \to R$ meets the following conditions:

1) Nonnegative: $\varepsilon(X) \geq 0$；

2) Invariance: R_1 and R_2 are two equivalent relations on U ,if $R_1 \approx R_2$,then $\varepsilon_{R_1}(X) = \varepsilon_{R_2}(X)$；

3) Limit monotonicity： R_1 and R_2 are two equivalent relations on U ,if $R_1 \prec R_2$,then $\varepsilon_{R_1}(X) \leq \varepsilon_{R_2}(X)$ .And called $\varepsilon : P(U) \to R$ satisfies the basic standards of uncertainty measure of rough sets.

If $\varepsilon : P(U) \to R$ also meet the following conditions:

1) if $\exists_{x \in U}([x]_{R_2} \cap X \neq \phi \wedge [x]_{R_2} \cap X \neq [x]_{R_1} \wedge ([x]_{R_1} \subseteq X \vee [x]_{R_1} \cap X \neq \phi))$ then $\varepsilon_{R_1}(X) < \varepsilon_{R_2}(X)$；

2) if $\forall_{x \in U}([x]_{R_1} \subset [x]_{R_2} \to ([x]_{R_2} \subseteq X \vee [x]_{R_2} \cap X = \phi))$， then $\varepsilon_{R_1}(X) = \varepsilon_{R_2}(X)$.

So called $\varepsilon : P(U) \to R$ satisfies the extended

standards of uncertainty measure of rough sets[6].

B. Improved Uncertainty Measure Method

Rough entropy $E_B(X)$ is a method that satisfying the basic standards, but not satisfying the extended standards. This paper proposed a proved measure method based on Rough entropy, its definition is:

Set (U,R) as a Pawlak approximation space, $U/B=\{X_1,X_2,\cdots,X_m\}$ is the partition of attribute set $B(B\subseteq A)$ on domain of discourse, $X\subseteq U$, then the improved rough entropy of X on the partition is defined as:

$$E_B'(X)=\rho_B'(X)E'(B) =\frac{|X-\underline{R}(X)|}{|BN_B(X)|}\sum_{i=1}^{m}\frac{|X_i\cap BN_B(X)\cap X|}{|U|}\log_2\frac{1}{|X_i|} \quad (8)$$

The following content proves that the improved measure method satisfies the basic standards of uncertainty measure of rough sets:

1)

a) if $BN_B(X)=\phi$, means X is certain in (U,R), $X_i\cap BN_B(X)\cap X=\phi$, then $E'(B)=-\sum_{i=1}^{m}\log_2\frac{1}{|X_i|}\geq 0$, $\overline{R_B}(X)=X=\underline{R_B}(X)$, $\rho_B'(X)=\frac{|X-\underline{R}(X)|}{|BN_B(X)|}=0$;

b) if $BN_B(X)\neq\phi$, $\phi\leq X_i\cap BN_B(X)\cap X\leq X_i$, then $0\leq-\sum_{i=1}^{m}\frac{|X_i\cap BN_B(X)\cap X|}{|U|}\log_2\frac{1}{|X_i|}\leq-\sum_{i=1}^{m}\log_2\frac{1}{|X_i|}$, $0\leq\rho_B'(X)=\frac{|X-\underline{R}(X)|}{|BN_B(X)|}\leq 1$;

Thus $E_B'(X)=\rho_B'(X)E'(B)$ meets the nonnegative property.

2) if $B_1\approx B_2$, then $\overline{R_{B_1}}(X)=\overline{R_{B_2}}(X)$, $\underline{R_{B_1}}(X)=\underline{R_{B_2}}(X)$, $BN_{B_1}(X)=BN_{B_2}(X)$, then $\rho_{B_1}'(X)=\rho_{B_2}'(X)$, $E'(B_1)=E'(B_2)$, $E_{B_1}'(X)=E_{B_2}'(X)$, thus $E_B'(X)=\rho_B'(X)E'(B)$ meets the invariance property.

3) if $B_1\prec B_2$, then $\overline{R_{B_1}}(X)\subseteq\overline{R_{B_2}}(X)$, $\underline{R_{B_1}}(X)\supseteq\underline{R_{B_2}}(X)$, $BN_{B_1}(X)\leq BN_{B_2}(X)$, then $\rho_{B_1}'(X)\leq\rho_{B_2}'(X)$, thus $E_B'(X)=\rho_B'(X)E'(B)$ meets the limit monotonicity property.

The following content proves that the improved measure method satisfies the extended standards of uncertainty measure of rough sets:

1) if $\exists_{x\in U}([x]_{B_1}\cap X\neq\phi\wedge[x]_{B_1}\cap X\neq[x]_{B_1}\wedge([x]_{B_1}\subseteq X\vee[x]_{B_1}\cap X\neq\phi))$, that means the partitions in the boundary domain are subdivided.

a) when $BN_{B_2}(X)=\phi$, *then* $E_{B_2}'(X)=\rho_{B_2}(X)E(B_2)=0$, X *is certain in* (U,B_2), $E_{B_2}'(X)<E_{B_1}'(X)$;

b) when $BN_{B_2}(X)\neq\phi$, then $\overline{R_{B_1}}(X)\supset\overline{R_{B_2}}(X)$, $\underline{R_{B_1}}(X)\subset\underline{R_{B_2}}(X)$, $BN_{B_2}(X)<BN_{B_1}(X)$, $\rho_{B_2}'(X)<\rho_{B_1}'(X)$, $X_i(B_2)\cap BN_{B_2}(X)\cap X\leq X_i(B_1)\cap BN_{B_1}(X)\cap X$, thus $E_{B_1}'(X)<E_{B_2}'(X)$.

Therefore $E_B'(X)$ meets the extended standard 1).

2) if $\forall_{x\in U}([x]_{B_1}\subset[x]_{B_2}\rightarrow([x]_{B_2}\subseteq X\vee[x]_{B_2}\cap X=\phi))$, that means the partitions in the positive domain or negative domain are subdivided.

a) When the subdivided partitions are belong to the positive domain, $\overline{R_{B_1}}(X)=\overline{R_{B_2}}(X)$, $\underline{R_{B_1}}(X)=\underline{R_{B_2}}(X)$, $BN_{B_1}(X)=BN_{B_2}(X)$, thus $\rho_{B_1}'(X)=\rho_{B_2}'(X)$; The intersection of X_i and $BN_{B_1}(X)$ is empty set.

b) As the same as the positive domain's partitions are subdivided, when the subdivided partitions are belong to the negative domain, $\overline{R_{B_1}}(X)=\overline{R_{B_2}}(X)$, $\underline{R_{B_1}}(X)=\underline{R_{B_2}}(X)$, $BN_{B_1}(X)=BN_{B_2}(X)$, then $\rho_{B_1}'(X)=\rho_{B_2}'(X)$; The intersection of X_i and $BN_{B_1}(X)$ is empty set, thus $E_{B_1}'(X)=E_{B_2}'(X)$.

Therefore $E_B'(X)$ meets the extended standard 2).

Ex: Set $U=\{x_1,x_2,\cdots,x_n\}$, $U/B_0=\{\{x_1,x_2,x_3,x_4\},\{x_5,x_6,x_7\},\{x_8,x_9,x_{10},x_{11},x_{12}\}\}$, $U/B_1=\{\{x_1\},\{x_2,x_3,x_4\},\{x_5,x_6,x_7\},\{x_8,x_9,x_{10},x_{11},x_{12}\}\}$, $U/B_2=\{\{x_1x_2,x_3,x_4\},\{x_5,x_6\},\{x_7\},\{x_8,x_9,x_{10},x_{11},x_{12}\}\}$, $U/B_3=\{\{x_1x_2,x_3,x_4\},\{x_5,x_6x_7\},\{x_8,x_9\},\{x_{10},x_{11},x_{12}\}\}$, $X=\{x_3,x_4,x_5,x_6,x_7\}$. Their rough entropy and improved rough entropy:

TABLE II. THE COMPARISON OF TWO ROUGH ENTROPIES

	U/B_0	U/B_1	U/B_2	U/B_3
$E_B(X)$	$\frac{2}{3}+\frac{1}{12}(\log_2 3^3+\log_2 5^5)$	$\frac{1}{12}(\log_2 3^6+\log_2 5^5)$	$\frac{5}{6}+\frac{1}{12}\log_2 5^5$	$\frac{5}{6}+\frac{1}{6}\log_2 3^3$
$E_B'(X)$	$\frac{1}{6}$	$\frac{1}{18}\log_2 3$	$\frac{1}{6}$	$\frac{1}{6}$

U/B_1 is the strict fine differentiate of the boundary domain partition of U/B_0, U/B_2 is the strict fine differentiate of the positive domain partition of U/B_0, U/B_{B_3} is the strict fine differentiate of the positive domain partition of U/B_0. $E_{B_1}'(X)<E_{B_0}'(X)$, $E_{B_0}'(X)=E_{B_2}'(X)=E_{B_3}'(X)$ are reasonable result.

IV. APLICATION OF THE IMPROVED METHOD

Fuzzy-rough algorithm based on fuzzy-rough theory is an important method applied to knowledge

acquisition, rough entropy can be used to choose the certain rule. The fuzzy-rough algorithm used the improved rough entropy this paper proposed is described as:

Input: knowledge expression system $S=(U,V,A,f)$

Output：the collection of fuzzy set with uncertainty measure

Steps：

1) According to instructions or concept clustering method ， determine the central point of the fuzzy partition, make the continuous attribute fuzzification. For any object of the discourse of domain, convert all continuous attributes $A_j(1\le j\le N)$ to fuzzy set.

2) According to decision attribute set D，calculate partition U/D.

3) For any condition attribute subsets B, calculate U/B.

4*)* For any classes of X_i of decision class，calculate the upper approximation and the lower approximation ,add the results to the certainty object set and uncertainty object set respectively.

