DCABES 2008 Proceedings

2008年国际电子商务、工程及科学领域的分布式计算和应用学术研讨会论文集

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

July 27~31, 2008, Dalian, China

Volume II

◎ 主　编：须文波
Editor in Chief: Wenbo Xu
◎ 副主编：刘　丹
Associate Editor: Dan Liu

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

DCABES 2008 Proceedings

2008年国际电子商务、工程及科学领域的分布式计算和应用学术研讨会论文集

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

Volume II

July 27~31, 2008, Dalian, China

主　编：须文波

Editor in Chief: Wenbo Xu

副主编：刘　丹

Associate Editor: Dan Liu

電子工業出版社

Publishing House of Electronics Industry

北京·BEIJING

内 容 简 要

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

图书在版编目（CIP）数据

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

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

中国版本图书馆 CIP 数据核字（2008）第 096214 号

责任编辑：秦绪军　　潘　娅
印　　刷：北京季蜂印刷有限公司
装　　订：北京季蜂印刷有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：880×1 230　1/16　印张：93.5　　字数：3000 千字
印　　次：2008 年 7 月第 1 次印刷
定　　价：398. 00 元（上、下册）

2008年国际电子商务、工程及科学领域的分布式计算和应用学术研讨会论文集

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

Volume II

责任编辑：秦绪军
潘　娅
封面设计：喻　晓

ISBN 978-7-121-07018-1

定价：398.00 元
（上、下册）

PREFACE

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

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

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

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

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

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

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

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

COMMITTEES

Steering Committee

Guo, Prof. Q. P. (Co-Chair), Wuhan University of Technology, Wuhan, China
Lai, Prof. C.-H. (Co-Chair), University of Greenwich, UK
Tsui, Dr. Thomas, Chinese University of Hong Kong, Hong Kong, China
Xu, Prof. W. B., Jiangnan University, Wuxi, China

Scientific Committee

Chair : Xu, Prof. W. B., Jiangnan University, China
Co-Chair : Lai, Prof. C.-H., University of Greenwich, U.K.
Guo, Prof. Q. P. , Wuhan University of Technology, China
Cai, Prof. X.C., University of Colorado, Boulder, U.S.A
Cai, Prof. Jiamei, Zhejiang Industry University, Hangzhou, China
Cao, Prof. J.W., Research and Development Centre for Parallel Algorithms and Software, Beijing, China
Chi, Prof. X.B., Academia Sinica, Beijing, China
Feng, Prof. Bin, Jiangnan University, Wuxi, China
Geiser, Dr. J.H. University at Berlin, Germany
He, Dr. H.W. Hohai University, China
Ho, Dr. P. T., University of Hong Kong, Hong Kong, China
Jesshope, Prof. C., University of Amsterdam, the Netherlands
Kang, Prof. L.S., Wuhan University, China
Keyes, Prof. D.E., Columbia University, USA
Khaddaj, Dr. S. Kingston University, UK.
Lee, Dr. John, Hong Kong Polytechnic, Hong Kong, China
Liddell, Prof. H. M., Queen Mary, University of London, UK
Lin, Dr. H.X., Delft University of Technology, Delft, the Netherlands
Lin, Dr. P., National University of Singapore, Singapore
Liu, Prof. Yuan, Jiangnan University, Wuxi, China
Loo, Dr. Alfred, Hong Kong Lingnan University, Hong Kong, China
Ng, Prof. Michael, Baptist?University, Hong Kong, China
Ng, Mr Frank C.k, Chinese University of Hong Kong, Hong Kong, China
Sloot, Prof. P.M.A., University of Amsterdam, Amsterdam the Netherlands
Sun, Prof. J., Academia Sinica, Beijing, China
Tsui, Mr. Thomas, Chinese University of Hong Kong, Hong Kong, China
Wang, Prof. Meiqing Fuzhou University, China

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

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

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

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

Zhuang Prof. Yueting, Zhejiang University,Hangzhou, China

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

Zhang, Prof. Wenyuan, Jiangnan Computer Research Institute

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

Local Organizer Committee

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

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

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

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

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

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

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

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

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

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

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

CONTENTS

Volume Ⅱ

Research of Multicast Method in FTTH *

Chuanqing Cheng[1] Cheng Chuanhui[2] Wang Li[3]

1 Department of Computer Science, Wuhan University of Science and Engineering , Wuhan, Hubei, 430074, China

Email：ccqcjl2005@126.com

2 Zhongnan University of Economics and Law, Wuhan, Hubei, 430074, China

Email: sammicch@hotmail.com

3 School of Electronic Information, Wuhan University, Wuhan, Hubei, 430074, China

Email：wl3833@126.com

Abstract

Ethernet Passive Optical Network (EPON) is a new technology which is considered one of the best solution of access network to support FTTH. In FTTH system， the different service is running on the same platform, data/voice/iptv must be supported in the system. For the iptv service, multicast is the important technology. This paper discussed multicast method in FTTH device, introduced different protocol, gave multicast filtering scheme .In the end, a multicast scheme base on LLID is introduced.

Keywords：Epon; llid; igmp ;multicast service

1 Introduction

EPON is the best solution of NGN(Next Generation Network) system with predominance of both Ethernet and PON. The outstanding advantage is great use scope, good extending performance and compatibility, lower cost. Since 2004.6 when the 802.3ah standard was brought out by IEEE802.3 EFM working group, the EPON technology has greatly developed. It is the both goals of network supplier and equipment supplier to supply multi-service access network. The pure data EPON equipment, which is according with IEEE 802.3ah standard, can not satisfied various services. EPON must have abundant operation function and maintenance function. So EPON must be a integrated broadband service access platform, can supply many kinds of services which will not interfere each other. EPON is up against some challenges: the technology which designed to Ethernet how to supply some guarantee mechanism to multi-services. EPON delivers data from a centralized Optical Line Terminal (OLT) to customer premises equipments, called Optical Network Units (ONUs).

EPON breaks the bandwidth of ordinary transmission line and can transmit multiple services, such as IPTV/DATA/VOICE. It is a novel optical access network technology, deploying point to multi-points structure, passive transmission on fiber, supplying service on Ethernet. It adopts PON technique on physical layer and Ethernet technique on data link layer and implements Ethernet access with PON structure. The services of multicast have the common character of that the single source information can be received by multi end station.

The section 2 introduces single LLID technology and multiple LLID technology. Section 3 introduces the filleting mechanism. a detailed controllable multicast scheme in EPON system . The multicast based on LLID mechanism is discussed in section 4. Section 5 is

* This work is supported by Hubei Prrovince Nature Science Foundation under Grant 2006ABA296

conclusion and future work.

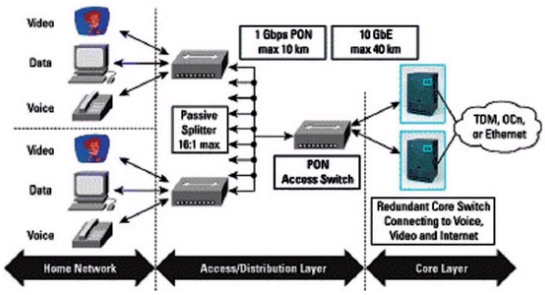


Figure 1 EPON supporting multi-sevice

2 Single Llid and multiple llid of epon

EPON network structure is point to multi-point(P2MP).It uses passive designated network(ODN) to carry ethernet data, connecting Optical Line Terminal（OLT）to Optical Network Unit（ONU）by fiber. Per OLT supports up to 64 ONUs. In the downstream direction, any signal from OLT can be transformed to all ONUs by broadcast communication. The maximum bandwidth can reach 1Gbits/s.In the upstream direction ,ONUs share the 1Gbits/s bandwidth by TDMA. Per ONU can passes at lease 15Mbits/s bandwidth.

LLID is a digital ID which is designated to the logic link built by P2PE Sublayer .by EPON systems.. Per logic link can get a different LLID. In EPON systems ,LLID is designated form OLT by network manager, then OLTknows clear which Onu the packet is from by identifying LLID. Moerover, packers can be forwarded to another ONU by modifying the LLID .Then we build a bi-directional path of OLT-ONU to communicate.

In a common EPON system, an ONU connects to OLT by only one LLID. Then the system only can supply basic media access control service(MAC).It means the system only supply a single service-data services. Moreover ,It have wake ability of QoS. Because the only way to realize Qos is 802.1p protocol. The 802.1p can improve the service quality a bit but not be adequate. The key problem is that the 802.1p is no t a link and have no ability of bandwidth control ability.

2.1 single LLID

EPON system which is according to 802.3ah standard can supply high speed of up to 1G bit. It is very hard for non-fiber system to do and is the physical fundament. The Multi-LLID technology can improve the multi-services supplying ability of EPON. EPON is a typical main-slave system. Ordinarily each ONU will be designated a LLID after it finish its registration, so the system can adjust some service performance agilitly, such as bandwidth or dithering.

If the ONU have single LLID of a EPON system, the system will be faced some problems like the following:

(1) The remote equipment(ONU)must supply more than one port to supply different service. Different service uses different port. Each port gets different services. On a common the ONU is installed at customer’s home. So the system must supply effective remote-management of ONU ,include queue distribution and port management.

(2) Traditional EPON only manage the ONU bandwidth and can not deal with QoS of different services. But if the network supplier run mutli-services, the SLA designated and bandwidth multiplex is necessary.

(3) When the service reach the focus equipment, the bandwidth constringency and services identification is pivotal functions of EPON to forwarding the stream. EPON is a L2 device, to supply multiple VLAN mode is a inevitable choice.

2.2 multiple LLID

According to what is introduced in 2.1,we know that in the single-LLID system, the system can’t confirm the bandwidth of a special priority and can’t bind service to priority well. Multiple LLD technology can give a well solution. Each LLID is corresponding to a different service. It means can control the QoS rigidly by LLID. Multiple LLID technology enhances the

EPON QoS. In fact, LLID is as a fix length mark, system to policy the packet according to the different mark. Moreover, Multi-LLID technology not only improve QoS but also be helpful of data safety and system management

The multiple LLID per ONU has many advantages:

Predigest designation of ONU

High employ of bandwidth

Supply delay-sensitive service support

TDM over IP service support

The Multi-LLID technology overcomes the Ethernet shortcoming, making the Ethernet be a integrated access platform. Multi-LLID can supply multi-services smoothly and is the trend of PON technology. The ONU port can be set one or more than one LLIDs, just like figure1.

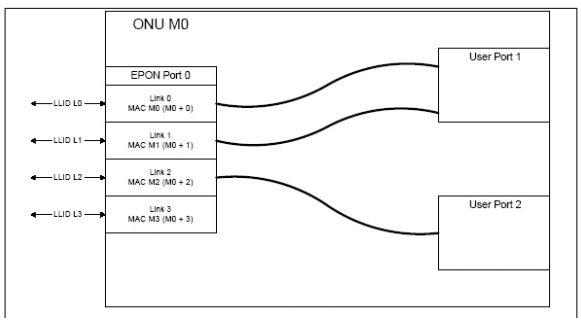


Figure 2 one port and more than one LLID

EPON system does filtering data mainly according to LLID. Each ONU have one or more LLID. Broadcast Frame brings a broadcast LLID. When the downstream data pass the RS sub layer, LLID will be added to the packet. Once ONU receives the packet, it will check the LLID of this packet, if the LLID is matched with itself , it will forward it to upper layer, otherwise it will discard the packet. If the LLID of the packet is broadcast LLID, the ONU will also forward it. So if an OLT will send data to multiple ONUs , only two ways : (1) send many copies of the data , in each copy data the different LLID is inserted.(2) send one copy of the data ,but broadcast LLID will be inserted to the packet. The downstream width will be wasted if the first way is used and ONU will be overloaded if the second way is used. So multicast LLID must be used in LLID multicast scheme.

IGMP(Internet Group Management Protocol） is the base to implement IP multicast. The Protocol runs between host and multicast router ,to support the both to forward multicast

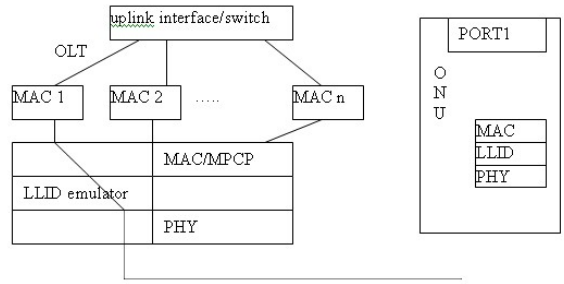


Figure 3 sub layer protocol reference model

3 Multicast filtering scheme

In EPON system, there are three downstream communication ways: unicast, multicast, broadcast. When the network running in multicast way, the packet from OLT can only be received by onus which is set in advance, other onu will not process the information. OLT only need to send a multicast stream, which can not only decrease the waste of bandwidth and improve the using rate of downstream bandwidth, but also distribute the pressure of EPON.

EPON deploys broadcast way in downstream direction. The downstream data broadcast to all of the onus. Each onu filter packets and receive own one. Both RS layer and MAC layaer can filter multicast packets. The difference is like following:

3.1 MAC Filtering

If multicast filter is on MAC layer, the broadcast LLID can be used. (1) Use Default LLID for all traffics with multicast MAC address(2) MAC Layer discard frames with unknown multicast MAC address. (3) RS Layer do nothing about all multicast traffics. (4)RS is responsible for filtering unicast packet. MAC is responsible for filtering multicast packet.

When the onu received the packet, all the multicast and broadcast LLID can pass the RS layer. Then the unknown MAC address which has passed the RS layer

will be dropped.