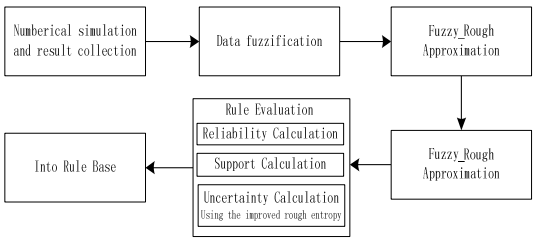
5) Repeat steps *2*)、*3*)、*4*)，until all the condition attribute subset and the decision classes are calculated.

6) Get the certainty rules from the certainty object set and get the uncertainty rules from the uncertainty object set, calculate their reliability 、support, and use the improved rough entropy this paper proposed to calculated the uncertainty.

7) Add these rules to fuzzy rule base after rules reduction.

This algorithm is shown below:

Figure 1. The Flowchart of Fuzzy-rough Algorithm



The data in the rules base can be used in simulation directly.

V. CONCALUTIONS

All sorts of rough set uncertainty measurement method measure the uncertainty of rough set from different ,rough entropy is an important method of them. But the rough entropy method cannot describe the relationship between the uncertainty and the subdivision of the knowledge partition. The uncertainty should be only influenced by the subdivision of partitions come from boundary domain. The basic standards of uncertainty measure of rough sets and the extended standards of uncertainty measure of rough sets are used to evaluate whether a measure method is a good method or not. A improved rough entropy satisfies the basic standards and the extended standards of uncertainty measure of rough sets the is proposed ,when the partitions belong to positive domain or negative domain are subdivided, its rough entropy will remain unchanged, but it will reduce when the subdivided partitions are belong to the boundary domain. Fuzzy-rough algorithm based on fuzzy-rough theory is an important method applied to knowledge acquisition, rough entropy can be used to choose the certain rule that make the final rules have low uncertainty.

VI. REFERENCES

[1] Wang Guoyin . Rough Set Theory and Knowledge Acquisition. Xi'an, China: Xi'an Jiaotong University Press，2001

[2] Pawlak Z. Rough Sets：Theoretical Aspects 0f Reasoning about Data[M]. Boston，USA：Kluwer Academic Publishers ，1991.

[3] Huynh V N，Nakamori Y. A Roughness Measure for Fuzzy Sets.Informatics and Computer Science： An International Journal，2005.

[4] Liang Jiye, Li Deyu. Uncertainty and Knowledge Acquisition in Information Systems.Beijing, China: Scicence Press, 2005

[5] Hujun, Wang Guoyin. Uncertainty Measure Rule Sets of Rough Sets. Pattem Recognition and Artifical Intelligence. 2010, 23(11): 606-615.

[6] Peng Yinghong，Hu Jie. KBE Technology and Application in Product Design of It. Shanghai：Shanghai Jiaotong University Press，2007.

[7] Chakrabarty K，Biswas R，Nanda S. Fuzziness in Rough Sets[J]. Fuzzy Sets and Systems，2000，110(2)：247—251

[8] Xie Bin，Li Leijun，Mi Jusheng. Uncertainty measure of Rough Sets Based on Knowledge Partitions. Computer Science ，2010，9(37)：225-228.

[9] Mi Ju-sheng，Yee Leung，Wu Wei-Zhi. An uncertainty measure in partition-based fuzzy rough sets[J]. International Journal of General Systems，2005，34：77-90

[10] Wang Guoyin，Zhang Qinghua.Uncertainty of Rough Sets in Different Knowledge Granularities.Chinese Journal of Computers, 2008,31(9):1588-1598

2011 10th International Symposium on Distributed Computing and Applications to Business, Engineering and Science

A Rule-based Framework of Metadata Extraction from Scientific Papers

Zhixin Guo, Hai Jin
Cluster and Grid Computing Lab
Services Computing Technology and System Lab
Huazhong University of Science and Technology, Wuhan, 430074, China
E-mail: guozhixin@mail.hut.edu.cn, hjin@mail.hust.edu.cn

***Abstract*—Most scientific documents on the web are unstructured or semi-structured, and the automatic document metadata extraction process becomes an important task. This paper describes a framework for automatic metadata extraction from scientific papers. Based on a spatial and visual knowledge principle, our system can extract title, authors and abstract from scientific papers. We utilize format information such as font size and position to guide the metadata extraction process. The experiment results show that our system achieves a high accuracy in header metadata extraction which can effectively assist the automatic index creation for digital libraries.**

Keywords- document metadata; information extraction; rule-based approach

I. INTRODUCTION

More and more people are used to obtain information through the Internet. Unfortunately, it is difficult for a user to quick access a wide variety of information sources on the web, because those sources are lack of semantic description. Extracting metadata from those sources becomes an important task to help people to access the desired information with semantic associations. However, automatic metadata extraction remains a challenging work. There are no perfect methods to automatically extract metadata from all kinds of documents. Most of the proposed tools for metadata extraction are only applicable to some particular type of documents.

In this paper, we focus on the problems of automatic metadata extraction from scientific papers such as journal articles and conference papers in PDF. The PDF format is quite suitable to be printed or shown on the screen, but it has very limited special descriptions for scientific papers. We present a rule-based framework for the automatic extraction of metadata, such as titles, authors, and abstracts, from scientific papers. We evaluate the performance of the proposed scheme in SemreX system, which is P2P based semantic literature sharing system offering the literature sharing services among computer science researchers [1].

The rest of the paper is organized as follows. Section II overviews the related work about metadata extraction. The software architecture of SemreX is presented in Section III, where the design of the document metadata extraction layer of SemreX is presented in detail. The metadata extraction algorithms are described in section IV. The results of experiments are presented in section V. Finally, we conclude the paper in section VI.

II. RELATED WORK

Most of the current web search engines such as Google.com and Yahoo.com attempt to index the scientific papers using the words and phrases in the documents for users to retrieve. Unfortunately, the abundance of documents leads to the retrieved results with many documents of little worth. Additionally, the use of the entire collection of words, in the document as keys for the index can cause a search to retrieve many irrelevant documents, particularly when the words in the search term are quite common [2]. Library collections such as CiteSeer attempt to solve the issues above. Documents are only included in the collection to be searched if they meet some criteria of relevance. The content of the document is not used as a single collection of equally weighted tokens and one can currently only search for terms in the authors' names, paper title, or body. Thus one can receive a list of valuable, relevant documents from a search. These library collections still have their own shortcomings. They rely on closed collections of documents. Addition of a document to the collection usually requires manual data entry or manual data correction which makes the processing quite slow [3, 4]. So the automatic metadata extraction from electronic documents is of significant importance.

In general, metadata of a scientific paper refers to the metadata from the paper header (the text before the *introduction* or the end of the first page) and the bibliographic fields [5]. There has been much previous work in automatic document metadata extraction. Seymore et al. used *hidden Markov models* (HMM) for the document header metadata extraction [6]. HMM learns a generative model over input sequence and labeled sequence pairs. Based on discriminatively trained SVM classifiers, Han et al. presented another method to extract the header metadata [5]. In [7], Peng et al. employed Conditional Random Fields for the task of extracting various common fields from the header and citation of research papers. Those methods are all based on machine learning technique. In general machine learning techniques are robust and adaptable and can be used on any document set. However, generating the labeled training data

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.14

is very time-consuming and costly.

Other researchers use the rule-based method for automatic metadata extraction. A. Kawtrakul and C. Yingsaeree presented a framework for automatic metadata extraction from electronic documents that can be both text documents and images of paper documents to ease metadata creation process [8]. H. Han et al. introduced a domain rule-based word clustering method for cluster feature representation to improve the classification performance of document header lines and bibliographic fields [9]. Y. Liu et al. presented an automatic table metadata extraction algorithm and tested it on PDF documents [10]. CiteSeer also can locate, download and parse scientific papers to extract citations from those documents in order to produce the citation index, but it does not extract other useful information such as title, authors, and abstract. The system introduced in [11] is the most similar one to our research. But it only copes with PostScript Files which format is quite different from PDF. Differently, our SemreX system extracts document metadata from both the header and bibliographic fields to get deeper information about the document, and stores the extracted information in a semantic database to assist the semantic search process.

III. SOFTWARE ARCHITECTURE

SemreX is a software system, including document metadata extraction, semantic classification of papers, semantic representation, syntax and semantic query, and P2P communications. Fig. 1 shows the system architecture of SemreX.

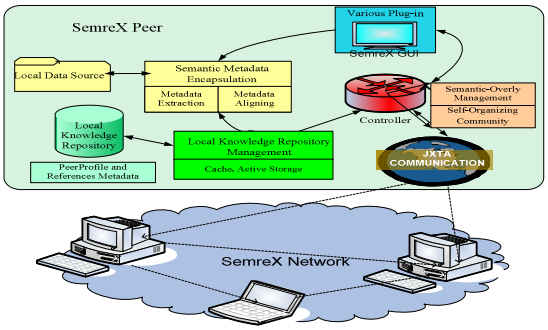


Figure 1. Architecture of SemreX

In this paper, we will introduce the extraction part of SemreX and give a view of the implementation of document metadata extraction. Fig. 2 shows the logical flow of document metadata extraction in SemreX. The crawlers search scientific papers on the web and when a crawler find a paper seems to be a scientific paper in PDF, it records the URL information of the paper and downloads the paper immediately. Then the File Conversion Module converts the file from PDF to other format. The File Identification Module uses the converted text file to estimate whether this paper is a scientific paper. If it is not, the system ignores it; otherwise the system starts the Metadata Extraction Module to extract the document metadata of this paper, saves the extracted metadata in the format of OWL, and stores the information in the ontology database to assist further semantic search.

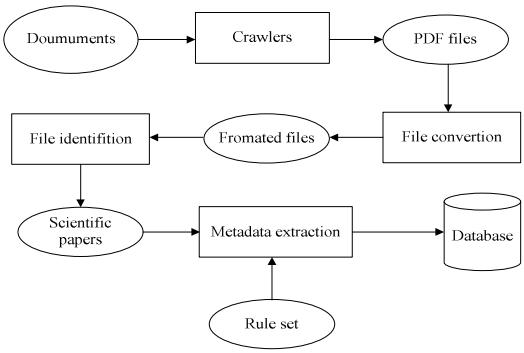


Figure 2. Logical flow of document metadata extraction

A. File Conversion

In order to process the documents and find the elements our system needs to index them and continue scanning the web. Before extraction, we need convert those downloaded PDF files to other files in different format which can be easier dealt with. PDF files make use of inner objects to describe information such as text, image, table. The Adobe Systems Incorporated has already released the Adobe Acrobat SDK for the developers to create software and plug-ins to interact and customize Acrobat. Those software libraries and components can help researchers to read and covert PDF files.

There are a lot of converting applications which have been released to handle the PDF files. According to our system, we try to convert the source PDF files to two kinds of files which are text files and XML files. The text files contain all the text extracted from the source files and they do not carry any format information. On the other hand, the XML files have the format information about the characters and paragraphs in the papers, such as the font size, the layout. We hope with the help of format information, the extracted metadata could be more accurate. Fig. 3 shows an example of a converted XML file.

```
<Page number="1" height="792.8394"
width="613.4392">
<Options> granularity=word </Options>
<Structure>
<Font name="TimesNewRomanPSMT" size="23.1" />
<Word box="[99.6 666.8 156.9863 689.9]"
textrendering="3">
<Text> Policy </Text>
</Word>
<Word box="[162.7994 666.8 262.0251 689.9]"
textrendering="3">
<Text> Evaluation </Text>
</Word>
<Word box="[266.6 666.8 293.4159 689.8]"
textrendering="3">
<Text> for </Text>
</Word>
```

Figure 3. An example of converted XML file

We can acquire the following important information from the source PDF file:

1) Position: the position of each word described as box="[*X1 Y1 X2 Y2*]" where *X* means *x*-coordinate and *Y* means *y*-coordinate.

2) Content: all extracted text of the PDF file.

3) Font type: the font type of each word.

4) Font size: the font size of each word.

5) Page: the page number that the word locates at.

Those characters can help us to identify whether this paper is a scientific paper and extract the header metadata and bibliographies more accurately.

B. File Identification

Our system tries to extract metadata from scientific papers in PDF. Since those PDF files are downloaded from the web, it is essential for us to identify whether those documents are scientific papers. Scientific papers are usually regular papers published in journals or proceedings. Although papers in different journals or proceedings have different format requirements, they may have some common characters in the format. We define several rules to identify the scientific papers. Those rules based on the analysis of the converted text files from sources PDF files are summarized in the following:

1) If the number of lines in the text file is less than 10, the origin PDF file is not a scientific paper.

2) If there are words like *IEEE*, *ACM*, or *PROCEEDINGS* in the first page, it should be a scientific paper.

3) If there are words like *abstract* or *introduction* in the first two pages, or *references* in the last two pages, it should be a scientific paper.

The first rule is to prevent the cases that some papers in PDF are made up of pictures. When those papers are put to the file conversion module, the text files we get only contain few words although they may really be the scientific papers. The other rules use some keywords to identify scientific papers. Although those rules are very simple, we find they are useful to identify the scientific papers in the experiments.

C. Metadata Extraction

When a paper is identified to be a scientific paper, it will be sent to the metadata extraction module. Then the document metadata including the header information and bibliographies will be extracted. The header information includes title, authors and abstract and the extract algorithms will be discussed in the next section. In our early works, we have presented the bibliographies extraction algorithm [12]. So in this paper, we only focus on the document metadata extraction in the header. Use those metadata extracted from the scientific papers, we can know the relations between those papers and authors such as whether those papers are written by the same author or whether the paper is cited by other authors. The extracted metadata could be helpful for the information retrieve in the next step.

IV. METADATA EXTRACTION ALGORITHMS

In our system, we propose a spatial knowledge based approach for document extraction. Our approach is based on exploiting the visual and spatial knowledge humans make use of when reading a document. In general, a certain visual layout can be identified for all documents within that category. But in scientific papers, the layout may be quite different due to lots of kinds of documents. For example some papers are formatted as two columns while others are formatted as only one column. After investigating hundreds of scientific papers, we define several rules to identify the layout and extract the metadata from scientific papers. Those rules which make up of our knowledge base cope with scientific papers appearing on conference proceedings and specialized journals in computer domain. Some spatial properties on the layout of a scientific paper are listed as follows:

1) The header information including title, authors and abstract are located at the first page.

2) The title is located on the upper portion of the first page and it usually uses the largest font on the first page.

3) Authors are listed under the title.

4) Affiliations follow the authors' list.

5) Abstract appears between titles and the main body of the paper and has a certain prefix such as *abstract*.

Those rules listed above work well in metadata extraction process. But we also find some exceptions as follows:

1) In some papers, the first letter of the main body uses the largest font on the first page.

2) Not all abstracts have the prefix such as *abstract*.

3) Sometimes authors are all listed on the same line while other times they are scattered across different lines.

4) Sometimes the affiliations follow all the authors' names while other times each author is spatially closer to his or her affiliation.

The exceptions like the ones listed above, are the hardest part of the metadata extraction. We spend lots of time to find those exceptions. Once we find one exception, we add a rule to our knowledge base. As the number of rules increases, our system can handle more kinds of scientific papers and the accuracy is also increased. Based on those rules we propose the extraction algorithms for title, authors and abstract extraction.

A. Title Extraction

At a first thought, one may think that the first line of text should be the title of the scientific paper; therefore, a text based extraction from those files could be easily applied. Unfortunately, this is not always true. The texts in the first line in some scientific papers are copyright information or the information of proceedings or journals that the paper is published on. So we can not only use the text information converted from PDF files, but also use the layout information such the font and position to extract the title. In scientific papers, the title is usually located at the upper area in the first page and its font size is always largest in the page. But sometimes there are some exceptions where the largest words in the paper are not belong to the title. So we draw out the rules and define the title extraction algorithm shown in Fig. 4. The word set *W1* includes the key words that may be appeared before the title, such as *IEEE*, *ACM*, *proceedings*, *copyright*, *2010*.

```
Input: the converted PDF files in text format f1 and in XML format f2;
the special word set W1
Ouput: the metadata of title
Begin
   f2←getfirstpage(f2);
   t1←getfirstlargestword(f2);
   line1←lineof(t1,f1);
   t2←gettextfromline(line1);
   if ((length(t2)<15) or
      (isintheuppermiddleofthepage(t2,f2)=false) then
       t1←getsecondlargestword(f2);
       line1←lineof(t1,f1);
       t2←gettextfromline(line1);
       if ((length(t2)<15) or
          (isintheuppermiddleofthepage(t2,f2)=false) then
           line1←1;
           while(line1 < 20)
              t2←gettextfromline(line1);
              if (hasword(t2,W1)=true) then
                  line1←line1+1;
               else break;
              end if
           end while
       end if
    end if
   if ((line1=20) or (length(t2)<15)) then
       t2←null;
    end if
  return t2;
End
```

Figure 4. Title metadata extraction algorithm

B. Abstract Extraction

Most of the scientific papers have an abstract in the header and the abstract usually has a certain prefix such as *Abstract:* or *ABSTRACT*. A simple way to extract the metadata of abstract is to find the keyword *abstract* and the text after such keyword in the same paragraph or in the next paragraph should be the abstract of this paper. Unfortunately, there are also some exceptions. Some papers have no abstract, so they do not contain the keyword *abstract*. But in some cases, the scientific paper which does not contain the keyword *abstract* also has an abstract. What is more, some papers which have not an abstract contain the keyword *abstract* in the first page. Based on those facts, the algorithm to extract the abstract is shown in Fig. 5.

```
Input: the converted PDF files in text format f1 and in XML format f2;
the text of extracted title T, the special word set W2
Ouput: the metadata of abstract
Begin
   t1←"abstract";
   if ( hasword(t1, stringtolower(f1)) then
      if ( pageof (t1,f2)=1) and ( lineof(t1,f1) <20 ) and (
          IsfirstwordofParagraph(t1,f1) = true ) then
       t2← strings after t1 in the same paragraph or the next paragraph if
            no strings after t1 in a line;
       end if
   else line1←lineof(T,f1) + 1;
       line2←lineof("introduction",f1)
              or lineof("1.",f1) or lineof("I.",f1);
       if (line2=0) then line2=20; end if
       while(line1 < line2)
          t2←gettextfromline(line1,f1);
          if (hasword(t2,W2)=true) then
              line1←line1+1;
           else break;
          end if
       end while
   end if
   if (length(t2) < 100 ) then
      t2←null;
   end if
   return t2;
End
```

Figure 5. Abstract metadata extraction algorithm

The word set *W2* includes the key words that may be appeared in the affiliations, such as *department*, *faculty*, *university*, *email:*, *@*, *China*. The size of *W2* is much larger than *W1* because of the large number of styles of affiliations.

C. Authors Extraction

After the title and the abstract metadata are extracted from the source PDF file, the next step is to extract the authors' names. Fig. 6 shows the authors extraction algorithm.