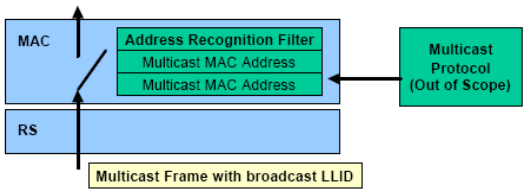


Figure 4 MAC filtering

3.2 RS filtering

Other method is to filter multicast packet on RS layer A multicast LLID must be defined with another mode bit.,RS Layer will discard frames which has unknown multicast LLID.MAC Layer may not require another filtering for the multicast. Mapping the multicast MAC address to the multicast LLID has to be defined. The map method can be Hash function or direct mapping

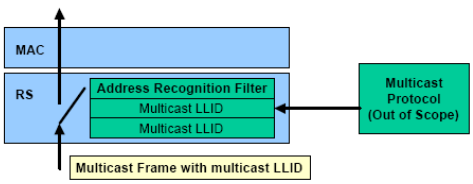


Figure 5 RS filtering

4 Multicast based on LLID

EPON system does filtering data mainly according to LLID. Each ONU have one or more LLID. Broadcast Frame brings a broadcast LLID. When the downstream data pass the RS sub layer, LLID will be added to the packet. Once ONU receives the packet, it will check the LLID of this packet, if the LLID is matched with itself , it will forward it to upper layer, otherwise it will discard the packet. If the LLID of the packet is broadcast LLID, the ONU will also forward it. So if an OLT will send data to multiple ONUs , only two ways :

(1) send many copies of the data , in each copy data the different LLID is inserted.(2) send one copy of the data ,but broadcast LLID will be inserted to the packet. The downstream width will be wasted if the first way is used and ONU will be overloaded if the second way is used. So multicast LLID must be used in LLID multicast scheme.

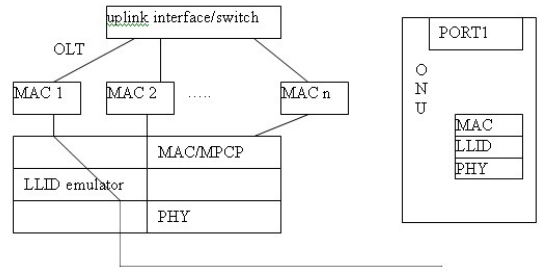


Figure 6 sub layer protocol reference model

4.1 IGMP snooping

IGMP Snooping is running in Layer-2. It snoops the IGMP packet between host and router., extract the L3 information, recur to the L2 group function to build and maintain the multicast table. When the device receives the multicast stream, only to copy stream to forward to the member in the group but not to the out.

In EPON, there are two pars to take part in the IGMP snooping procedure, OLT and ONU. The IGMP membership report and leave message from user is sent to ONU firstly and to OLT in succession. So the OLT and ONU also need to maintain multicast table .

If the IGMP snooping works in RS sub layer. The IGMP packet will be analysis in RS sub layer. OLT will decides add the LLID in the frame to the group or delete it from the group.

4.2 Form multicast table

After OLT analysis the IGMP protocol packet, the join and quit information of the multicast group can be get. Then the multicast table of OLT is formed. But the ONU also need LLID filter table. Only by the table ,the ONU can receive multicast data of it own. How to form and refresh the ONU multicast table are the key technology in implementing LLID multicast in EPON.

If there are multiple user has joined the multicast table, then one of the users quit the multicast table. OLT will delete the LLID from multicast table ,which will

affect other users of this ONU. A scheme can be implemented as following:

1) When OLT get the IGMP join packet, it create the multicast group and send a OAM frame to the ONU.(multicast JOIN OAM).ONU will add a entry to its multicast table.

2) When the OLT get the IGMP leave packet, it will not delete the LLID directly, but send a OAM frame to the ONU to inform the leave informaiton ,the ONU will delete the entry according to the OAM.

Aging mechanism is used on OLT. If the new multicast group is created, a timer will be created at the same time. The period of the timer is set to t. If in t , there is no membership report , the OLT can think that all user of the ONU don't need multicast service. It will delete the multicast group and the LLID of all member of the group. Then a OAM (multicast LEAVE) is sent to all the ONU to inform the ONU deleting the group.

3) When the ONU receives the multicast JOIN , it check the multicast table. If there is no the group LLID in the table ,a new entry is added, and the user MAC address is added too. If there is the group LLID already, check if there is user MAC address, if no, add it.

When the ONU receives the multicast LEAVE, it checks the multicast table and delete the user MAC address from the group. If there is no entry in the group, ONU must send a multicast LEAVE OAM to OLT, to inform the OLT to delete the LLID from the group.

The multicast-inform-OAM can refresh the multicast table between ONU and OLT. The user MAC address is not the integrant part of the multicast table. It is only to mark the users of the same ONU to avoid the circs that if a member of a ONU leaves, all the users can not receive the multicast frame.

5 Conclusion

IEEE 802.3 ah has not defined how to implement multicast technology. This paper discusses the multicast mechanism Using MAC filtering or RS filtering ,how to use RS filtering can be different in different EPON device. The multicast technology can improve multiple service supporting ability

References

[1] G Kramer , B Mukherjee , G Pesavento. IPACT :a dynamic protocol for an Ethernet PON (EPON)[J] . IEEE Commun ,2002 ,40 (2) :74 - 80

[2] Deering S. Host extensions for IP multicasting. RFC 1112, Stanford University, 1989

[3] Fenner W. Internet group management protocol. Version 2, RFC 2236, Xerox PARC, 1997

[4] Biswas S, Haberman B, Cain B. IGMP multicast router discovery. Nortel Networks and Cereva Networks. Internet-Draft, 2001

[5] B.Cainetal，Internet，Group Managem ent Protocol，Version 3, RFC3376,October 2002

[6] Cheng Chuanqing, Wang Li IP multicast group management protocol and implementation in L2 《 information technology》， 2003，9

[7] Fenner B, He HX, Haberman B, Sandick H. IGMP-Based multicast forwarding (IGMP proxying). AT&T-Research, Nortel

[8] [Parkhurst WR. Cisco Multicast Routing And Switching. McGraw Hill, 1999. 25~42, 43~53

[9] IEEE P802.3ah task force.IEEE Draft P802.3ahTM/ D1.9,Operations,Administration and Maintenance (OA M).http://www.ieee802. org/3/efm,2003-02-31

[10] X. Ma and L. Ping, "Coded modulation using superim posed binary codes," IEEE Trans. Inform. Theory, vol. 50, no. 12, pp. 3331–3343, Dec. 2004

A Fair Electronic Cash Scheme Based on Elliptic Curve Cryptography*

Shouzhi Xu Qiaoli Liu Hen Yu Huan Zhou

College of Electrical Engineering & Information Technology, China Three Gorges University
Yichang, Hubei, 430074, P.R. China
Email: xsz@ctgu.edu.cn

Abstract

Fair tracing is an important mechanism to prevent or detect crimes in anonymous electronic cash schemes. Current fair electronic cash schemes are mostly based on discrete logarithm problem on finite, which are not efficient for new applications on mobile electronic payment owing to time cost and storage requirement of exponentiations. In this paper, a fair electronic cash system based on Elliptic Curve Cryptography is proposed. An opening protocol, a withdrawal protocol and a payment protocol of the electronic cash scheme are illustrated. With a blind signature technique based on Discrete Logarithm Problem (ECDLP), electronic coins can't be forged without secret key of bank, and customers' privacy is protected well. But the anonymity of participants can be revoked if the e-coin is spent more than one times. In comparison with recent similar research schemes based on RSA, our scheme has 30% reduction in the storage requirement of e-cash.

Keywords: Electronic cash system; elliptic curve cryptography; restrictive blind signature; anonymity; fair tracing

1 Introduction

Electronic commerce is one of the most important applications for the internet. The prerequisite for establishing an electronic marketplace is secure and fair payment [1]. Customers' privacy should be well protected if they are embedded in legal commercial transactions or payments. Owing to the unlinkability properties, blind signature techniques [2] have been widely used to protect the right of an individual's privacy in the untraceable electronic cash (e-cash) systems [3]. However, it is easy to make multiple copies of the electronic coins, which is in the form of number strings. Therefore, fair blind signature techniques [4] [5] are developed to withstand the possible abuse of unlinkability. If users misuse the unlinkability property, such as money laundering and blackmailing, then a trusted third party can get enough information to reveal their identity with specific protocols.

Recently, much work has been performed in the area of such fairness. There are three main methods to detect double-spending: 1) Checking each electronic coin whether double spent on-line with an central bank database; 2) Using tamper-resistant hardware; 3) digital signature with restrictive blindly signature techniques. In practice, we need some anonymity revocation mechanisms to prevent criminal activities or to trace the criminals. On the other hand, we need another trusted third party (TTP, in brief) to perform this mechanism to protect the honest participants of the system. So, it makes electronic cash system very complicate. Those above electronic cash schemes based on RSA cryptosystem can ensure strong secrecy for e-commerce, but they are not efficient enough in modern applications such as mobile payment owning to a lot of exponential computation needed. Terminals in mobile payment scenario are portable devices such as PDAs, cell phones, smartcards, and many others, which possess less power capability, lower computing ability and smaller storage. So, it is very urgent to find a new efficient way for such

applications.

Elliptic curve cryptography (ECC), introduced by Miller[6] and Koblitz[7] in the 80's, has attracted increasing attention in recent years due to its shorter key length requirement in comparison with other public-key cryptosystems such as RSA[8]. For example, 160-bit Elliptic-curve Digital Signature Algorithm has a security level equivalent to 1024-bit DSA. Such advantages make ECC a better choice for public-key cryptography in resource constrained systems [9]. In this paper, we propose a new fair blind signature scheme based on elliptic curve discrete logarithm problem and apply it to electronic payment. The anonymity of participants can be revoked in our double-spending-resistant system, and our system has the ability to trace both the electronic coin and the owner of the electronic coin.

The rest of context is organized as follows: The next section, Section 2, presents elliptic curve discrete logarithm problem over finite fields. In Section 3, we introduce a model of e-cash system and a blind signature scheme based on ECDLP, then we propose a withdrawal protocol and a payment protocol of the e-cash scheme. Subsequently, in section 4, the security of which is proved and the efficiency of which is analyzed in contrast to recent references. Finally, we conclude our work for this paper in the last section.

2 Preliminaries

We define a Weierstrass equation over a prime field $GF(p)$ (where p is a large prime):

$$E: y^2 = x^3 + ax + b \qquad (1)$$

Where $p \in (2^k, 2^{k+1})$, $a, b \in GF(p)$, $\Delta = 4a^3 + 27b^2 \neq 0$.

The set of pairs (x, y) that solves function (1) and the point at infinity O form an abelian group, which is used to implement the Elliptic Curve Cryptosystem (ECC).

For efficient operation methods of point addition and point multiplication of $E(\mathrm{F}_q)$, refer to [7] for more details. Given an ECC system, it is necessary to find a right generator point G, which makes its order n big enough (n is the minimum number which makes nG=O)[6]. As $E(\mathrm{F}_q)$ is an Abel group, it is closed for addition operation of any element in group $E(\mathrm{F}_q)$. That means there is a point $R = kG \in E(\mathrm{F}_q)$ for an integer $k \in_{\mathrm{R}} \mathbf{Z}_n$ (Select an integer for field $\mathbf{Z}_n$ at random), so $E(\mathrm{F}_q)$ is a finite cyclic Abel group, and n is its order $\#E(\mathrm{F}_q)$.

Definition 1: Elliptic Curve Discrete Logarithm Problem (ECDLP), defines as: Let $E(\mathrm{F}_q)$ be an Abel group of

elliptic curve over a finite field F_q, and a generator point $G \in E(\mathrm{F}_q)$, whose order is n. Given any point $R \in E(\mathrm{F}_q)$, the elliptic curve discrete logarithm problem is to find an integer k, which satisfies $R = kG$.

Given the point G and an integer k, it is easy to figure out $R = kG$, whereas it is very difficult to find the integer k when given the point G and R only.

This paper use following notations:

Table 1

Notation	Description
B	Bank
M	Merchant
C	Customer
T	Trustee (TTP, third trusted party)
P	Public key(e.g. P_B: public key of bank B)
S	Secret key (e.g. S_B: secret key of bank B
$E(\mathrm{F}_q)$	Abel group of elliptic curve
$k \cdot G$ (or kG)	Point multiplication operation
P+Q	Point addition operation
$H(\square)$	An one-way hash function(SHA-1 or MD5)
Encry(m, *key*)	Encryption function
Decry(m, key)	Decryption function

3 The Proposed Electronic Cash Scheme

3.1 System model

This paper applies an e-cash model as Figure 1. In this model, an electronic wallet consists of an observer

and a computer of user. The bank **B** releases its system parameters (p, a, b, n, G, P_{1B}, P_{2B}) for users.

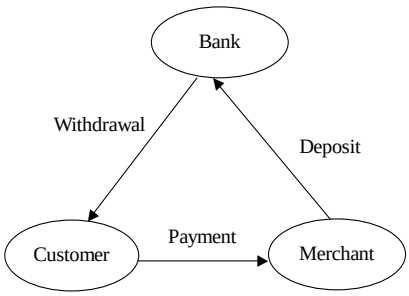


Figure 1 A model of e-cash scheme

3.2 Blind signature technique based on ECDLP

Let $\phi \in E(\mathrm{F}_q)$ be a message. The main steps of blind signature perform are as follows:

(1) The signer chooses $S_w \in_R \mathbf{Z}_q$ (in this paper, $\mathbf{Z}_q$ denotes $\{1,2,...,q-1\}$), computes $a = H(S_S \| \phi)$ (where a is blind version of message ϕ), $z = H(S_w G \| a)$ and $b = S_w + zS_S \pmod n$, and encrypts the signature $\{a,b,z\}$ by $s = Encry(\{a,b,z\}, S_S P_R)$ and then sends it to the recipient.

(2) The recipient decrypts the encrypted message by $Decry(s, S_R P_S)$ to obtain $\{a,b,z\}$, then checks the signature of the signer by verifying quality $z = H(bG - zP_s \| a)$.