In all the scientific papers, the location of the authors' names is always after the paragraph of the title and is before the paragraph of the abstract. So the texts that contain the authors' names can be located soon. The difficulty to extract authors' names is that the authors' names usually appear together with their affiliations. Fortunately we find that the affiliations are usually separated with authors' names using some punctuation such as commas or they are not in the paragraphs where the authors' names locate at. Since the author's affiliation generally introduces the information of the department, university, email, country or postcode of the author, the affiliation has some special keywords that should not be the name of the author, such as *department*, *university*, *email*, or *@*. In our system, we define nearly one hundred keywords that should be part of affiliations. Using those words, we can identify the authors' names from their affiliations.

```
Input: the converted PDF files in text format f1; the text of extracted
title T and abstract A, the word set W2
Ouput: the metadata of authors
Begin
    t1←null;
   if ( T=null) then return null;
      else line1←lineof(T,f1) + 1;
   end if
   if ( A=null) then line2←20;
      else line2←lineof(A,f1);
   end if
  while(line1 < line2)
     while (not reach the end of the line1)
        t2← the next strings before the punctuation as comma;
        if ((hasword(t2,W2) = false ) and (hasnumbers(t2) = false) and
              (length(t2)>4) )  then
           t1←t1+" ,"+t2;
        end if
     end while
    line1←line1+1;
  end while
  return t1;
End
```

Figure 6. Authors metadata extraction algorithm

V. Experimental Results

We test our system over a set of 500 scientific papers in computer domain downloaded from digital libraries of IEEE, ACM, and LNCS. Those papers are all from conference proceedings or journal papers. The test papers are all in PDF and could be converted from PDF files to text files and xml files. We use our system to extract such metadata as title, authors and abstract. For each paper in our sample we test whether the automatically extracted metadata are correct. Such test is based on manual verification of the correctness of the extracted metadata against the original paper. Because this process is tedious and boring, we have to limit our sample size. Table I shows the accuracy results of our test.

TABLE I. Metadata Extraction Accuracy

Metadata	Accuracy		
	IEEE	***ACM***	***LNCS***
Title	89%	92%	90.5%
Authors	84.5%	87.6%	87%
Abstract	93%	95%	92%

From the results we find that the extract accuracy of abstract is the highest because the abstract of a scientific paper usually has some obvious prefix like *abstract:* or *abstract*. The failures of extraction of abstract in the test are due to the absences of such words and the affiliations, abstract and introduction of the papers are joint together which make it hard to be identified by the system. The accuracy of the metadata of authors' names is the lowest due to the difficulty to identify the names from affiliations and in our system the extraction of authors' names also depends on the information of title and abstract. From the comparison we also know that the accuracy of metadata in ACM is higher than the others because the format of ACM is more regular.

VI. Conclusions

In this paper, we propose a rule-based metadata extraction system for scientific papers. The metadata extracted by our system can help people find the papers they need quickly and set users free from the pains in reading through the huge amount of documents. One problem about our system is that it often fails to extract metadata from papers that have somewhat complicated structures. We are now trying to revise our system and make it more robust to retrieve desired information successfully. A wide variety of documents types can also be considered in addition to the PDF papers. Those are our avenues for ongoing and future research.

References

[1] H. Jin and H. Chen, "SemreX: efficient search in semantic overlay for literature retrieval," *Future Generation Computer Systems*, Vol.24, 2008, pp.475-488.

[2] S. Bradshaw and K. Hammond, "Automatically indexing documents: content vs. reference," *Proceedings of the International Conference on Intelligent User Interfaces*, 2002, pp.180-181.

[3] S. Lawrence, C. L. Giles, and K. Bollacker , "Digital libraries and autonomous citation indexing," *IEEE Computer*, Vol.32, 1999, pp.67-71.

[4] H. Li, I. Councill, W. Lee, and C. L. Giles, "CiteSeerx: an architecture and web service design for an academic document search engine," *Proceedings of the International Conference on World Wide Web*, 2006, pp.883-884.

[5] H. Han, C. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. Fox, "Automatic document metadata extraction using support vector machines," *Proceedings of Joint Conference on Digital Libraries*, 2003, pp.37-48.

[6] K. Seymore, A. McCallum, and R. Rosenfeld, "Learning hidden markov model structure for information extraction," *Proceedings of the AAAI'99 Workshop on Machine Learning for Information Extraction*, 1999, pp.37-42.

[7] F. Peng and A. McCallum, "Accurate information extraction from research papers using conditional random fields," *Proceedings of Human Language Technology Conference*, 2004, pp.329-336.

[8] A. Kawtrakul and C. Yingsaeree, "A unified framework for automatic metadata extraction from electronic document," *Proceedings of the International Advanced Digital Library Conference*, 2005, pp.71-77.

[9] H. Han, E. Manavoglu, H. Zha, K. Tsioutsiouliklis, C. L. Giles, and X. Zhang, "Rule-based word clustering for document metadata extraction," *Proceedings of ACM Symposium on Applied Computing*, 2005, pp.1049-1052.

[10] Y. Liu, K. Bai, P. Mitra, and C. L. Giles, "TableSeer: automatic table metadata extraction and searching in digital libraries," *Proceedings of the ACM Conference on Digital Libraries*, 2007, pp.91-100.

[11] G. Giuffrida, E. C. Shek, and J. Yang, "Knowledge-based metadata extraction from postscript files", *Proceedings of the ACM Conference on Digital libraries*, 2000, pp.77-84.

[12] Z. Guo, H. Jin, and H. Chen, "Semantic document reference metadata extraction in SemreX", *Journal of Computer Research and Development*, Vol.43, 2006, pp.1368-1374 (in Chinese).

Dynamic Sliding Mode Control Applied in Parallel Robot System

Zhu Caihong, Zhang Hongtao
Department of Electronic Information Engineering
Suzhou Vocational University
Suzhou, China
zhuzhang_2000@yahoo.com.cn

Abstract—**Parallel robot has many good features, such as high rigidity, great load and high precision. Directed against the parallel robot mechanism with AC servo-motor drive, a model of controlling system was established. A kind of dynamic sliding mode control algorithm was designed, and the stability of this control algorithm was analyzed. A simulative experiment of trajectory tracking was made on the Matlab / Simulink. The simulation result was given to prove the validity of the variable structure controller, and a good performance in tracking, and the high-accuracy real time control of the parallel robot mechanism is implemented.**

Keywords- parallel robot; sliding mode control; servo motor; trajectory tracking

I. INTRODUCTION

Compared with the serial robot, the parallel robot has the stiffness, load capacity, high precision, compact structure, and can be widely used in industrial, aviation, military and other fields [1]. In recent decades, the characteristics, the institutional learning, and the kinematics aspects of the parallel robot is researched in a wide range of in-depth by domestic and foreign scholars. Parallel robot is a complex, multi-variable, multi-degree of freedom, multi-parameter nonlinear coupling system, and the control strategy is extremely complex. When the control system originally is designed, mostly, the various branches of the parallel manipulator are taken as a completely independent control system, the control strategy is the traditional PID control strategy for the control, and the control results are not very satisfactory. Fuzzy control method can realize robot's control, does not require the precise model of the robot, but the fuzzy control design of fuzzy rules is more important, the quality of rule design will directly affect the control results [2], and that the rule set will be obtained by expert knowledge or after several trial, therefore, in the absence of appropriate conditions, this method may not play a better control effect.

Sliding mode control is essentially a special class of nonlinear control, and the nonlinear performance of the control is a discontinuity. The difference between this control strategy and other control strategies is that the "structure" of the system is not fixed, but in the dynamic process of the current state of the system (such as the deviation and its derivative, etc.) are ever-changing destination, and force the system in accordance with the state trajectory of the scheduled "sliding mode"[3]. Studies have shown that [4-6]: Sliding mode control owns many merits of fast response, insensitivity to parameter changes and disturbances, no system of online identification, physical implementation and simple. The sliding mode control is a comparison method for robot control [7-8], a new variable structure path tracking control method is used, it is a dynamic sliding mode control method based on dynamic switching function. That is, the switching function s through new design or from the conventional sliding mode control constitutes a new switching function σ through the differential link. The switching function is related with the first order or higher order derivatives of the system input. you can transferred to the continuous control items to a first-order or higher order derivatives of the control, then get the nature continuous dynamic sliding mode control law in time, and can effectively reduce the chattering.

II. DESIGN OF THE SLIDING MODE CONTROLLER

The sliding mode controller design requires the following work:

(1)Striking the switching function $s(x)$; (2)Guarantee the existence of sliding mode; (3)Determine the stability of sliding mode; (4)Variable structure robust control approach and dynamic phase of quality assurance; (5)seeking the variable structure control $u^{\pm}(x)$.

Considering a single IO n order affine nonlinear system which is following:

$$\begin{cases} \dot{x}_i = x_{i+1}, i = 1,2,\cdots,n-1 \\ \dot{x}_n = f(x) + g(x)u + \eta \\ y = x_1 \end{cases} \tag{1}$$

In the formal, $x \in R$ is an observable state variable; u, $y \in R$ are the input and output of the system; $f(x)$ and $g(x)$ are the knowing smoothing functions; η is an uncertain item including the model uncertainty and external disturbance[9].

Error function and switching function are following:

$$\begin{cases} e = y - y_d \\ s = c_1 e_1 + c_2 e_2 + \cdots + c_{n-1} e_{n-1} + e_n = \sum_{i=1}^{n-1} c_i e_i + e_n \end{cases} \tag{2}$$

In the formal, $e_i = e^{(i-1)}$(i=1, 2, …, n) is tracking error and its all-order derivatives. Constant c_1, c_2, …, c_{n-1} is selected to

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.39

guarantee that the multinomial $p^{n-1}+c_{n-1}p^{n-2}+\ldots+c_2p+c_1$ is *Hurwitz* stable. p is a *Laplace* operator. Then:

$$\dot{s} = f(x) + g(x)u + \eta - y_d^{(n)} + \sum_{i=1}^{n-1} c_i e_{i+1} \tag{3}$$

New switching function:

$$\sigma = \dot{s} + \lambda s \tag{4}$$

In the formal, λ is a strict normal number.

When σ =0, $\dot{s}$ +λs=0 is an asymptotically stable one order dynamic system, s is close to zero.

Hypothesis 1: The uncertainty meets the boundary condition, a boundary function $B_n(x)$ is existed, such that

$$|\eta| \le B_n(x) \quad \forall\, x \in R^n$$

and the symbol of g(x) is a constant.