(3) The recipient chooses $S_{w'} \in \mathbf{Z}_q$ at random, computes $a' = H(P_R \| a)$, and encrypts the signature by $s' = Encry(\{a', S_{w'}\}, S_R P_S)$ and sends it to the original signer.

(4) The original signer decrypts the encrypted message by $Decry(s', S_R P_S)$ to obtain $\{a', S_{w'}\}$, then verifies the signature of the signer by checking equality $a' = H(S_R P_S \| a)$ .If acceptant, he will sends back $Encry(S_{w'}, S_R P_S)$.

With the proposed signature technique, a receiver can get a correct signature on a original message which the receiver can't read out (it is blinded owning to hash function), so the receiver can testify the identity of the signer, but anyone can't forge the signature or the original message if the hash function is strong enough for anti-collision.

3.3 Opening an account

We denote the bank by B, a customer by U, and a merchant by M. Bank B sets up two account databases for customer U and merchant M. At the setup stage, bank B selects $S_1, S_2, S_3 \in_R \mathbf{Z}_q$ as the secret key of a signature, and computes a corresponding public key of the signature $P_1 = S_1 G, P_2 = S_2 G, P_3 = S_3 G$. Then B proclaims system parameters $\{q, a, b, G, n, H(\square), Encry, Decry, P_1, P_2, P_3\}$ in bulletin.

Both customers and merchants need open their accounts in a bank. When a user U request to open an account in a bank B,

B generates a unique integer $ID_U \in_R \mathbf{Z}_q$ as the secret identity of the customer. Then B computes and verifies equation $pk' = ID_U P_1 + P_2 \neq O$, if it is accepted, computes $\theta_U \equiv ID_U S_1 + S_2 \pmod q$ sequentially, then records $\{name, ID_U, \theta_U, pk', balance\}$ and informs account $Encry(\{name, ID_U, pk', balance\}, S_B P_U)$ to U. The bank, the customer and the merchant can get their pairs of secret and public keys $\{S_B, P_B\}, \{S_U, P_U\}$, $\{S_M, P_M\}$ respectively.

3.4 The withdrawal protocol

When customer U wants to draw electronic cash form bank B, the main steps of the withdrawal protocol follows as:

1) The customer sends his request to the bank, which includes his ID and the amount of withdrawal. The request message is signed with his secret key;

2) B verifies the request whether the ID of U is legal and the requested amount of cash *d* is under the rule of bank, for example, the needed cash *d* must not be more than his total saving in the bank.

3) If the request is verified, the bank draws the e-cash for the customer, which is signed by the bank. Before the customer U computes the e-cash, U must verify the signature of Bank firstly.

The details of the withdrawal protocol are indicated in Figure 2.

The e-cash is represented by $(d, pk_1, pk_2, (c, r))$, which has been signed by customer U and B. Parameters $(\alpha, \zeta_1, \zeta_2)$ are signature keys of the e-cash. The e-cash is saved in a client software, such as electronic wallet.

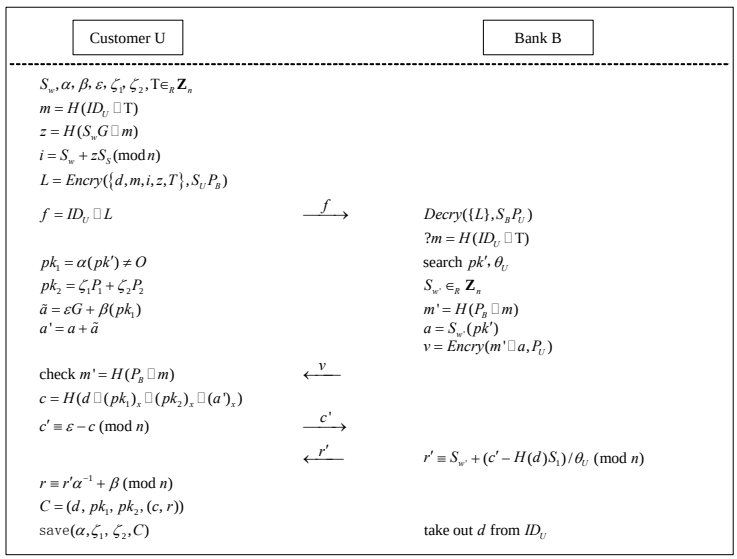


Figure 2 Withdraw protocol based on ECC

(Note: $(P)_x$ means the x-coordinate of point P on an elliptic curve)

The payment protocol

When customer U wants to buy some goods from merchant M with his e-cash, he should send his buying request (list of goods and their amount) and his e-cash to the merchant M. the main steps of the payment protocol follows as:

1) The customer U sends request of buyment includes the list of goods and their quantity, and pays it with an e-cash *C*.

2) The merchant M verifies the request on business rules, for example, checking if there is anything wrong with the ID of goods, price, and total cost. If acceptable, M must validates the e-cash *C*.

3) M sends a claim of payment to U, and U sends M the signature of the e-cash. If it is accepted, the transaction is done.

The details of the payment protocol are indicated in figure 3.

In this protocol, signatures ρ_1 and ρ_2 can help to reveal the identity of user if the e-coin is spent twice.

4 Security and efficiency analysis

In this section, we will analyze the security and efficiency of the electronic payment scheme based on the proposed blind signature scheme.

4.1 Security analysis

The security of the scheme can be analyzed by the following propositions:

Proposition 1(**Correctness**): if the recipient follows the blind signing protocol and accepts it, then $\{a, b, z\}$ is a correct signature on blinded message of ϕ.

Proof: The signature $\{a, b, z\}$ is a correct signature on ϕ if the equality ($z = H(bG + zP_s \| a)$) is verified. If it can be assumed that $H(\square)$ is collision-resistant, then this is equivalent to proving that $bG - zP_s = S_w G$. Obviously, the relation follows from: $bG - zP_s = (S_w + zS_S)G - zP_S = S_w G$ because $b = S_w + zS_S$ if the recipient accepts.

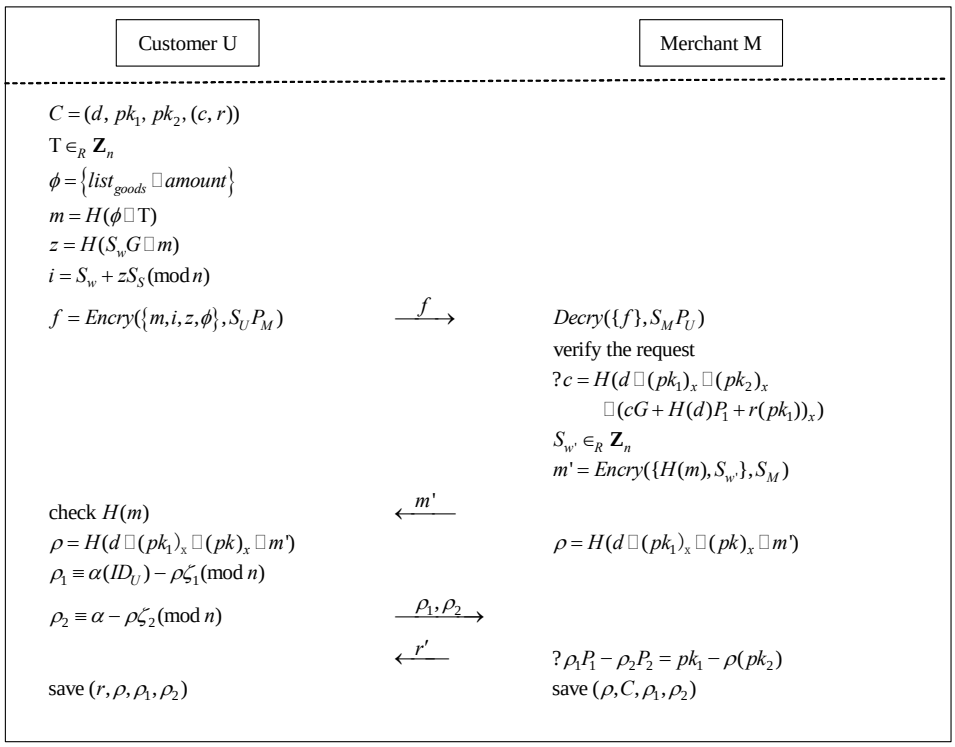


Figure 3 Payment protocol based on ECC

Proposition 2(**Non-forgeability**): the recipient and other third parties can't form the same signature of the signer if they follow the blind signing protocol.

Proof: Two evidences here. Firstly, other third parties can't form the signature message without the secret key of the signer, because the recipient decrypts the encrypted message with the public key bearing an asymmetric encryption function. Secondly, the recipient can't forge a correct signature with the information of $\{a, b, z\}$, because the secret key of the signer must be embed into signature element *b*. Anyone who wants to get the secret key of the signer has to solve $P_S = S_S G$, which belongs to ECDLP.

Proposition 3(**Blindness**): the recipient can't access the original message when he gets the signature of the signer.

Proof: the recipient accepts the signature $\{a, b, z\}$ if he. Since $H(\cdot)$ is collision-resistant, the recipient can't recover the original message ϕ from the blind version $a = H(P_S \| \phi)$.

Proposition 4: No one can forge a correct e-cash without secret keys of the bank.

Proof: If an amount of e-cash $C = (d, pk_1, pk_2, (c, r))$ can be used to buy goods following the proposed payment protocol, it must satisfy equality:

$$cG + H(d)P_1 + r(pk_1) = S_{w'}(pk') + \varepsilon G + \beta(pk_1) \quad (2)$$

Since $pk_1 = \alpha(pk')$ and $pk' = ID_U P_1 + P_2$, we have:

$$\begin{aligned} cG - \varepsilon G &= S_{w'}(pk') + \beta\alpha(pk') - r\alpha(pk') - H(d)P_1 \\ &= ((S_{w'} + \beta\alpha - r\alpha)(ID_U S_1 + S_2) - H(d)S_1)G \end{aligned} \quad (3)$$

So $c - \varepsilon = (S_{w'} + \beta\alpha - r\alpha)(ID_U S_1 + S_2) - H(d)S_1$.

But any attacker who wants to get secret keys S_1, S_2 of the bank must solve the ECDLP problem described in section 2, so he can't forge such parameters in function (4) to make a correct e-cash *C*.

Proposition 5: If an amount of correct e-cash is spent more than once, its owner can be discovered by the bank.

Proof: Suppose $C = (d, pk_1, pk_2, (c, r))$. If a customer U spend it to buy some goods from a merchant M, M will send (m', C, ρ_1, ρ_2) to bank B to deposit it, where *m'* is the buying request of U, (ρ_1, ρ_2) is the signature of the cash. If U spends C more than once, B will discover it when later merchants deposit such e-cash. Since $S_{w'} \in_R \mathbf{Z}_n$, the values of this signature parameter

are different each time when M validates the e-cash. In such a case, even if U has the same buying request *m*, $m' = Encry(\{H(m), S_{w'}\}, S_M)$ and $\rho = H(d \cdot (pk_1)_x \cdot (pk)_x \cdot m')$ will be different.

Here let's discuss a case when U spends double times.

Suppose B has (ρ, ρ_1, ρ_2) and $(\rho', \rho_1', \rho_2')$ from different deposition operation of M. Since $\rho_1 \equiv \alpha(ID_U) - \rho\zeta_1$, $\rho_2 \equiv \alpha - \rho\zeta_2 (\mathrm{mod}\, n)$, $\rho_1' \equiv \alpha(ID_U) - \rho'\zeta_1 (\mathrm{mod}\, n)$, and $\rho_2' \equiv \alpha - \rho'\zeta_2 (\mathrm{mod}\, n)$, so B has:

$$\begin{cases} \rho'\rho_1 - \rho\rho_1' \equiv (\rho'\alpha(ID_U) - \rho\alpha(ID_U)) (\mathrm{mod}\, n) \\ \rho'\rho_2 - \rho\rho_2' \equiv (\rho'\alpha - \rho\alpha) (\mathrm{mod}\, n) \end{cases} \quad (4)$$

Since $pk_1 = \alpha(pk') \neq O$, $\alpha \neq 0$, so

$$ID_U \equiv (\rho'\rho_1 - \rho\rho_1')/(\rho'\rho_2 - \rho\rho_2') \ (\mathrm{mod}\, n) \quad (5)$$

From the above proposition 4 and 5, we can see that no one can forge a correct e-cash, for the merchant can discover the identifier of a customer in the bank, and the customer can't spend the same e-cash a second time. So it is a fair protocol in sense of security (the foundation of our system is that the bank is reliable enough.). This assumption corresponds to financial rules in real society (The bank must keep privacy of customers by law). Threatens from the bank is limited, for the anonymity of a customer can't dispose if he executes protocols normally.

4.2 Efficiency analysis

The time cost and storage are considered critical factors of efficiency. In our scheme, both signature and encryption operations have no exponentiations, which makes it more efficient than schemes based on RSA or DSA, since 160 bits key in ECC is as strong as 1024 bit key in RSA. For example, the message of e-cash is 5144 bit in [5], 1464 bit in [9] and 960 bit in our system. Therefore, the proposed electronic payment scheme of ours has 30% reduction in the storage requirement. In addition, compared to scheme in [9], we have no trusted third party, so our scheme is simpler and more efficient.

5 Conclusions and Future Work

The main contribution of this paper is the electronic cash scheme based on elliptic curve discrete logarithm problem. The main security features of the proposed electronic cash scheme are proofed as well. We have proposed a withdrawal protocol and a payment protocol, and proved that e-cash is strong enough to resist attacking and that double-spent e-cash can be traced. Compared with recent research work, our electronic payment schemes simplify the system structure and act more efficiently.

As mobile payment is a fashion application nowadays, more work needs to be done on how to reduce the computation cost.

References

[1] J. Camenisch, J. Piveteau, M. Stadler, An efficient fair payment system, Proceedings of ACM Conference on Computer and Communications Security, pp. 88-94, 1996

[2] D. Chaum, Blind signatures for untraceable payments, Advanced in Cryptology, Proc Crypto'82, Springer-Verlag, 1983: 199-203

[3] S. Brands, Untraceable off-line cash in wallet with observers, Advances in Cryptology CRYPTO'93, Lecture Notes in Computer Science, Springer-Verlag, 1993, vol. 773: 302-318

[4] C. Wang, H. Xuan, A Fair Off-line Electronic Cash Scheme Based on RSA Partially Blind Signature, 1st International Symposium on Pervasive Computing and Applications, 2006:508-512

[5] Y. Tan, Z. Yang, S. Wu, An efficient restrictive blind signature scheme with applications to electronic cash, High Technology Letters, 2002, vol.8(4):60-63

[6] V. Miller, Uses of elliptic curves in cryptography, Advances in Cryptology, Proceedings of Crypto'85, Lecture Notes in Computer Sciences, Springer-Verlag, 1986, pp. 417-426

[7] MENEZES, Elliptic Curve Public Key Cryptosystems, Kluwer Academic Publishers, 1993

[8] L. Patrick, M. Ali, Fast and Flexible Elliptic Curve Point Arithmetic over Prime Fields, Transactions on Computers ,2007, (99):1-13

[9] C. Popweau, A fair off-line electronic cash system based on elliptic curve discrete logarithm problem, Studies in Informatics and Control, 2005, vol.14(4):291-297

A New Invasion Detection System Based on Mobile Agent and Fuzzy Synthetical Assessment Technology

Jianjun Liu Changyou Guo

Department of Computer, Dezhou University, Dezhou, Shandong, China

Email: ljj@dzu.edu.cn

Abstract

In the information era, the network has become not only a primary means of date exchanges and business transactions but also a major target of attackers. Therefore, an intrusion detection technology which can discover and report non-authorized access or anomalies of system comes into being. Traditional intrusion detection systems are inefficient and unsatisfactory in detecting ability, so in this paper, a new invasion detection system based on mobile agent and fuzzy synthetical assessment technology is presented. This system can detect not only known attacks accurately but also unknown attacks. Detection accuracy rate is improved. At the same time, it can also decrease systematic communication load and improve work efficiency tremendously.

Keywords: Invasion Detection System; Mobile Agent; Fuzzy Synthetical Assessment Technology; Information-collecting agent; track agent

1 Introduction

The invasion detection can collect and analyze information from key points in the computer network or system, discover behaviors whose accesses are non-authorized or behaviors threatening systemic security, and at the same time respond to these behaviors to ensure system security. The invasion detection is considered the second security strobe after the firewall. It can monitor the network without affecting network performance and can provide real-time protection against interior and exterior attacks and misoperations.

Invasion detection system (IDS) is a series of hardware and software that can realize automatically invasion detection. It can draw character values of flux statistics from the collected texts,, to be dealt with by intelligent analysis and to be compared and matched with the built-in invasion knowledge library. According to the preinstalled threshold value, the text flux whose match coupling degree is high will be considered as an attack, the invasion detection system will carry on warning or have limit counterattack according to the corresponding disposition.

The invasion detection system can be divided into two kinds based on the adopted analytical technique: singularity detection and Misuse detection [2].

2 Traditional invasion detection system and existed flaw

Dorothy E. Denning put forward a general Invasion detection expert system frame for the first time in 1987. It was called IDES. Figure 1 shows an IDES invasion detection system model[1].

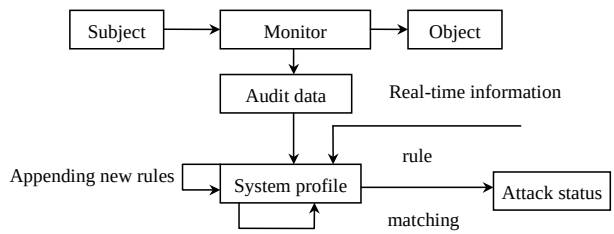


Figure 1 invasion detection system model

This model can detect hacker's invasions, unauthorized operations and abnormal uses of the

computer system. This model is based on such a supposition: it can recognize invasion behaviors that use abnormal system through inspecting system's audit records. The IDES model is made up of subject, object, audit record , profile, anomaly records and activity rules.

1) Subject is the active initiator in the system operation, for example program in computer operating system, network service connection and so on.

2) Object refers to the resources that the system can manage including file, command, equipment and so on.

3) Audit record refers to the data that system produces when subject operates object like user registration, order execution and file access and so on.

4) Profile can score behavior character that subject operates object using random variable and statistical model

5) IDES renews dynamic the profile and examines the unusual behavior. If an unusual behavior is discovered, an anomaly record is produced.

6) Activity rules indicate movements that the system should adopt when an audit record or anomaly record is produced. The rules have two parts: first is activity; second is movement.

The IDES model that Denning proposed is in fact a system based on the regular pattern matching. This conventional model and tool played the major role in the early time, but with the lapse of time, the flaws of invasion detection system based on the conventional model are discovered by more and more people. Firstly, IDES is unable to describe accurately attack behaviors in the network for example TearDrop attack. Secondly, attack behavior in the network is complex. Some aggressors dig systemic weaknesses by cooperated way, and IDES model is unable to describe aggressor's operating process. For example the sensitive password document is gained through WWW server's disposition weakness. Thirdly, IDES model can only protect the single host computer, but the goals of network invasion are diverse, The mail servers, the route servers and the domain name servers, the network communication link and so on are all the objects that intruders choose. Attacks are conjoint, so IDES is unable to judge hidden attack activities. Fourthly, IDES model limits the sources of invasion indication information to audit record, but it is insufficient. In fact, any auditing system has its own limit, and can not supply the information which the invasion detection needs.

In order to overcome the above flaws, this paper improves on the traditional invasion detection system, and proposes an invasion detection system based on mobile agent and fuzzy synthetical assessment technology.

3 An invasion detection system based on mobile agent and fuzzy synthetical assessment technology

3.1 Mobile agent technology[8]

COAST group in Purdue university proposed Autonomous Agent for Intrusion Detection, and was the first to apply agent to invasion detection system. In fact the agent is a series of software units that can complete certain detecting functions. This system has used hiberarchy, each agent running on host computer gathering the information of operating system as well as network activity. Single agent cannot make any resolution and can only transmit the information to the transmitter. The transmitter combines information which are collected from each agent to constitute the condition of main computer at that time, and makes the warning judgment or transmits results to high-level monitor. The monitor constructs the condition of whole network and makes judgment.

The mobile agent is a kind of special agent. With the development of computer network, agent can move and execute in the network to finish certain work It is concept of mobile agent. In mobile agent technology, the service request the agent to execute by moving to server. This agent can face directly the resources of serve which must be visited, without depending too much on network transmission. So it is reduced the system’s dependence on network bandwidth. Mobile agent does not need to be attempered uniformly The agent established by the user can run asynchronously on the different node, and after finishing mission, mobile agent transmits results to user. In order to complete some tasks, the user may establish

many agents that run simultaneously on one or many nodes to form the ability of parallel solution. Mobile agent also has characteristics of autonomy and intelligent route and so on.

These obvious characters of mobile agent qualify for constructing distributed invasion detection model especially.

3.2 Fuzzy synthetical assessment technology

There are many network connection attributes that can by used in detecting invasions. It is very difficult to give accurate expression to detect characters which different types of attacks aim at. Therefore the fuzzy synthetical assessment method may be used.

The singularity degree of network connection is regarded as language variable, and the singularity degree of network connection is assessed with linguistic value "normal" or "abnormal". An evaluated set is defined as $E=\{e_1,e_2,e_3,e_4\}$, thereinto e_1 = "normal", e_2 = "a little abnormal", e_3 = "abnormal", e_4 = "very abnormal". Membership function may be indicated using continual or discrete form. The membership degree belonging to the ith grade of evaluation may be calculated according to the membership function. The plan x includes m characteristic goal and 4 levels of evaluation set. Its fuzzy assessment matrix is as Eq. (1)

$$R=(\gamma_{ij})(i=1,2,...,m;\ j=1,2,...,4) \tag{1}$$

And $\gamma_{ij}=\mu_{ij}(x)$ expresses that membership degree of the i*th* goal relative to the j*th* grade in the plan x. When we carry on fuzzy synthetical assessment to many objects, we must give respectively each goal to weight. Suppose the weight value of the i*th* goal is W_i, then

$$\sum_{i=1}^{m} W_i=1, W_i\geq 0 \tag{2}$$

The weight vector is $A=(W_1,W_2,\ldots W_m)$.

Finally the fuzzy assessment matrix is as Eq. (3):

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{14} \\ r_{21} & r_{22} & \cdots & r_{24} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{n4} \end{bmatrix} \tag{3}$$

The formula is as Eq. (4)

$$B_i=A_i\cdot R=(W_1,W_2,\ldots W_m)\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{14} \\ r_{21} & r_{22} & \cdots & r_{24} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & r_{m4} \end{bmatrix} \tag{4}$$
$$=(b_1,b_2,\ldots b_4)$$

3.3 A new Invasion detection system based on mobile agent and fuzzy synthetical assessment technology

A material invasion detection system based on mobile agent and fuzzy synthetical assessment technology is illustrated in Figure 2. This kind of new invasion detection model is composed of intermediate agent, track agent, information-collecting agent, many monitors, system database and database and so on.

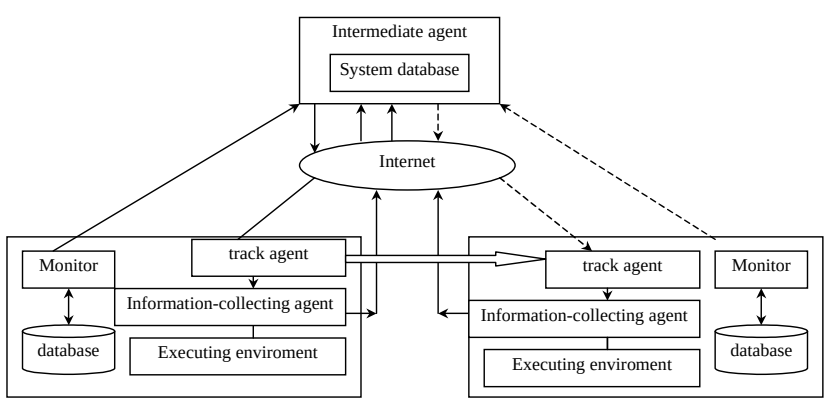


Figure 2 Invasion detection system based on mobile agent and fuzzy synthetical assessment technology

(1) Intermediate agent

Intermediate agent accumulates and assesses all kinds of information collected by information-collecting agent using fuzzy synthetical assessment technology. If these information has surpassed certain threshold value, intermediate agent thinks that the invasion has happened. Intermediate agent manages track agent and system database, and provides the surface for communication between manager and system. Intermediate agent resides in each network section.

(2) Monitor

Monitor appears in every target system It seeks for MLSI using system log in the monitor database. If the monitor has discovered MLSI, it would report to intermediate agent. Monitor might also report the type of MLSI.

(3) track agent

Track agent tracks the path of invasion and locates its initial station. Intermediate agent, monitor and track agent cooperate in the following way: Monitor detects a MLSI and reports to intermediate agent; Intermediate agent allocates a track agent to the target system using Agent Transfer Protocol; Track agent transfers automatically among computers, and does not require that the intermediate agent track invasion independently. When many MLSIs are discovered in some target system in a short time, intermediate agent will distribute many track agents to target system to collect all information of MLSI. Single track agent can not make the judgment of invasion, neither can it determine whether the invasion has happened. Only the combination of several track agents can enable intermediate agent to make fuzzy systhetical assessment. Track agent can move to any system installing the executing environment of agent, so it can carry on the invasion route track intelligently.

(4) Information-collecting agent

Information-collecting agent is mobile, and it collects information relative to MLSI in the target system. When track agent investigating intruder is allocated to the target system, it will activate information-collecting agent residing in this system; The information-collecting agent collects information according to the type of MLSI, and returns results to intermediate agent. If track agent shifts to another target system, it will activate information in the present target system. Many different track agents in the same target system can activate many information-collecting agents.

(5) System database and database

System database is located at intermediate agent computer, and is used to record information collected from the target system as well as integrate information from each track route. Database appears in each target system, and is used to deposit system log.

The work process of invasion detection system based on mobile agent and fuzzy synthetical assessment technology is as follows: When the monitor detects the MLSI in target system, it will report it to the intermediate agent, and the intermediate agent will allocate a track agent to the target system. The track agent arrives at the target system and activates an information-collecting agent residing in this system. The information-collecting agent collects information relative to the MLSI, and the track agent will investigate the initial station of MLSI to distinguish where invaders are making their attack. The track agent may carry on the fuzzy synthetical assessment to infer these knowledge using data relative to network connections and program executing in the system and so on. After collecting the information, the information-collecting agent will return the report to the intermediate agent and announce the collecting information in the system database, then the track agent shifts to another target system in the track route and activates a new information-collecting agent. If the track agent arrives at the beginning of invasion route, or it cannot shift to other places, or other track agent has already arrived at the destination route, it will return to the intermediate agent. When the monitor detects that many MLSIs are discovered in the identical target system in the short time. or, if the monitor detects many MLSIs in many target systems, intermediate agent will allocate track agent to all MLSIs and carries on the track according to the above work process.

4 Merit of invasion detection system based on mobile agent and fuzzy synthetical assessment technology

Invasion detection system based on mobile agent and fuzzy synthetical assessment technology can solve more complicated and fuzzy problems with precise mathematical tool by using the fuzzy synthetical assessment technology to integrate numerous detection characters, and obtain quite precise results. The false alarm rate of this method is low, and can detect accurately known attacks and unknown attacks as well.