Hypothesis 2: The derivative of the uncertainty has boundary.

$$|\dot{\eta}| \le \overline{B}_n(x) \quad \forall\ x \in R^n \tag{5}$$

Hypothesis 3: A positive real number ε is existed, and can satisfy the following formal.

$$\varepsilon > (c_{n-1}+\lambda) + \overline{B}_n \tag{6}$$

Therefore, the dynamic sliding mode control law is as follows:

$$\dot{u} = \frac{1}{g}[-((c_{n-1}+\lambda)g + \frac{dg}{dx}\dot{x})u - (c_{n-1}+\lambda)f - \frac{df}{dx}\dot{x} + (c_{n-1} + \lambda)y_d^{(n)} + y_d^{(n+1)} - \sum_{i=1}^{n-2} c_i e_{i+2} - \sum_{i=1}^{n-1} \lambda c_i e_{i+1} - \varepsilon\,\mathrm{sgn}(\sigma)] \tag{7}$$

Stability analysis: Stability is a basic structural characteristic of the system, the stability is an important issue to the theory of control system.

It can be obtained by Equation (2) into (4) :

$$\dot{\sigma} = \sum_{i-1}^{n-1} c_i e_{i+1} + \dot{e}_n + \lambda(\sum_{i-1}^{n-1} c_i e_{i+1} + e_n) \tag{8}$$

Then

$$\dot{\sigma} = \sum_{i-1}^{n-2} c_i e_{i+2} + c_{n-1}\dot{e}_n + \ddot{e}_n + \lambda(\sum_{i-1}^{n-1} c_i e_{i+1} + \dot{e}_n) \tag{9}$$

Equation (1), (2) into (9), finishing

$$\dot{\sigma} = \sum_{i-1}^{n-2} c_i e_{i+2} + \sum_{i-1}^{n-1} c_i e_{i+1} + (c_{n-1}+\lambda)f + \frac{df}{dx}\dot{x} - (c_{n-1}+\lambda)y_d^{(n)} - y_d^{(n+1)} + [(c_{n-1}+\lambda)g + \frac{dg}{dx}\dot{x}]u + g\dot{u} + (c_{n-1}+\lambda)\eta + \dot{\eta} \tag{10}$$

The control law equation (7) into (10)

$$\dot{\sigma} = (c_{n-1}+\lambda)\eta + \dot{\eta} - \varepsilon\,\mathrm{sgn}(\sigma)$$

It can be obtained based on the assumption

$$\sigma\dot{\sigma} = \sigma(c_{n-1}+\lambda)\eta + \sigma\dot{\eta} - \varepsilon|\sigma| = \sigma[(c_{n-1}+\lambda)\eta + \dot{\eta}] - \varepsilon|\sigma| < \sigma[(c_{n-1}+\lambda)\eta + \dot{\eta}] - [(c_{n-1}+\lambda)B_n + \overline{B}_n]|\sigma| \le 0 \tag{11}$$

By Lyapunov stability analysis, a new dynamic switching function σ is revealed to meet the function $\sigma\dot{\sigma} \le 0$. The sliding mode control theory is met the reaching condition, which verifies the stability of the system, and will ensure the robustness and the dynamic quality of the controller.

III. CONTROL MODEL OF THE PARALLEL ROBOT

The less DOF parallel robot has many characteristics of each simple branch agencies, no more virtual, and big work space is studied in this case [10].

The full Lagrange robot dynamics model can be expressed as the following form:

$$M(q)\ddot{q} + C(q,\dot{q})\dot{q} + G(q) = \tau \tag{12}$$

Here, q and $\dot{q}$ represent the robot position and velocity of each joint, τ is the torque vector. *M*, *C*, *G* were the order function matrix of $n \times n$, $n \times n$ and $n \times 1$ determined by the specific structure of the robot. τ is a vector-order moment of $n \times 1$. It looks more concise as the style has been written in the form to be linear. In fact, the system started in the form of fairly complex. The distributed control system based on joint or slip on the robot model is one of the most designing applied widely in the actual process, and in industrial engineering, most systems used this type of robot design.

Mutually parallel branches of the parallel robot we researched can be expressed in Figure1.

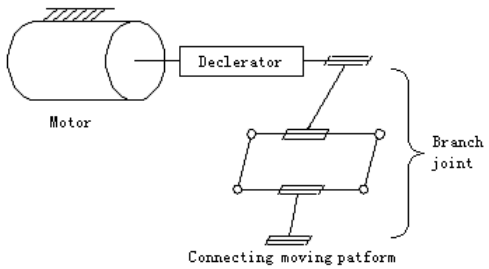


Figure 1. Branch model of robot

The driving device of the robot joint is a AC servo-motor, the equivalent interference torque is neglected, then the mathematical model of the robot various branches may use the following transfer function representation.

$$G(s)=\frac{\theta(s)}{I_d(s)}=\frac{K}{s(j's^2+B's+W')+K_n} \qquad (13)$$

In the formula: $J'=Lp(J_a+J_m+i_2J_0)$; $K=3K_AK_{pre}K_{tp}/2$; $B'=(R_p+K_AK_i)(J_a+J_m+i_2J_0)+L_p(B_m+i_2B_0)$; $K_x=3K^2_{tp}/2$; $W'=(R_p+K_AK_i)(B_m+i_2B_0)$.

IV. SIMULATION AND ANALYSIS

The parameters of the robot servo motor are as follows:

K_{pre}=88, K_i=2.2, K_A=6, K_{tp}=3.41 N·m/A, L_p=0.03837 H, R_p=5.09Ω, J_a=0.19 kg·m^2.

The speed ratio of deceleration devices is i=40. Joint part of the device driver in the slow side for the moment of inertia is 0.1kg·m2. Since the coupling of inter-agency, the equivalent moment of inertia and the factor of the equivalent load resistance of the system is J_0=40, B_0=0, and J_m=0, B_m=0.

The model transfer function of the robot joint based on the AC servo drive is follows:

$$G(s)=\frac{\theta(s)}{I_d(s)}=\frac{0.0125}{s^3+18.29s^2+0.007} \qquad (14)$$

The conversion equation of state is followed:

$$\begin{cases} \dot{x}_1=x_2 \\ \dot{x}_2=x_3 \\ \dot{x}_3=-0.007x_1-18.29x_3+0.0125u+\eta \\ y=x_1 \end{cases} \qquad (15)$$

Here $f(x)$ = -0.007x_1-18.29x_3, $g(x)$ =0.0125.

Let the desired tracking signal is y_d =sin (t), the tracking error is e=x(1)-y_d. When n=3, the definition is s=c_1e+c2 $\dot{e}$ +c3 $\ddot{e}$, and c1=c2=100, c3=1, λ=3000. The initial condition is x (0) = [0.5 0 0], the dynamic control law is:

$$\dot{u}=\frac{1}{g}[-(c_2+\lambda)gu-(c_2+\lambda)f-\dot{f}+(c_2+\lambda)\ddot{y}_d-y_d^{(4)} \qquad (16)$$
$$-c_1\ddot{e}-\lambda(c_1\dot{e}+c_2\ddot{e})-\varepsilon\,\mathrm{sgn}(\sigma)]$$

In the formula ε= (c_2+λ)B_n+ $\overline{B}_n$ +100, B_n+ $\overline{B}_n$ =50.

The simulation model including two S-Function is composed on the Matlab / Simulink software. The simulation use ode45, and the step is 0.001s. The simulation results are shown below.

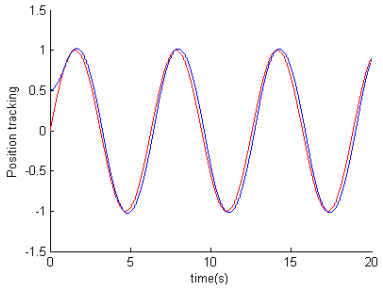


Figure 2. Location tracking

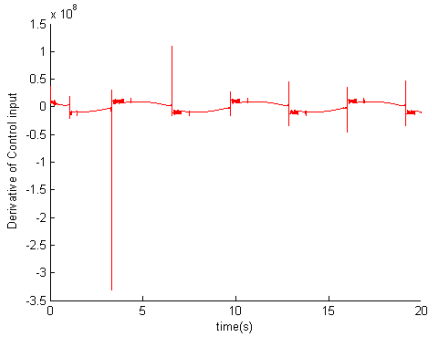


Figure 3. Input control signal $\dot{u}$

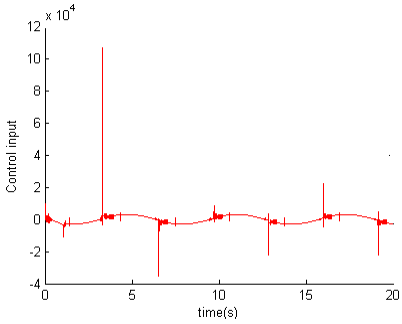


Figure 4. Input control signal u

V. CONCLUSION

The simulation results show that the trajectory tracking of the parallel robot is realized by dynamic sliding mode control based on dynamic switching function. This method of the sliding mode control owns a good anti-interference effect and shadowing property. The study laid the foundation further to achieve the high precision and real-time control on the parallel manipulator.

ACKNOWLEDGMENT

At the point of finishing this thesis, I'd like to take this opportunity to show my sincere gratitude to, Mr. Zhang Hongtao, who has given me so much useful advices on my writing, and has tried he best to improve my thesis. Without his help, it would be much harder for me to finish my thesis.

REFERENCES

[1] PIEPER J. First Order Dynamic Sliding Mode Control[C]. Proceedings of the 37th IEEE Conference Decision & Control, Tampa, Florida USA, 1998, 2415-2420.

[2] HWANG C, HAN S, YU Y. Network-based Fuzzy Decentralized Sliding Mode Control for Car-like Mobile Robots[C]. IEEE Trans. Industrial Electronics, 2007, 54: 574-585.

[3] HU Yueming. Variable Structure Control Theory and Applications[M]. Beijing: Scientific Publishing house, 2003:156-165.