In many traditional network invasion detection systems, each target system transmits its system log to the invasion detection server, and the entire system log is analyzed by the server to judge whether invasion has happened. Vast data of system log are transmitted daily, but most of these data has nothing to do with the invasion. Therefore, in the large network, this kind of invasion detection system is extremely low in efficiency and wastes massive system resources. However the mobile agent can shift automatically to the target system and collect only the information relative to the invasion, so the system log is no longer transmitted to the server. The communication load of system is reduced enormously, and the working efficiency is increased.

References

[1] Zh.J.Tang,J.H.Li, Invasion Detecting Technology[in Chinese], Tsinghua University press, Beijing, 2004.4

[2] G.K.Wei, Brief Analysis Based on Unusual Invasion detection Technology[in Chinese], Journal, Computer Engineering and Design, 2005.1

[3] J.Chen,M.Luo,H.G.Zhang, the Outline of Invasion detection technology[in Chinese], Journal, Computer Engineering and application,2004.2

[4] Catalyst3560Switch Software Configuration Guide,Cisco JOSRelease12.2(20)SE.[EB/OL] .http://cisco.cexx.net/ch/45.html. 2004.3

[5] Galen A Grimes. Network security managers‘preferences for the Snort IDS and GUI,Journal .Network Security, 2005.4

[6] Anderson J P. Computer security thread monitoring and surveillance .Fort Washington,PA: Jame P Anderson Co, 1980

[7] Denning D E. An,intrusion-detection model Journal.IEEE Transaction on Software Engineering, 1987

[8] Snapp S R,Brentano J,Dias G V,et al. A system for distributed intrusion detection[A] .Proceedings of the IEEE COMPCON91. San Francisco,CA: IEEE, 1991

[9] Hoglund G,Butler J. Rootkits:Subvertingthe Windows Kernel .Boston: Addison-Wesley, 2005

[10] Leu F Y,Yang T Y. A host-based real-time intrusion detection system with data mining and forensic techniques .IEEE37th Annual2003International Carnahan Conference on Security Technology. Tai wan. 2003

[11] Deng P S,Wang J H,Shieh W G,et al. Intelligent Automatic Malicious Code Signatures Extraction .IEEE37th Annual2003International Carnahan Conference on Security Technology. Tai wan. 2003

[12] Mukherjee B,Heberlein L T,Levitt K N,et al. Net workintru-sion detection .IEEE Network, 1994, 8 (3) :26～41

Cloud-based Trust Management Model in Open Networks*

Zhang Changlun[1] Liu Yun[1] Zeng Ping[2]

1 School of Electronics and Information Engineering Beijing Jiaotong University, Beijing, 100044, China
Email: 05111045@bjtu.edu.cn

2 Department of Communication Engineering, Beijing Electronic Science and Technology Institute
Beijing, 100070, China

Abstract

Trust cloud model integrates the fuzziness and randomness in a linguistic term to a unified way, and solves the problem of uncertainty in the description and reasoning of trust properly in open networks. Aiming at the limitation of the reasoning mechanism of the existing cloud-based trust model, new trust management model and its reasoning mechanism are proposed for the discrete and continuous trust metric. The reasoning mechanism of trust cloud can deal with the trust recommendation and synthesis of multiple trust paths, and can implement the propagation of trust relationship.

Keywords: trust management; cloud model; linguistic; fuzziness; open networks

1 Introduction

As an important strategy to improve the security in open networks, trust management has been applied widely in public key management, e-commerce, p2p networks, and ad hoc networks.

Although there is no clear consensus on the definition of trust in distributed networks until now, most of the researchers regard that trust is subjective and uncertain. Using various mathematic methods, researchers established many models [1-9] to deal with uncertainty in trust. These models can mainly be divided into probability, fuzzy and cloud-based model.

Probability-based model [2-4] is the basic and extensively applied trust model, which uses simple probability statistics and posteriori probability approach to describe trust relationship. Fuzzy-based model [5-7] considers uncertainty, more accurately, fuzziness, and uses fuzzy logic to deal with trust related problems. Both probability and fuzzy-based trust models are accurate theory, since they assign each element a certain value as probability or membership degree.

Distinguished from previous trust models, cloud-based trust management model [8,9] takes account of trust uncertainty and describes trust degree and trust uncertainty in a uniform form i.e. cloud [10]. It seems more rational than other models. However, the previous cloud-based trust models have many limitations in trust reasoning, such as no trust combination for discrete metric and no consideration the weight of each cloud in the trust reasoning for continuous metric.

In this paper, we aim at the limitation of the reasoning mechanism of the existing cloud-based trust model, proposed new trust management model and its reasoning mechanism for discrete and continuous trust metric.

The remainder of the paper is organized as follows. Section 2 will review preliminary of cloud model. Section 3 describes discrete metric trust cloud model. Section 4 describes continuous metric trust cloud model. Section 5 gives trust decision approaches for trust cloud model, and the conclusion follows in Section 6.

2 Cloud Model

Cloud model [10] is a model of the uncertain

* The research is supported by National Natural Science Foundations of China under Grant Nos.60572035 and the Open Fund of Key Laboratory of Information Security and Privacy (KYKF200703), Beijing Electronic Science and Technology Institute.

transition between a linguistic term of a qualitative concept and its numerical representation. In short, it is a model of the uncertain transition between qualitative and quantitative. It effectively integrated the fuzziness and randomness in a linguistic term to a unified way. Until now, the cloud model has been applied in many fields successfully, such as automatic control and data mining, etc.

Let X be a set as the universe of discourse, and T a linguistic term associated with X. The membership degree of x in X to the linguistic term T, $C_T(x) \in [0,1]$ is a random number with a stable tendency. A compatibility cloud [10] is a mapping from the universe of discourse X to the unit interva1 [0, 1]. That,

$$C_T(x) \to [0,1], \quad \forall x \in X \quad x \to C_T(x). \tag{1}$$

A trust cloud [10] is a normal cloud to quantify trust relationship between two entities, indicating how much and how surely one is trusted by the other. Formally, trust cloud can be denoted as:

$$C(Ex, En, He), \quad 0 \le Ex \le 1, 0 \le En \le 1, 0 \le He \le 1. \tag{2}$$

Here, Ex is the trust expected value, En is trust entropy, and He is the trust hyper entropy.

3 Discrete Metric Trust Cloud Model

3.1 Trust model

Trust can be evaluated in very different ways. PGP [11] employ linguistic descriptions of trust relationship, three levels of trust are assigned to someone else's public key: Complete trust, Marginal trust, Untrusted. In some other schemes, discrete numerical values are assigned to measure the level of trustworthiness.

Generally, trust level can be described in linguistics T= {T_1,T_2,…,T_M }. Here, we employ following linguistic descriptions:

T= {Distrusted, Minimal trust, Average trust, Good trust, Complete trust}.

The traditional trust level can be described in trust base cloud.

Definition 1 Trust base cloud (TBC) is the description to traditional trust level in cloud model.

TC{TC_1 (Ex_1, En_1, He_1), TC_2 (Ex_2, En_2, He_2),…, TC_M (Ex_M, En_M, He_M)},TC_i (Ex_i, En_i,He_i)($1 \le i \le M$).For example, TC_1 (0, 0.1, 0.01),TC_2 (0.125, 0.1, 0.01), TC_3 (0.375, 0.1, 0.01),TC_4 (0.625, 0.1, 0.01),TC_5 (0.875, 0.1, 0.01), TC_6 (1, 0.1, 0.01). Figure 1 shows the trust base cloud.

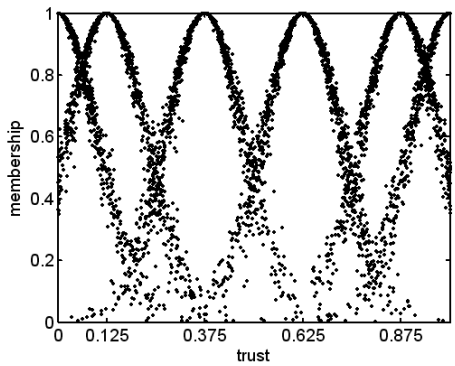


Figure 1 Trust base cloud

The parameters of trust base cloud can be acquired by many approaches such as experts' experience or fuzzy synthetic evaluation [7, 8].

From figure 1, we can see that one same trust degree can belong to different trust base cloud. For example, trust degree 0.125 belongs to Distrusted, Minimal trust and Average trust, but the memberships are different to different base cloud. To the same base cloud, same membership can has different trust degree.

In order to describe the relationship of trust degree and trust base cloud, we introduce the accept factor according to the principle "3En" of normal distribution function.

Definition 2 (**Accept Factor)** Given trust degree t and trust base cloud TC_i (Ex_i, En_i, He_i), denote $\varphi = (Ex_i - 3En_i\ ,\ Ex_i + 3En_i)$, then

$$\delta(t, TC_i) = \begin{cases} ((t + 3En_i - Ex_i)/(6En_i)) \times 100, & t \in \varphi \\ 0, & t \notin \varphi \end{cases} \tag{3}$$

is the accept factor of trust degree t to trust base cloud TC_i.

The value of accept factor is in [0,100]. It shows the degree of a trust relationship to trust base cloud and the transition of trust relationship from low level to high level. The default value of accept factor is 50.

3.2 Trust reasoning

Entity always needs to interact with stranger, which has not direct trust relationship. In this case, it always needs to get recommendation from others, so trust propagation is needed.

3.2.1 Trust recommendation

Definition 3 (Trust Logic Operator) If the trust relationship of entity A to entity B is $TC_{AB}(TC_i\ (Ex_i, En_i, He_i), \delta_1)$, and the trust relationship of entity B to entity C is $TC_{BC}(TC_j\ (Ex_j, En_j, He_j), \delta_2)$, then trust logic operator of entity A to entity C is:

$$Ar^{scsr}(TC_{AB}(TC_i,\delta_1)),TC_{BC}(TC_j,\delta_2)) = \begin{cases} Ar^{scsr}(\delta_1,TC_i,TC_j), & Ex_i \geq Ex_j \\ Ar^{scsr}(\delta_2,TC_j,TC_i), & Ex_i < Ex_j \end{cases} \quad (4)$$

Here, Ar^{scsr} is a single condition and single rule operator [10].

Definition 4 (Trust Recommendation) If the trust relationship of entity A to entity B is $TC_{AB}(TC_i\ (Ex_i, En_i, He_i), \delta_1)$, and the trust relationship of entity B to entity C is $TC_{BC}(TC_j\ (Ex_j, En_j, He_j), \delta_2)$, then trust recommendation of entity A to entity C is indicate by $\otimes$:

$$TC_{AC}\ (TC_k\ (Ex_k, En_k, He_k), \delta_3) = TC_{AB}\ (TC_i\ (Ex_i, En_i, He_i), \delta_1) \otimes TC_{BC}((TC_j\ (Ex_j, En_j, He_j), \delta_2) \quad (5)$$

if $Ex_j \geq Ex_i$,

$TC_k\ (Ex_k, En_k, He_k) = TC_i\ (Ex_i, En_i, He_i)$;

Else,

$TC_k\ (Ex_k, En_k, He_k) = TC_j\ (Ex_j, En_j, He_j)$.

$\delta_3 = \min(\delta(t, Ar^{SCSR}), \delta_1, \delta_2)$, trust logic operator of entity A to entity C, $\delta(t, Ar^{SCSR}) = \delta\ (mean(t_1, t_2), TC_k)$.

In the reasoning of trust recommendation, the result can not exceed the trust degree and accept factor of any one, trust recommendation result less than the trust of recommenders strictly, we say it pessimism trust; if we define the accept factor

$$\delta_3 = \max(\delta(t, Ar^{SCSR}), \delta_1, \delta_2). \quad (6)$$

The value of δ_3 can not exceed the trust degree and accept factor of any one, trust recommendation result does not less than the trust of recommenders strictly, we say it optimism trust.

Figure 2 shows the process of trust recommendation.

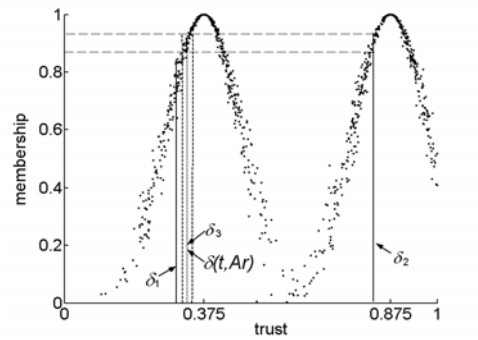


Figure 2 Trust recommendation

3.2.2 Trust Combination

Definition 5(Trust Combination) There are two trust path from entity A to entity B, one is $TCp_1\ (TC_i\ (Ex_{iB}, En_i, He_i), \delta_1)$, the other is $TCp_2\ (TC_j\ (Ex_j, En_j, He_j), \delta_2)$, then the trust combination of entity A to entity B is indicate by $\oplus$

$$TC_{AB}(\ (TC_k\ (Ex_k, En_k, He_k), \delta_3) = TCp_1(TC_i\ (Ex_i, En_i, He_i), \delta_1) \oplus TCp_2(TC_j\ (Ex_j, En_j, He_j), \delta_2), \quad (7)$$

Here:

1) $\delta_3 = max\ (\delta\ (mean\ (\delta(\delta_1^{-1}, \delta_2^{-1})), TC_m))$, TC_m is trust base cloud.

2) TC_k is the base cloud that has maximum accept factor.

4 Continuous Metric Trust Cloud Model

Different from discrete trust base cloud, continuous trust metric means the parameters Ex, En, He of cloud are continuous values in [0, 1].

In this section, we give the approach of computing trust cloud and trust reasoning.

4.1 Trust computation

Cloud does not exist naturally, so it is necessary to compute the parameters Ex, En and He from objective data. The acquiring of objective data can be from existing schemes [3, 4, 7, 8]. Here, we define the computing result of objective data as basic trust value,

which is a continuous real number in [0, 1], and can be simple probability value or statistic results.

Based on trust value, we propose an approach to compute the parameters of trust cloud. Given a serial basic trust values $t_i, 1 \le i \le n$, the detailed computing processing of trust cloud is described in the following Procedure:

1) According to basic trust values t_i, $1 \le i \le n$, compute

Average: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} t_i$,

First order absolute central moment: $\frac{1}{n}\sum_{i=1}^{n} | t_i - \overline{X} |$,

Variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(t_i - \overline{X})^2$.

2) Compute trust expectation: $E_x = \overline{X}$.

3) Compute trust entropy: $E_n = \sqrt{\frac{\pi}{2}} \times \frac{1}{n}\sum_{i=1}^{n} | t_i - E_x |$.

4) Compute trust hyper entropy: $H_e = \sqrt{S^2 - E_n^2}$.

4.2 Trust reasoning

Similar to the discrete trust metric, trust recommendation and trust combination for continuous metric are defined.

4.2.1 Trust Recommendation

Supposing there are three nodes *A*, *B* and *C*, trust relationship of *A* to *B* is $TC_1(Ex_1, En_1, He_1)$, and *B* to *C* is $TC_2(Ex_2, En_2, He_2)$. We define the trust relationship of *A* to *C* as follows:

$$T\mathrm{C}(Ex, En, He) = T\mathrm{C}_1(Ex_1, En_1, He_1) \otimes T\mathrm{C}_2(Ex_2, En_2, He_2). \quad (8)$$

Where,

$$\begin{cases} Ex = Ex_1 Ex_2 \\ En = En_1 Ex_2 + En_2 Ex_1 \\ He = He_1 Ex_2 + He_2 Ex_1 \end{cases} \quad (9)$$

For example, the trust clouds are

$$T\mathrm{C}_1(Ex_1, En_1, He_1) = (0.4, 0.1, 0.01),$$

$$T\mathrm{C}_2(Ex_2, En_2, He_2) = (0.7, 0.15, 0.02).$$

The recommendation trust cloud is computed as follows:

$$Ex = 0.4 \times 0.7 = 0.28,$$

$$En = 0.1 \times 0.7 + 0.15 \times 0.4 = 0.13,$$

$$He = 0.01 \times 0.7 + 0.02 \times 0.4 = 0.015.$$

$$T\mathrm{C}(Ex, En, He) = T\mathrm{C}(0.28, 0.13, 0.015).$$

Figure 3 shows the cloud of trust recommendation.

4.2.2 Trust computation

If there are two different trust paths from node *A* to *C*, one is from node *B* to *C*, the other is from node *D* to *C*, the two trust clouds are $TC_1(Ex_1, En_1, He_1)$ and $TC_2(Ex_2, En_2, He_2)$ respectively. We define the trust relationship of node *A* to *C* as Eq.(10).

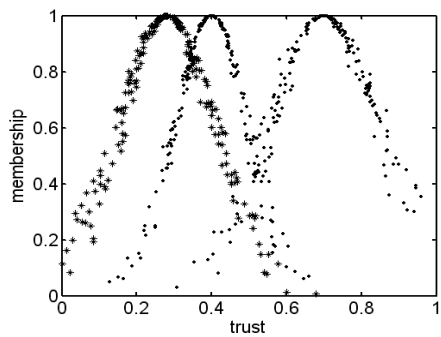


Figure 3 Trust cloud of recommendation

$$T\mathrm{C}(Ex, En, He) = T\mathrm{C}_1(Ex_1, En_1, He_1) \oplus T\mathrm{C}_2(Ex_2, En_2, He_2) \quad (10)$$

Where,

$$\begin{cases} Ex = \dfrac{Ex_1 \times En_1^{'} + E_{x_2} \times En_2^{'}}{E_{n_1}^{'} + E_{n_1}^{'}} \\ En = E_{n_1}^{'} + E_{n_2}^{'} \\ He = \dfrac{He_1 \times En_1^{'} + He_2 \times En_2^{'}}{En_1^{'} + En_2^{'}} \end{cases} \quad (11)$$

$En_1^{'}$ and $En_2^{'}$ are defined as follows:

$$\begin{cases} En_1^{'} = \frac{1}{\sqrt{2\pi}} \int C_{TC_1}^{'}(x)dx \\ En_2^{'} = \frac{1}{\sqrt{2\pi}} \int C_{TC_2}^{'}(x)dx \end{cases}, \quad (12)$$

$C_{TC_1}^{'}(x)$ and $C_{TC_2}^{'}(x)$ are computed according to the formulas (13) and (14) respectively:

$$C_{TC_1}^{'}(x) = \begin{cases} C_{TC_1}(x) & if \quad C_{TC_1}(x) \ge C_{TC_2}(x) \\ 0 & other \end{cases}, \quad (13)$$

$$C_{TC_2}^{'}(x) = \begin{cases} C_{TC_2}(x) & if \quad C_{TC_2}(x) > C_{TC_1}(x) \\ 0 & other \end{cases}. \quad (14)$$

In order to simple the computation of $En_1^{'}$ and $En_2^{'}$, we can replace the above integral with subsection linear approximation. According to the "3En" principle of normal distribution, we select Ex, $Ex \pm En$, $Ex \pm 2En$, $Ex \pm 3En$, and to build subsection linear approximation

function.

For example, the values of TC_1 and TC_2 are same as above. The curves $C_{TC_1}(x)$ and $C_{TC_2}(x)$ intersect at $x' = 5.2$. Compute En_1'=0.0885, En_2'=0.1327, then

$$Ex = \frac{0.4\times 0.0885 + 0.7\times 0.1327}{0.0885 + 0.1327} = 0.58,$$

$$En = 0.0885 + 0.1327 = 0.2212,$$

$$He = \frac{0.01\times 0.0885 + 0.02\times 0.1327}{0.0885 + 0.1327} = 0.016,$$

$$TC(Ex, En, He) = TC(0.58, 0.2212, 0.016).$$

Using above subsection linear approximation to compute the approximation values of En_1' and En_2':

$$En_1' \approx 0.08725,$$

$$En_2' \approx 0.1309.$$

Then,

$$Ex = \frac{0.4\times 0.08725 + 0.7\times 0.1309}{0.08725 + 0.1309} = 0.58,$$

$$En = 0.08725 + 0.1309 = 0.218,$$

$$He = \frac{0.01\times 0.08725 + 0.02\times 0.1309}{0.08725 + 0.1309} = 0.016.$$

$$TC(Ex, En, He) \approx TC(0.58, 0.218, 0.016).$$

From the example, we can see that the approximation value of trust cloud is very similar to the integral value. The approach of replacing the integral with linear approximation is feasible.

Figure 4 shows the cloud of trust combination in multi-path recommendation.

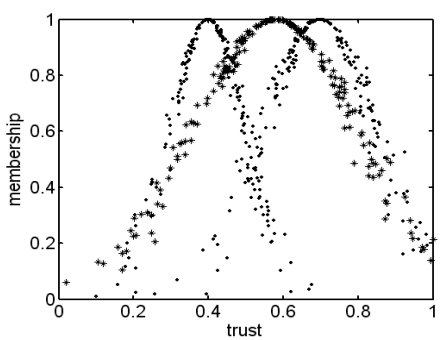


Figure 4 Trust cloud of combination

5 Trust decision

Threshold values are often used in security mechanisms. In order to make a cloud-based trust decision, we also employ the simple policy.

The first approach is to give the thresholds values of *Ex*, *En*, and *He* respectively. According to the thresholds, the node makes decision.

The second approach is to map trust clouds to linguistic descriptions of trustworthiness. Since cloud model describes trust from aspects of trustworthiness and uncertainty, linguistic descriptions also contain these two aspects, for example, “it can be trusted very much, but this is uncertain”. The details of the approach can refer to [9].

6 Conclusions

To provide efficient approach for trust management in open networks, we proposed new trust management model and its reasoning mechanism for discrete and continuous trust metric to evaluate the trustworthiness among entities. For discrete metric, trust base cloud and accept factor is proposed to describe the uncertainty of trust. Based on the basic trust values from objective data, a simple method to compute the trust cloud is given for continuous metric. Considering the weight of trust cloud, a serial trust cloud operation to deal with the trust recommendation and trust propagation are proposed, which seems more rational than previous models.

References

[1] Blaze M, Feigenbaum J, and Lacy J, “Decentralized Trust Management,” Proc. of the Symposium on Security and Privacy. IEEE Computer Society Press, pp.164~173, 1996

[2] Abdu1-Rahman A, Hailes S, “A Distributed Trust Model,” Proc. of the 1997 New Security Paradigms Workshop. Cumbria: ACM Press, pp.48~60, 1998

[3] Ruidong Li, Jie Li, and Peng Liu et al, “An objective trust management framework for mobile ad hoc networks,” Proc. VTC2007, pp.56~60

[4] Sun Y, Yu W, and Han Z, et al, “Information theoretic framework of trust modeling and evaluation for ad hoc networks,” IEEE Journal on Selected Areas in Communications, Vol. 249, No.2, pp.305~319, 2006

[5] Song SS, and Hwang K, “Fuzzy trust integration for security enforcement in grid computing,” Proc. NPC 2004, LNCS 3222, pp. 9~21

[6] C Castelfranchi, R Falcone, and G Pezzulo, “Trust in

Information Sources as a Source for Trust: A Fuzzy Approach,"Proc. AAMAS'03, pp.89~96

[7] Tang Wen, Chen Zhong, "Research of Subjective Trust Management Model Based on the Fuzzy Set Theory," Journal of Software, Vol. 14, No.8, pp.1401~1408, 2003

[8] Meng Xiangyi, Zhang Guangwei, and Liu Changyu, "Research on subjective trust management model based on cloud model," Journal of System Simulation, Vol. 19, No.14, pp.3310~3317, 2007

[9] He Rui, Niu Jianwei, and Hu Jianping, "Modeling trust with uncertainty for open networks," Journal of Beijing University of Aeronautics and Astronautics, Vol. 30, No.11, pp.1125~1128, 2004

[10] Li DY, Meng HJ, and Shi XM, "Membership clouds and membership clouds generator," Journal of Computer Research and Development, Vol. 32, No.6, pp.15~20, 1995

[11] P. R. Zimmermann, The Official PGP User's Guide. Cambridge, MA: MIT Press, 1995

Graph Partitioning Technique for Separating Nets in Single-row Networks

Norazaliza Mohd. Jamil[1] Noraziah Ahmad[1] Shaharuddin Salleh[2]

1 Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang,
Locked Bag 12, 25000, Kuantan, Pahang, Malaysia
Email: norazaliza@ump.edu.my, noraziah@ump.edu.my

2 Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310, Johor Bahru, Malaysia
Email: ss@mel.fs.utm.my

Abstract

Single-row routing is fundamentally a routing technique for pairs of nodes arranged in a single-row axis. It contributes in the printed circuit board (PCB) design. The main objective in single-row routing is to achieve minimum congestion arising from the number of horizontal tracks in the network. Optimal results for a single layer network have been achieved through our previous model called ESSR. However, a single layer model suffers from non-tolerable lower bound values with high congestion depending on the network size. These results may further be improved by partitioning the network into two or more layers. In this paper, we propose a technique based on the graph partitioning concept for partitioning the nodes from a single-row network into several layers of planar graphs. The experiment result shows that the proposed technique is able to decrease the network congestions.

Keywords: Single-row routing, ESSR, graph partitioning, maximum clique, Kernighan-Lin algorithm

1 Introduction

The problem of designing a complex multilayer printed circuit board (PCB) is divided into two main subproblems; layering and routing[1]. This is very crucial in wire interconnection for efficient electronic systems. As the circuit complexity is increasing, the PCB routing problem becomes more challenging. Usually, manual routing effort is applied when traditional routing algorithm can not handle these challenges effectively[2]. The main target in the layering problem is to achieve the minimum number of layers such that there is no wire intersection on each layer. This is because the additional number of layers has a tendency to increase the cost production and reduce the reliability[1]. One of the routing methods in PCB design is single-row routing. The objective is to find the optimum wiring from a set of pins which are aligned in a single node axis. Basically, graph theoretic is used widely in this problem.

In 1999, Salleh and Zomaya proposed a solution to the optimum single-row network, namely enhanced simulated annealing for single-row routing (ESSR) [3]. The algorithm is based on the simulated annealing technique. The design of multi-layer printed circuit boards is very significant in the construction of complex electronic systems [4]. This has led to the use of graph partitioning methods for layering problems. One of the methods is Kernighan-Lin algorithm (K-L algorithm) which is an iterative, two-way, and balanced partitioning (bisectioning) heuristic [5].

In this paper, we propose a technique for partitioning a net list into several layers by applying a maximum clique approach and the K-L algorithm. This paper is organized into seven sections. We describe the problem formulation in Section 2. The single-row routing problem is discussed in Section 3. Section 4 presents the graph partitioning algorithm, maximum clique and the K-L algorithm. This is followed by a description of our approach for solving the layering problem of single-row routing in Section 5. The

experimental results from the simulations are shown in Section 6. Finally, we conclude this research in Section 7.

2 Problem Formulation

Given a set of nets $\{N_i\}, i = 1, 2, ..., m$ where each net N_i is made up of a pair of pins v_j, $j = 1, 2, ..., m/2$, aligned in a single-row axis as shown in Figure 1. A single-layer row-routing network is obtained by drawing the nets using horizontal and vertical lines from left to right in such a way that the lines do not cross each other.

In this paper, we extend the problem by partitioning the nets into two or more layers in order to reduce the congestion in the network. A technique based on graph partitioning method using the K-L algorithm has been deployed to produce optimum results. We also deploy a maximum clique approach in determining the number of layers needed for the routing.

For example, consider the single-layer row routing network of 10 nets as shown in Figure 1. The nets in the network have been partitioned into two layers for optimum designs and the layout for both layers is shown in Figure 2.