[4] MOON J, KIM K, KIM Y. Design of Missile Guidance Law Via Variable Structure Control [J]. Journal of Guidance, Control and Dynamics, 2001, 24 (4): 659-664.

[5] MATSUNO F, SATO H. Trajectory Tracking Control of Snake Robots Based on Dynamic Mode[C]. IEEE Conf. Robotics and Automation, Apr. 2005: 3029-3034.

[6] MEI Hong, WANG Yong. Fast Convergent Sliding Mode Variable Structure Control of Robot [J].Information and Control, 2009, 38(05):552-557.

[7] WU Bo, WU Shenglin, ZHAO Keding. Current Status and Development Trend of Stewart Platform Control Strategy [J]. Machine Tool & Hydraulics, 2005, 10:5-8.

[8] WANG Hongbin, WANG Hongrui, XIAO Jinzhuang. Study on Trajectory Following of Parallel Robot Using Integral Variable Structure Control [J].Journal of Yanshan University, 2003,27 (1):25-28.

[9] LIU Jinkun. MATLAB Simulation of Sliding Mode Control [M].Beijing: Tsinghua University publishing house, 2005:283-285.

[10] XU Chunshan, SUN Xingjin, CAO Guangyi. A Novel Variable Structure Control for the Tracking of Robot[J].Computer Simulation, 2004, 21(7): 115-118.

Mathematical Model and System Simulation of the Brushless DC Motor

Zhu Caihong, Zhang Hongtao
Department of Electronic Information Engineering
Suzhou Vocational University
Suzhou, China
zhuzhang_2000@yahoo.com.cn

***Abstract*—The brushless DC motor (BLDCM) is a multi-variable and non-linear system. The mathematical model of the BLDCM is founded, the simulation model of the double closed-loop control system is designed using Matlab6.5, and the rotational speed profiles of the electrical machinery are given on the control strategy of traditional PI algorithm and fuzzy-PI algorithm, then the validity of the proposed BLDCM model and the rationality of the control system are all confirmed.**

Keywords- BLDCM; double closed-loop control system; PI algorithm; Fuzzy-PI algorithm

I. INTRODUCTION

It takes the electrical machinery as the mechanical and electrical energy conversion installment, the brushless DC motor is the high-tech product which is the collection electrical machinery and the electronic integration. The BLDCM has the merits of the little volume, the light weight, the high efficiency, the small inertia and the higher control precision, Simultaneously, the BLDCM retained the fine physical characteristics of the ordinary direct current motor, and it widespread application in national economy each domain. The fast development of modern industry requires that the performance of the BLDCM can been enhanced unceasingly [1-3]. Therefore, the research on the control system of BLDCM, which has the speed of response to be quick, the adjustment ability is strong, the control precision is high, have the very vital significance. The mathematical model of the BLDCM is established. The design and the simulation of the fuzzy PI control system are completed. The fuzzy PI control theory sophistication and DSP superiority are further synthesis. In order to construct the fuzzy PI control system of the BLDCM based on DSP, the design and the research of are studied, and the advanced control thought can move toward the production practice from the laboratory.

II. MATHEMATICAL MODEL OF THE BRUSHLESS DC MOTOR

The effect circuit diagram of the BLDCM main body is usually like chart 1. In order to simplify the analysis, the supposition is followed [4-5]: (1) The stator winding for 60° has the full-pitch winding; (2) The slot effect and the saturated magnetic circuit are not considered; (3) The magnetic lag, the turbulent flow, the collection skin effect, and the temperature to the parameter the influence are neglected; (4) The three-phase stator winding is symmetrical, $R_a = R_b = R_c$, $L_a = L_b = L_c$, $M_{ab} = M_{bc} = M_{ca}$.

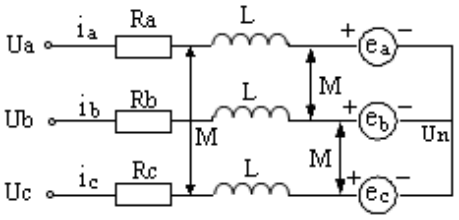


Figure 1. The equivalent circuit diagram of the BLDCM

Like this, each resistance is supposed R, each self inductance is supposed L, two mutual inductances are willfully M, the voltage equilibrium equation may be written:

$$\begin{bmatrix} u_a \\ u_b \\ u_c \end{bmatrix} = \begin{bmatrix} R_s & 0 & 0 \\ 0 & R_s & 0 \\ 0 & 0 & R_s \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} L & M & M \\ M & L & M \\ M & M & L \end{bmatrix} p \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} \quad (1)$$

Also because in the three-phase symmetrical electric motor exists $i_a + i_b + i_c = 0$, so $Mi_a + Mi_b + Mi_c = 0$, therefore after reorganizes:

$$\begin{bmatrix} u_a \\ u_b \\ u_c \end{bmatrix} = \begin{bmatrix} R_s & 0 & 0 \\ 0 & R_s & 0 \\ 0 & 0 & R_s \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} L-M & 0 & 0 \\ 0 & L-M & 0 \\ 0 & 0 & L-M \end{bmatrix} p \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} \quad (2)$$

The movement equation is:

$$T_{em} - T_L - B\omega = J\frac{d\omega}{dt} \quad (3)$$

$$T_{em} = (e_a i_a + e_b i_b + e_c i_c)/\omega \quad (4)$$

In the formula, ω is the angular speed of the electrical machinery rotor. T_L is the load torque; B is the viscous damping coefficient; J is the rotor and load rotor inertia.

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.57

Thus, the completely mathematical model of the three-phase BLDCM has been constituted.

III. SIMULATION DESIGN OF THE CONTROL SYSTEM OF THE BLDCM

A. Control system overall design

Conventional PI control has the merits of simple algorithm, high precision, reliable strong to the control system which can be established the precise mathematical model. But regarding electric motor velocity modulation system, the establishment of the precise model is difficult, therefore the control system's robustness is unsatisfactory. In many control Strategies, the fuzzy control does not rely on the mathematical model of the controlled plant, whose control rule obtained through carrying on the induction and the optimization of the operator's experiences. So, the fuzzy control obtained the widespread application in the commercial process. In this control system, the fuzzy control and the PI control unified [6]. It both has the fuzzy control to be nimble, and compatible strong merit, and has the PI control precision high characteristic. It can improve system's static state and the dynamic property obviously, and has the good control effect.

The fuzzy PI controller's design is the process which a relapse debugs, and the off-line computation load is very big. For saving of time, the computer-aided design is generally used to speed up the design speed, and the Matlab simulation software is used mostly in project application multipurpose.

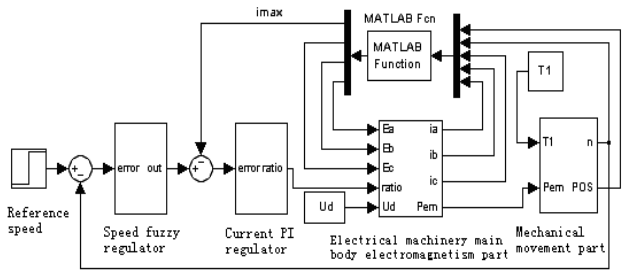


Figure 2. The overall simulation diagram of the BLDCM control system

Under Matlab6.5 Simulink environment, using Simulink and Fuzzy Logic Toolbox rich module storehouse, based on the analysis the mathematical model of the BLDCM, the simulation model of the BLDCM fuzzy PI control system has been established. This control system simulation model uses the double closed-loop control plan. The rotational speed link uses the fuzzy PI regulator constitution. The electric current link constitutes by the traditional PI regulator. Figure 2 is the overall simulation diagram of the BLDCM control system. According to modular model thought, the control system is division for each function independent submodule, such as the main body of the BLDCM, the speed adjustment module, the current adjustment module [7-8], the phase change module, and the mechanical movement module.

B. Speed fuzzy PI control module

The fuzzy PI control used is the increase type method adjusts the PI parameter, which is very difficult to realize under the Matlab Simulink environment. In order to save the simulation to realize the time, here use directly replaces *Kp*, *Ki* value method to realize the PI parameter adjustment. According to the input value, the *Kp*, *Ki* obtained value directly, not the □*Kp*, □*Ki* value. It takes the speed jitter signal *e* and error signal's rate of change □*e* to the input value, *Kp*, *Ki* takes as the output value. The linguistic value and the fuzzy universe of discourse of the input value *e*, □*e* the fuzzy language variable can unification define as {NB，NM，NS，ZO，PS，PM，PB} and [-3,-2,-1,0,1,2,3]. The linguistic value and the fuzzy universe of discourse of the output value *Kp*, *Ki* the fuzzy language variable can unification define as {VS，S，M，L，VL} and [1,2,3,4,5]. The membership function graph takes is a triangle. According to the manufacture principle of the □*Kp*, □*Ki* fuzzy control rule table manufacture principle, the control rule table of the *Kp*, *Ki* is manufactured. The complete speed fuzzy PI adjustment module can be constituted by choice the appropriate input and the output quantification factor finally.

C. Simulation result analysis

The parameter of the simulation electrical machinery model is: tthree-phase 23 antipodes, star connection, rated voltage *Ud* =36*V*, rated speed n_N =300 *r/min*, resistance *R* =1 Ω , inductance *L* =0.0023*H*, mutual inductance *M* =0.0004*H*, rotor inertia *J*=0.005*kg*· m^2, coefficient of potential *Ke*=0.062 *v/rad*·s^{-1}.

The simulation step is specifically as follows: firstly the simulation of the BLDCM electric current double closed-loop control system tradition PI control carried on. In the speed PI link, *Kp*=10, *Ki*=5; in the electric current PI link, *Kp* = 5, *Ki* = 7, the electrical machinery rotational speed simulation curve can be obtained in figure 3. Then, the simulation of the fuzzy PI control system is carried on, in the electric current PI link, *Kp*=5, *Ki*=7, input, output factor, *Ke*=0.01, *K*□*e*=0.005, *Kpu*=1 , *Kiu*=0.5, the simulation curve of the electrical machinery rotational speed is shown in figure 4.