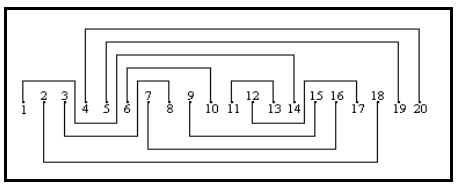


Figure 1 Single-layer row routing network

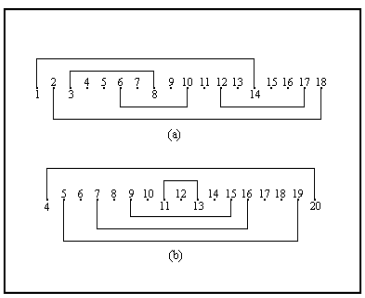


Figure 2 (a) Realization for the first layer, and (b) Realization for the second layer

3 Single-Row Routing Problem

The single- row routing technique is applied to design the routes of wires so that each electronic component could communicate to each other. The single-row routing problem can be put in the context of a single-layer PCB such that each net in the network is not overlap.

In single-row routing problem, we are given a set of nodes, $V = \{1, 2, 3, ..., n\}$ arranged horizontally along a single-row axis from left to right. Basically, each net in the single-row routing consist of two nodes. The nodes are assumed to be the electrical components, while the nets are visualized as the conductor paths that exist between them. The objective in single-row routing is to route a net list $L = \{N_1, N_2, N_3, ..., N_m\}$ with minimum number of tracks and the nets are not allowed to cross each other [4].

In ESSR [3], the energy function is expressed as the total lengths of all tracks as follows:

$$E = \sum_{i=1}^{m} \sum_{r=1}^{m_i} \left| h_{i,r} \right| \quad (1)$$

In Eq. (1), $h_{i,r}$ is the height of segment r in net i, while m is the number of nets and m_i is the number of segments in the net N_i, for $i = 1, 2, ..., m$. The number of overall street congestion and the number of doglegs are denoted by Q and D respectively.

4 Graph Partitioning Algorithm

The graph partitioning problem is the problem of finding a partition of vertices into a number of disjoint subsets in such a way that the number of edges crossing the subsets is minimized. We consider the problem of finding a balanced k-way partition where the size of the largest subset and the smallest subset differ by at most one. Many approaches have been proposed to solve the graph partitioning problem, such as the Kernighan-Lin method, K-means algorithm, neural network, or combination of different methods [5].

Maximum clique

A clique is a subset S of a graph G where every pair of vertices in S is joined by at least one edge and there is no proper superset of S has this property. In

other words, a clique of a graph *G* is a maximal subset of mutually adjacent vertices in *G*. The maximum clique problem is known to be NP-hard, which appears in many theoretical and practical applications [6].

The problem illustration is shown in the Figure s below. There are two cliques in the graph of Figure 3. In particular, each of the vertex subsets, {1, 4, 5, 6} and {2, 3}induced complete subgraphs, and no proper superset of any of them induces a complete subgraph. The maximum clique of Figure 3(a) is shown by black vertices in Figure 3(b).

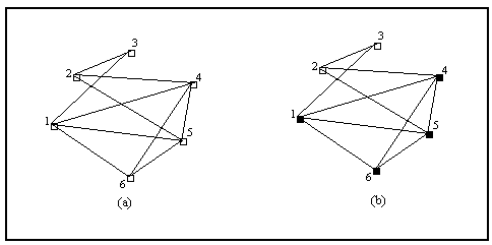


Figure 3 (a) A 6-node 10-edge undirected graph, and (b) the maximum clique consisted of the black vertices

Kernighan-Lin algorithm

K-L algorithm was originally proposed in 1970 for graph partitioning problem. However, later on, it has been used in a number of different fields with a lot of variations [5]. In general, KL algorithm iteratively finds a better solution. It is used to view external against internal cost between nodes of a graph from the given dataset to be clustered. The nodes of the graph are assigned to each of the partitions. The internal cost is the cost of an edge between two nodes within the same partition. The external cost is the cost of the edge of a node in one partition to a node in the other partition.

This algorithm tries to move each node between partitions until it reaches the lowest cost between nodes. In each iteration it carries out a sequence of best moves. The solution that has the lowest cost in the sequence is selected as the solution for the next iteration. However, the solution with inferior cost may also be accepted (uphill moves) in order to help this algorithm escape from local minima.

K-L algorithm is a well-known method for partitioning a graph and much used as a tool for assigning electronic components on circuit boards. It has been proven that this algorithm is very good in producing near- optimal result in relatively short time.

5 Our Approach

We are given a net list $L=\{N_1, N_2, ..., N_m\}$. Each net consist of exactly two pins that are to be made electrically equivalent. For example, consider a net list $L=\{N_1, N_2, ..., N_{16}\}$ in Figure 4. The single-row routing representation of these nets in a single layer which is obtained by using ESSR model is shown in Figure 5 with

E = 54, Q = 6, D =10.

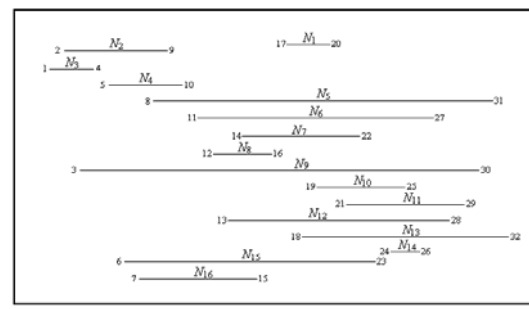


Figure 4 Net list $L=\{N_1, N_2, ..., N_{16}\}$

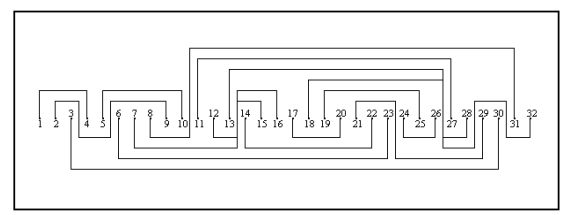


Figure 5 Single-row routing realization of L with E = 54, Q = 6, D =10

The problem is to find a partition of L into the minimum number of subsets $L_1, L_2, ...L_T$ such that each partition L_i $(i=1,2,...T)$ can be realized on a single layer. The nets are to be realized by single row routing in a minimum number of layers by the use of non-overlapping wires that are composed solely of horizontal and vertical segments.

The algorithm begins by constructing a containment graph to visualize the connection between the net lists. Each net is represented as a node in the containment graph. In the final solution, we can get lower number of doglegs by using the proposed technique compared to interval and overlap graph. Recall the containment

graph, two nodes are adjacent if one of the corresponding intervals is strictly contained in the other one. The containment graph associated with the net list in Figure 4 is depicted in Figure 6.

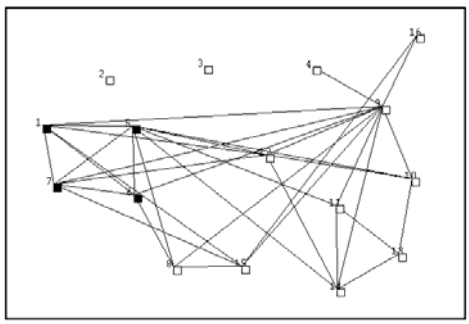


Figure 6 Maximum clique of the containment graph

The partition among the layers is determined following the principle of the maximum clique. All nodes in the maximum clique should be assigned layers first. In this stage, we can estimate the number of layers needed for the routing and the solution of initial partitions. The algorithm that we propose for the layering problem determines a subset of L that can be realized in a single-row routing representation.

We begin with finding the maximum clique of the containment graph. The maximum clique of is shown by black nodes in Figure 6. This algorithm constructs a wire layout by applying all nodes in the maximum clique into the first layer, L_1. The number of layers required for the routing is obtained by dividing the number of nets in the network by the size of the maximum clique.

Once L_1 has been determined, we determined a subset from $L - L_1$ to layout in layer 2. In this example, our algorithm selects nets $N_1, N_5, N_6,$ and N_7 for L_1 and the size of the maximum clique is 4. Since there are 16 nets in the problem, therefore we need 16/4 = 4 layers for the routing.

To determine L_2, the nodes corresponding to layer 1 are removed from the containment graph. Then the maximum clique for the remaining nodes is determined and all nodes in that maximum clique are imposed on the second layer. This process is repeated until we have determined $L_1, L_2, ...L_T$, and there are no more cliques left in the containment graph.

Figure 7 shows the remaining nodes after the nets in layer 1 has been removed from the containment graph. Then, the maximum clique for this Figure is determined and we have $L_2 = \{N_9, N_{11}, N_{14}\}$. This process is repeated until we have decided L_1, L_2, L_3, and L_4, and there are no more cliques left in the containment graph. Now, we can organize all the nets into their partition as follows: $L_1 = \{N_1, N_5, N_6, N_7\}$, $L_2 = \{N_9, N_{11}, N_{14}\}$, $L_3 = \{N_{10}, N_{12}\}$ and $L_4 = \{N_8, N_{15}\}$.

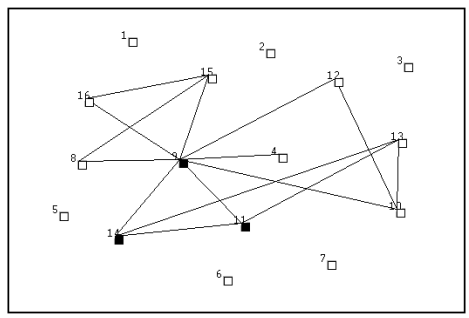


Figure 7 Layer 2: {N_9, N_{11}, N_{14}}

To avoid partitions from becoming heavily unbalanced, the remaining nets are then randomly distributed to occupy available subsets such that $L = L_1 \cup L_2 \cup ... \cup L_T$. Therefore, we can get the initial partitions for this problem example which are: $L_1 = \{N_1, N_5, N_6, N_7\}$, $L_2 = \{N_9, N_{11}, N_{14}, N_2\}$, $L_3 = \{N_{10}, N_{12}, N_3, N_4\}$, and $L_4 = \{N_8, N_{15}, N_{13}, N_{16}\}$.

The partition found by the maximum clique approach is usually near a local optimum. Since a k-way partition has already been obtained, the Kernighan-Lin algorithm appears to be a laudable choice. At the end of the algorithm, we optimize the best partitioning found so far using the Kernighan-Lin algorithm, and return the result as our final k-way partition. This restricted the size of k to a power of two. The Kernighan-Lin algorithm consists of several iterations. Optimization is performed until no improvement in the cut size can be found.

The modification is accepted only if the total sum of energy value, E for each layer that used the Kernighan-Lin algorithm is lower than the current solution. Otherwise, we stick to the current solution where the maximum clique approach is applied. Although the cut sizes between k subsets may not be

optimal, we believe that it is good enough since the congestion in the original problem is reduced greatly. The layout shown in Figure 8 was obtained by applying the above algorithm on the previous problem example.

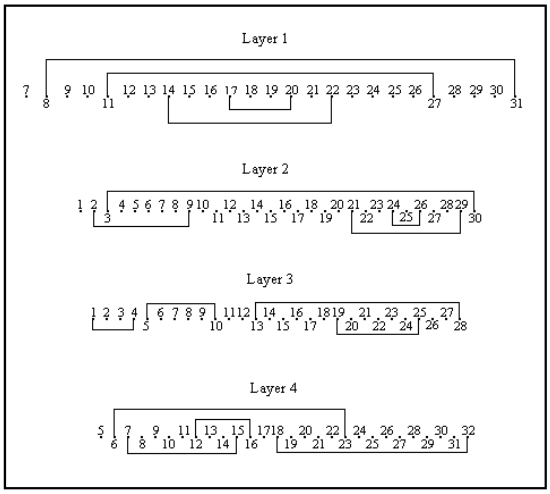


Figure 8 The final solution

6 Results

The results of our experiment are tabulated in Table 1. The number of nets to be connected by single-row routing is shown in the second column. The number of cut edges between partitions is presented in column 7. The output obtained by using ESSR technique is in the form of energy *E*, overall street congestion *Q* and number of doglegs *D* which are shown in columns 8, 9, and 10 respectively. The value of *E*, *Q*, and *D* in the table represents the total sum of the corresponding value for each layer. Nine random data sets are generated to contains 8 to 40 nets and has been partitioned into 2, 3,

Table 1 Sample Results from the Simulations

No.	# nets	Single-layer			# layers	Cut edges	Multi-layer		
		E	*Q*	*D*			*E*	*Q*	*D*
1	8	25	4	4	3	0	11	5	1
2	10	31	5	4	2	7	16	5	0
3	12	31	5	3	3	18	16	6	0
4	16	54	6	10	4	23	20	7	0
5	20	82	6	14	5	11	27	9	0
6	24	168	10	32	4	67	41	11	0
7	28	172	8	34	4	73	56	13	4
8	32	238	12	47	4	130	56	13	4
9	40	441	14	81	6	174	79	18	9

gather the data for each entry in the table. Each graph 4, 5, or 6 layers. The result obviously show that the energy in the single-layer network is reduced greatly when the original problem is transformed into multi-layer representation.

7 Conclusions

This paper presents on how to use maximum clique approach and Kernighan Lin heuristic partition algorithm within the layering problem of single-row routing. We presented the algorithm for partitioning the net lists into several layers. This is because the wiring of single-layer PCB is more congested rather than the multi-layer PCB of the same network. By using this technique we can reduce the congestion even in a large network.