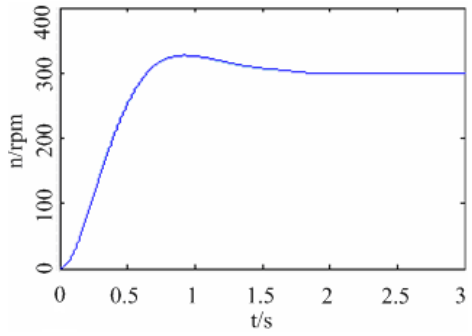


Figure 3. The speed gram of used the traditional PI control

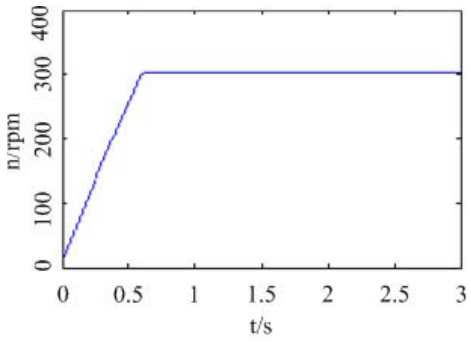


Figure 4. The speed gram of used the fuzzy PI control

The simulation result indicated that speed of response of the control system speeds up, the dynamic property is improved, and without the ultraharmonic vibration and strong robustness after using the fuzzy PI control algorithm. This is that the pure PI control can be realized with difficulty

IV. CONCLUSION

The simulation of the BLDCM double closed-loop control system under the Matlab simulation software environment is carried on in this article. The electric current link uses traditional the PI control, and the speed ring uses the fuzzy PI control. When designing the fuzzy PI controller simulation, the fuzzy control rule's adjustment is easy and feasible, and then fuzzy PI the controller of the fine performance of the BLDCM is designed. The simulation result indicated that the BLDCM control system has the quicker speed of response, and does not have the ultraharmonic vibration by using the fuzzy PI controller compared to the traditional PI controller. In the simulation, the fuzzy control rule has the very good theory instruction function to the BLDCM fuzzy control system's design based on DSP.

ACKNOWLEDGMENT

I would like to acknowledge and extend my heartfelt gratitude to following persons who have made the completion of this thesis possible:

Mr. Zhang Hongtao.

REFERENCES

[1] SHI Hao, PAN Zaiping. The Fuzzy Control System and Simulation of the Brushless DC Motor [J]. Micromotors, 2005, 38 (5):42~44.

[2] CHEN Ya-bing, ZHOU Zhi-ping, JI Zhi-cheng. Study on Fuzzy PI Intelligent Control Strategy of Brushless DC Motor [J]. Journal of Jiangnan University: Natural Science Edition, 2005,4(1):14~18.

[3] WU Xuemei, JING Zhanrong, SHI Yongqi. Study of Control Technology for Brushless DC Electromotor Based on DSP [J]. Machinery & Electronics, 2005 (3):50~52.

[4] JI Zhicheng, SHEN Yanxia, JIANG Jianguo. A Novel Method for Modeling and Simulation of BLDC System Based on Matla [J]. Journal of System Simulation, 2003,15(12):1745~1749.

[5] SUN Jianbo, GONG Shiying, DONG Yahui.Simulation of the Permanent Magnetic Brushless DC Motor Speed-adjustable System[J]. Micromotors, 2001,34 (02):19~23.

[6] LIU Ruifang, YAN Dengjun, HU Minqiang. Field circuit and movement coupled time stepping finite element analysis on permanent magnet brushless DC motors [J]. Proceedings of the CSEE, 2007, 27(12): 59-64.

[7] TANG Renyuan. Modern permanent magnet machines theory and design [M]. Beijing: Mechanical industry publishing house. 1997.

[8] WANG Qunjing. Rare earth neodymium iron boron permanent magnetism synchronous motor's design theory and computer simulation [M]. Beijing: Mechanical industry publishing house. 1997.

System-Level Power Management for Low-Power SOC Design

Zhu Jing-jing
IoT Engineering
Jiangnan University
Wuxi, Jiangsu, China
zhujingjing0918@163.com

Lu Feng
IoT Engineering
Jiangnan University
Wuxi, Jiangsu, China
luf@cetc58.com

***Abstract*—From the system-level of low power consumption view, this article puts forward in software management proposal. Power management unit are introduced. Especially power states transition scheme has been worked out. And the scheme for device power management is elaborate. We applied the method to a complicated SoC and the experimental results prove that this method could reduce system power very greatly with no performance penalty on the chip.**

Key-words: System-on-chip; Low-power design; power management;

I. INTRODUCTION

Along with the development of the integrated circuit technology, chip design scale and integration in unceasing enhancement, power consumption problems in system design become more and more important. Great power will affect the operation of the system speed and stability, shorten battery life, increase chip packages cost. Therefore, low power design has become an inevitable trend of next generation SOC design [1].

In CMOS circuit, dynamic power consumption is the main component of power system, which can be expressed as:

$$P = 1/2 \times \alpha \times C \times V_{DD}^2 \times f \quad (1)$$

Among them: C for load capacitance; VDD voltage; α for flip chance per clock cycle, namely the charging and discharging cycle happens in number; F for access clock frequency. From the formula 1, can from reducing C, decreasing VDD, reasonably managing work frequency three aspects reduce dynamic power throughout the CMOS circuit [2].

According to different levels of abstraction layers, integrated circuit design method can be divided into system level, structural level, register transfer level, door level, territory level and so on the different levels [3]. In view of the different levels, have different low power consumption methods, from the abstract level higher, to consider the power consumption of SoC chip, the greater the amplitude and the higher power optimization efficiency.

This design from the system-level levels consideration, formulate a low-power management plan.

II. SYETERM LEVEL LOW POWER MANAGEMENT PLAN

According to the system's demands, the design and customize a power management processes, general framework is built on the basis of message, communications are basically rely on news way transmission. Application through the message sent to strategy modules way, strategy modules will call power management function module in corresponding function and send the corresponding management command control power management module PMU when receiving message, and then control the hardware unit clock source. Low power consumption management system scheme of frame is as shown in figure 1.

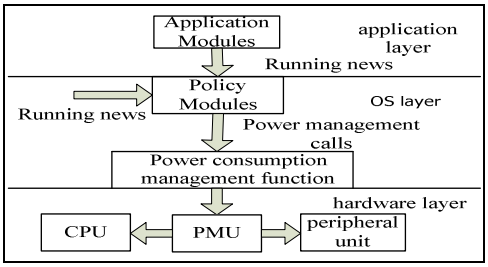


Figure 1 system of low power management scheme frame-chart

978-0-7695-4415-1/11 $26.00 © 2011 IEEE
DOI 10.1109/DCABES.2011.87

This scheme provides within the PMU functions, from the two aspects of CPU and peripherals are analyzed, and respectively formulate a different management strategy, reducing power consumption of the chip.

A. Clock and power management unit (PMU)

SoC system working condition is diversiform, application complex, which can be high speed, also can undertake low-speed operations. According to the formula 1, dynamic power consumption and working frequency is proportional to the relationship; the SoC chip can reduce work frequency decrease dynamic power consumption when in low performance or waiting mode. So scheme added a clock with power management module (PMU) in chip design, through the PMU module, chip can effectively manage resources in the system of layer. This module contains clock control circuit of modules, according to the different application, unified manage clock source supply of the chip internal modules, timely adjust system work frequency, thus reduces the overall power consumption chip.

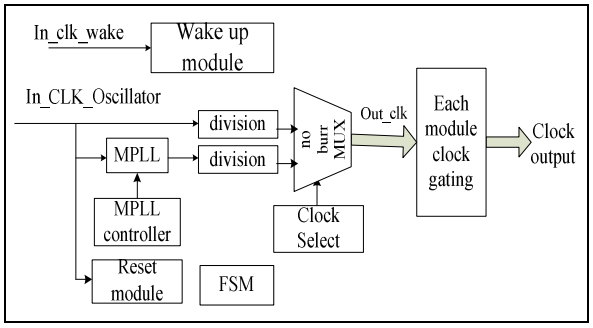


Figure 2 PMU module clock circuit diagram

Where, 5MHz as chip clock crystal excitation source, it provides the whole all modules of the clock SoC source; chip clock can use crystals clock, also can adopt the clock after PLL octave. PLL module receive the 5MHz as input clock, then the input clock for octave to a very high frequency, then frequency division, providing more appropriate frequency clock to CPU and the surrounding IP module. The responsibility of state machine is to perform four models of switching by configuration the system work mode registers (PMDR) and also send out of PLL control signal, change working frequency. MUX selector mainly chooses crystals clocks and PLL clock, in front of MUX, joining latches can eliminate possible burr, the chosen clock also gave to the CPU and IP module, according to the needs of each module, by configurating internal clock source control register (PCSR) complete internal each child module clock source supply management, the gated control sent out the clock.

B. CPU low power consumption

1) CPU work status and conversion

According to the PMU module providing functions, four modes are defined for processor: NORMAL, SLOW, IDLE, and SLEEP. The switching among four models is completed by work mode state machine, switching process as shown in figure 3.

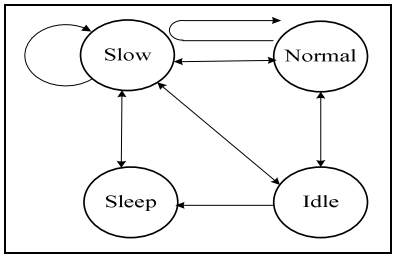


Figure 3 working mode conversion figure

a) SLOW mode

When the system is reset or after awakening from SLEEP mode, the system into SLOW mode. In SLOW mode, the processor clock is from 5MHz crystals source. Then operator can configure the PLL parameters, The configuration of PLL clock parameters in SLOW mode will automatically get application after into NORMAL mode.

b) NORMAL mode

In NORMAL mode, system all clock source are from PLL. In dynamic configuration PLL process, because there is a clock locked the time, namely within frequency divider working time, the clock waveforms output is irregular. In order to guarantee the normal operation of the system, the scheme is setting a relatively stable time 20 us. In the stability of the set time, system will automatically enabled SLOW mode，after the time automatically switch back NORMAL mode.

c) IDLE mode

Close the processor clock source, the other modules

keep the state ago. During IDLE mode, it nor can dynamically manage the modules clock source, because CPU has dormancy, unable to configure power management unit register inside.

d) SLEEP mode

In Sleep mode, closed CPU and all module clock sources. When it exited Sleep mode, , the system can only first back Slow mode, because in Sleep mode it also close PLL.

According to working mode conversion figure, define: after completion of initialized，Slow state as system starting point, then according to the needs of user, system switches to Normal state; When operating system detects idle tasks, if prediction time reaching set time, system will immediately be Idle state. If there are awakened events, system will return to a prior state; After entering Idle state in, if keeping time reaching setting or when the user manual hanging a system, the system will be into Sleep state. If there are awakened events, system will recover from sleep, after waking up, the system will firstly return to the Slow state, and then into the Normal state.

2) CPU work state transition management strategy

Chip's job is to keep the state in different between the conversions. As shown in figure 4, between normal condition and low power consumption state there is a switching delay, and in this period of time it will consume more energy. Therefore adopt what strategies to better reduce power consumption is the key point.

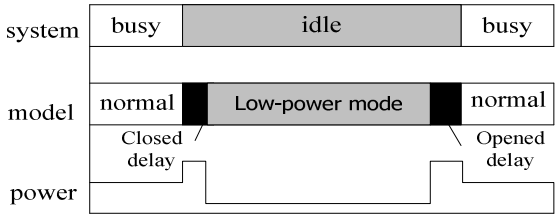


Figure 4 the influence of delay and energy when state switching

Set: T_I is idle time, T_O is sum of time entering low power consumption status (T_{in}) and waking up (T_{wakeup}), respectively, P_W and P_L are running state and low power consumption state power consumption, E_O is the extra energy entering low consumption and waking up. Then:

$$CASE\,1: E_1 = P_W \times T_I \qquad (2)$$

$$CASE\,2: E_2 = P_L \times (T_I - T_O) + E_O \qquad (3)$$

Because $P_W > P_L$, as long as T_I long enough, it is very easy to achieve the purpose of reducing power consumption. But Case2 increase delay. Due to the need for extra pay two aspects the price from running mode into low power mode: One is increased delay (T_O); another is the extra energy (E_O). So only in case of idle time long enough, choosing to enter low-power mode will save power consumption. The least idle time length calls critical closing time (T_{th}).So:

$$T_{th} \geq (E_O - P_L \times T_O)/(P_W - T_I) \qquad (4)$$

In this scheme design, the state transition management strategy adopts the method of static prediction [4, 5, 6]. It will be the past idle time as observation adjusting to predict the current total duration of idle time, The algorithm is as follows:

$$T_{pred}^{n} = \alpha T_{idle}^{n-1} + (1-\alpha) T_{pred}^{n-1} \qquad (5)$$

But when the system is awakened, because a idle time before is very long, next forecast time (T_{pred}) may also far outweigh the actual idle time, taking this factor into consideration, the algorithm set a constant c on the mend, so:

$$if\,(\alpha T_{idle}^{n-1} + (1-\alpha) T_{pred}^{n-1} > c T_{pred}^{n-1}) \quad T_{pred}^{n} = c T_{pred}^{n-1} \qquad (6)$$

Forecast time (T_{pred}) growth rate was limitation of the constant c automatically refreshed. And if $T_{pred} < T_{th}$, to keep the state system busy. At the same time open the watchdog timer tracking the actual idle time, if in after T_{th} still no request, the system start to predict. This algorithm has two advantages: ① it is easy to realize up; ② Set the constant c can reduce excessive predictive (namely predictive idle time than the actual time too long) possibility.

C. Device module management strategy

SoC chip integrates a lot of peripheral modules, some module related with system operation, as the clock, interrupt controller etc, these are not closed; Also some modules, related with application, may be closed according to the application. How to timely and effective management of these peripherals are very important to reduce power consumption. In the PMU, clock source control register

(PCSR) controls clock supply to the internal modules, every one of PCSR is for each module clock source control bit. In order to clear the use of device module information, this plan devised a global structure variable which contains the module using information, Structure is as follows:

```
Typedef struct
{
U32 dev_id;              // Device ID
ID owner_id;             // ID for task using the module
U32 called_count;        // cited frequency
U32 power_state;         // power consumption status
} dev_struct;
```

Module structure array is convenient to initialization, query and modification. There is corresponding relation between each application and device module, giving the corresponding relation set a mark (such as MP3_mod). This sign is defined as integral type macro. Each bit represents a module, and correspond with the clock control bit of PSCR register in the PMU. So using of PMU can effectively manage peripherals.

When an application program startup, operating system founds the name of application with ID task number, after finding the application of a matching task, let the macro (such as MP3_mod) logic AND corresponding clock bits of PSCR register, and determine whether open related modules, then modify their using of information structure variables. The operating system will go to read the module global variables structure every once in a while by timing query function, if found a device module not currently used, namely called_count = 0, power_state will be set to dev_off.

III. SIMULATION RESULT

The scheme has been applied in a special SOC chip, which was designed under the normal operating frequency, 100MHz. The system power consumption has been assessed in different working mode using Power Compiler, which is a power assessment software of Synopsys Corp. The assessment result is shown in table 1. In Normal mode, the clock frequency was set to 100MHz, and all the module clocks were open. When the system working mode went to Idle mode through Normal mode, CPU was closed and the clock frequency was 100MHz, and clocks in other modules were all open. In Slow mode, the system operation frequency was 5MHz, and all of the module clocks were open. When the system went to Idle2 mode through Slow mode, the system operation frequency was 5MHz, and clocks in other modules were all open. In Sleep mode, only RTC and PMC were open.

Table.1 Table of power comparisons in different modes.

module	Normal	Idle1	Slow	Idle2	Sleep
power （mW）	312.5	248.3	42.1	36.7	0.134

Through table 2, we can get the conclusion that cutting off the clock of UART and SPI, in Normal and Slow mode, can effectively reduce the system power consumption.

Table.2 Power comparisons table of the shutoff part in Normal and Slow mode.

module	Normal	Normal 1	Slow	Slow 2
power （mW）	312.5	282.4	42.1	26.7

Through the above simulation results, it is obviously proved that the system level design scheme has great effect to chip power management.

IV. CONCLUSION

This paper, starting from the clock frequency, has analyzed how to reduce the system dynamic power consumption from strategy view. This scheme combines with the PMU functions, from the two aspects of CPU and peripherals to analyze, and respectively formulates a different management strategy, realizes the CPU working state conversion and device module dynamic management by using the operating system. Through the simulation results can see, this scheme played the obvious effect to reduce the chip power consumption.

[1]. Zhong Tao, Wang Hao-cai. Power consumption optimizing and low power design technology of CMOS integrated circuit [J]. Microelectronics, 2003,30(2): 106-112.

[2]. Toshihiro Hattori. Challenges for Low-power Embedded SOC's [J]. 2007 International Symposium on VLSI Design, Automation and

Test (VLSI-DAT) 2007:1-4.

[3]. Yu Zhi-guo, Wei Jing-he. Low power design and implementation for a SoC. Solid-State and Integrated-Circuit Technology, 2008. ICSICT 2008. 9th International Conference on，2009,9(5): 20-23.

[4]. Benini L, Bogliolo A. A survey of design techniques for system-level dynamic power management [J].IEEE Transactions on VLSI Systems, 2000, 8(3):299-315.

[5]. Chi-Hong Hwang; Allen C.-H. Wu. A predictive system shutdown method for energy saving of event-driven computation [J]. Computer-Aided Design, 1997. Digest of Technical Papers, 1997 IEEE/ACM International Conference on, 1997:28-32.

[6]. L. Benini A. Bogliolo G. De Micheli. System-level Dynamic Power Management [J]. Low-Power Design, 1999. Proceedings. IEEE Alessandro Volta Memorial Workshop on, 2003:23-31.

Author Index

Author Index

Author Index

IEEE Computer Society Conference Publications Operations Committee

IEEE Computer Society Publications

The world-renowned IEEE Computer Society publishes, promotes, and distributes a wide variety of authoritative computer science and engineering texts. These books are available from most retail outlets. Visit the CS Store at *http://www.computer.org/portal/site/store/index.jsp* for a list of products.

IEEE Computer Society *Conference Publishing Services* (CPS)

The IEEE Computer Society produces conference publications for more than 250 acclaimed international conferences each year in a variety of formats, including books, CD-ROMs, USB Drives, and on-line publications. For information about the IEEE Computer Society's *Conference Publishing Services* (CPS), please e-mail: cps@computer.org or telephone +1-714-821-8380. Fax +1-714-761-1784. Additional information about *Conference Publishing Services* (CPS) can be accessed from our web site at: *http://www.computer.org/cps*

Revised: 1 March 2009

CPS Online is our innovative online collaborative conference publishing system designed to speed the delivery of price quotations and provide conferences with real-time access to all of a project's publication materials during production, including the final papers. The ***CPS Online*** workspace gives a conference the opportunity to upload files through any Web browser, check status and scheduling on their project, make changes to the Table of Contents and Front Matter, approve editorial changes and proofs, and communicate with their CPS editor through discussion forums, chat tools, commenting tools and e-mail.

The following is the URL link to the ***CPS Online*** Publishing Inquiry Form:
http://www.ieeeconfpublishing.org/cpir/inquiry/cps_inquiry.html