References

[1] J.F. Naveda, K.C. Chang, and H.C. Du, “A new approach to multi-layer PCB routing with short vias,” Proceedings of the23rd Conference on Design Automation, 40(1), pp. 696-701, 1986

[2] M.M. Ozdal and M.D.F. Wong, “Simultaneous Escape Routing and Layer Assignment for Dense PCBs,” Proceedings of the 2004 IEEE/ACM International conference on Computer-aided design, pp. 822-829, 2004

[3] S. Salleh, B. Sanugi, H. Jamaluddin, S. Olariu and A.Y. Zomaya, “Enhanced Simulated Annealing Technique for the Single-Row Routing Problem,” The Journal of Supercomputing, Kluwer Academic Publishing, Netherlands, pp. 285-302, 2002

[4] A.Y. Zomaya, D. Patterson, and S. Olariu, “Sequential and Parallel Meta-Heuristics for Solving the Single Row Routing Problem,” Journal of Cluster Computing, 7(2), pp. 123–139, 2004

[5] E.S. Smestad, Kernighan-Lin Heuristic in an IDC, Gjovik University College, 2006

[6] D. Kumlander, “An Approach for the Maximum Clique Finding Problem Test Tool Software Engineering,” Proceeding of the IASTED International Conference on Software Engineering(SE’07), pp. 297 – 301, 2007

Research of Algorithm for Network Topology Discovery *

Qiu Jianlin He Peng Gu Xiang Chen Jianping Li Feng

School of Computer Science and Technology, Nantong University, Nantong, Jiangsu 226019, P.R. China
Email: qiu.jl@ntu.edu.cn

Abstract

This paper gives the method of network topology discovery. Three kinds of centralized automatic discovery algorithms for network topology were introduced, which include building network top logy base on ICMP or ARP protocol and utilizing SNMP protocol to visit in MIB (management information base) to construct network topology. A kind of better network top logy algorithm was proposed. It was based on the three algorithms which can discover network topology accurately, unabridged and efficiently. In this paper, we give the data structure and steps of this algorithm that was detailed described.

Keywords: Topology Discovery, MIB, SNMP, ARP, ICMP, OSPF

1 Introduction

Network topology discovery is the base for network management functions. It is very important to monitor the whole network, acquire information of network completely, insure network to work efficiently and work stably.

This paper first analyses several popular network topology discovery algorithms, discusses their length and limitation, describing the applying scope of each one. Then this paper introduces the improved network topology discovery algorithm, which is based on ICMP, ARP and SNMP. And the paper also gives a implementing model of the improved algorithm.

2 Topology Discovery Algorithm Based on ICMP

ICMP[1] protocol is a important tool to get route information. People always make use of the ICMP echo reply news to examine the activity and attainability of the network devices. And make use of the information of ICMP time exceed or the port unreachable, and the information of TTL data-field in IP protocol, to search information of assigned host. If “ping” every IP address available in a net segment by polling way, it is easy to find out active devices in this net segment at current time according to reply for “Ping” command. Then “Traceroute” every active device IP address and record the results of every “Traceroute” operation. Finally, we can get the whole network topology by analyzing the information returning by former operation.

Almost all networks devices support ICMP protocol who are support TCP/ IP. So it is easy to implement the network topology discovery algorithm based on ICMP. And at the topology discovery moment, it tests the current activity status of network devices. Because only the active devices could response the ICMP reply messages. In addition, we can take broadcasting way or multi-threading way to cut the cost and mitigate burden of management work’s station and network evoked by “ping” every IP address. Of cause, this method also has limitation as follow:(1) " Ping" operation can easily find out network devices, but it is not able to discover the devices ICMP messages passed

* This work was supported in part by the Science Foundation of JiangSu Grants BK2001130 and the Science Foundation of JiangSu Education Department Grants (03KJB520103, 05KJD520166, 06KJB520090) and the Science Foundation of Nantong of JiangSu Grants (AL2007033, K2006008, C3018, C3024) and the Modern Education Technology Found of JiangSu Education Department Grants 2004-METR-27.

by when it moved from source node to destination node. “Traceroute” operation an help us know which router devices ICMP messages passed by, but it is hark to help us make connecting relation of devices clear. (2) This algorithm is not targeted. Event if we get response messages “Ping” command reply, it is still hard to judge the circumstance of the subnets. (3) Because some fire walls forbid ICMP message pass through, this algorithm is not able to search the devices behind the firewall. Although this algorithm can find out network devices quickly, it is difficult to search the topology of network backbone.

3 Analysis of the Method Based on ICMP

Almost all networks devices support ICMP protocol who are support TCP/ IP. So it is easy to implement the network topology searching method based on ICMP. And at the topology searching moment, it tests the current activity status of network devices. Because only the active devices could response the ICMP reply messages. In addition, we can take broadcasting way or multi- threading way to cut the cost and mitigate burden of management work’s station and network evoked by “ping” every IP address.

Of cause, this method also has limitation as follow:

(1) " Ping" operation can easily find out network devices, but it is not able to discover the devices ICMP messages passed by when it moved from source node to destination node. “Traceroute” operation an help us know which router devices ICMP messages passed by, but it is hark to help us make connecting relation of devices clear.

(2) This method is not targeted. Event if we get response messages “Ping” command reply, it is still hard to judge the circumstance of the subnets.

(3) Because some fire walls forbid ICMP message pass through, this method is not able to search the devices behind the firewall. This method can discover network devices quickly, but it is difficult to search the topology of network backbone.

4 Topology Discovery Algorithm Based on OSPF

OSPF is a kind of open shortest path first protocol. It is used broadly in current time. OSPF has two types of the router selecting way. It will choose internal route principle when source node and destination node are in the same AS (autonomous system). And it will choose external route principle when source node and destination node are in different AS.

In the OSPF protocol, there is a special kind of router, which has several ports and the ports’ IP address are in different area. This kind of router is called boundary router. They govern and maintain independent topology database for every area. A topology database could be considered as a global network structure what pictures the relation ship of routers. The topology database records LAS data from all the routers in the same area. We can get a AS’s topology structure by only visiting the boundary routers, because the routers in same AS share the same topology database and share the same information.

OSPF backbone area shoulders the task to broadcast route information to all AS areas. OSPF backbone area is also a area of OSPF, and its ID is Area 0. Area 0 comprises of all boundary routers while each of boundary routers belong to it’s own AS too. Therefore, accessing routers’ topology database in Area 0 equals to accessing topology database of all routers in the AS.

It introduces OSPF protocol into this method that means it implement OSPF protocol by itself. It uses OSPF protocol to communicate with route devices, access topology database of all the boundary routers in Area 0. And makes use of route information obtained the network topology from topology database.

5 Analysis of the Method Based on OSPF

This method only need to access routers in Area

0.It is unlike the method based on SNMP protocol need to access all the routers' MIB in

As. So it is much quick and efficient. In fact, the network of a common enterprise is hardly out of a AS. Thus this method is practical. And it is able to make use of update messages of OSPF link status, detect and report local change of network topology without searching network again. But it has limitation as follow:

(1) Require all the devices in searching area support OSPF protocol.

(2) The searching is only limited in one AS scope.

(3) OSPF has no topology information at port level.

(4) This method is difficult to implement. Because Algorithm OSPF took is very complicated.

This method is quick and efficient. But it is difficult to achieve, and it only available in the scope that all network devices in it support OSPF protocol. In addition, OSPF protocol has no idea about subnet, so it is not able to searching network connections at port level.

6 Topology Discovery Algorithm Based on ARP

6.1 Method based on ARP

All the devices Ethernet interface support ARP[4] (Address Resolution Protocol), and each of them with maintain a ARP table by themselves. This ARP table is used for convert IP address to physical Ethernet address. We can use ARP Table of router or exchanger to search devices linking Ethernet ports. Then tell the routers or exchangers from these devices with other information. Later on, search the ARP table again and again of new founded routers or exchangers. Finally, we can get network topology of the whole Ethernet.

6.2 Analysis of the method based on ARP

The IP addresses in the ARP are active IP, and the ARP table almost has no redundant data. So the topology searching methods base on ARP is very efficient. But it has limitation as follow:

(1) Request the network equipments to all support the agreement of ARP.

(2) If network scale is too large, ARP table may probably record all the existing network devices in the network. So this method is just suitable to search network topology of LAN.

7 Topology Discovery Algorithm Based on SNMP

The SNMP[3] (Simple Network Management) is a kind of protocol based on TCP/IP. We can use it as tool to get the network information from MIB[4]. Today's network devices do basically support SNMP Agent. So we can get network information from MIB by SNMP Agent, then we conclude network topology by the information.

There is a table called ipRouteTable in MIB. It records the router information of this device(ipRouteDest, ipRouteIfIndex，IpRouteNextHop，IpRouteType…). ipRouteDest is the index of a record. The data of ipRouteDest data-field present the possible destination address or range of destination address if regard this device as the root node and it may associate. The ipRouteIfIndex record interface port index related ipRouteDest. IpRouteNextHop record next hops address of the gateway device. IpRouteType record the relationship of destination subnet and this subnet. Because IpRouteNextHop (a datafield of IpRouteTable) represents network the node which has router function. So from the default getway, read IpRouteNextHop from IpRouteTable in devices' MIB, we can find out all the router decvices and the connection among them. We will also find ports by ipRouteIfIndex and tell the type of subnets gateways connecting by IpRouteType. Thus, we can get the backbone of network topology easily.

This method does totally base on SNMP. It is simple, efficient and low cost. But this method has limitation as follow: (1) The method can't discover the network devices who do not support SNMP or be not installed SNMP Agent. (2) There is much redundancy information in ipRouterTable. (3) A router device always has several IP address (one interface biding one IP address), and a router accustomed uses IP address as

it's ID. That means a router become to be several routers logically. As a result it is very difficult to reflect connections in topology. So this method is suitable to searching the topology of network backbone, but not suitable to reflect the whole network

8 Improved Topology Discovery Algorithm

By analyzing and comparing three network topology discovery algorithm, one based on ICMP, one based on ARP, one based on SNMP, we tell their length and limitation. If we combine them for using, they would performance much better. This paper takes level-division discovery idea. The idea break topology discovery down into two phases. In first phase what the algorithm does is to discover network backbone including route devices and subnets, we call this first-level topology discovery. In second phase what the algorithm does is to discover common devices such as personal computers and their relationship in subnet, we call this second-level topology discovery. In order to representing devices better, we ought to get the other information, for example, service information, to reflect network level better.

8.1 First-Level topology discovery

First-level discovery mainly takes the algorithm, which is based accessing route table of route device by SNMP protocol. This method was introduced above, start from the default gateway of subnet where network management workstation is in, poll every route device in whole network by，analyze information of route table synthetically, then we can get topology of the network bone. But this algorithm [5] has limitation that regards a route device, which has multi-ID as several route devices.

To resolve this limitation: Variable ipAdEntAddr is defined in ipAddrTable of MIB, it is used to identify IP address of interface on route device. Obviously, we are able to get every interface's IP addresseby polling ipAddrTable. According to RFC1519, the IP address range assigning for net of type C is 202.0.0.0—203.255.255.255. All the legal IP address in China starts form 202.0.0.0. In fact, illegal subnet IP address always chooses 192.168.xxx.xxx. So we think choosing the max IP address in ipAddTable to be the ID of route device is more reasonable, because the possibility of IP address's validity is higher.

Polling of First-level topology discovery follows the principle[6]: (1) If ipRouteType is not connecting directly (ipRouteType=4), that means the gateway whose ipRouteTalbe is accessed is the neighbor of ipRouteNextHop in its ipRouteTable. (2) If ipRouteType is connecting directly (ipRouteType=3), that means the gateway whose ipRouteTalbe is accessed is the neighbor of subnets in its ipRouteTable. (3)Discovering over time is designed to end discovering process.

8.1.1 Key data structure of first-level topology discovery

Key data structure to implement First-level topology discovery as follow[8-11]:

(1) ipaddress CurrentRouter;

//Current route device: the unique ID of current route device accessing, the max of IP addresses in its ipAddrTable.

(2) ipaddress CurrentGateway;

//Current route interface: route interface is accessing currently

(3) struct RouterQueueItem

{

//Route device queue：record the ID of all route devices discovered

ipaddress RouterAddr; //route address

*struct RouterQueueItem *next;*

//next pointer of queue

}

(4) struct VisitedGatewayQueueItem

{ipaddress RouterAddr; //gateway address

*struct RouterQueueItem *next;*

//next pointer of queue

} VisitedGatewayQueueItem，UnVisitedGatewayQueueItem;

//VisitedGatewayQueueItem: queue of gateway

accessed； UnVisitedGatewayQueueItem: queue of gateway unaccessed

(5) struct SubnetQueueItem

{//SubnetQueueItem：record the ID of all subnets discovered

ipaddress SubnetAddr; //subnet address

ipaddress SubnetMask; // subnet mask

*struct SubnetQueueItem *next;*

//next pointer of queue

} ;

(6) struct ConnectionQueueItem

{//queue of connection：record the connection between routes and routes, routes and subnets

ipaddress From; // address of one connection end

ipaddress To; // address of another connection end

*struct ConnectionQueueItem *next; //next pointer of queue*

} ;

8.1.2 Implementing of first-level topology discovery

(1) Initialize RouterQueueItem, VisitedGateway-QueueItem, SubnetQueueItem, ConnectionQueueItem, discovery time.

(2)Get the default gateway's address by reading local system file"/ctc/sysconfig/network"in management workstation.

(3)Insert default gateway's address from(2)into UnVisitedGatewayQueueItem;

(4)Get a node form UnVisitedGatewayQueueItem as current gateway, and kick off time starts;

(5)Access ipAddrTable of current gateway, get all ipAdEntAddr, and insert them into VisitedGateway-QueueItem. Insert the max ipAdEntAddr into Router-QueueItem, and assign it as current route.

(6)Is it overtime now? Yes, goto (7). No, goto(13);

(7)Bind ipRouteNextHop, ipRouteDest, ipRouteMask, ipRouteType，take GetNextRequest operation of SNMP, get this four numerical value in first row of current gateway's ipRouteTablen;

(8)Analyze ipRouteNextHop obtained by (7),if it is not in VisitedGatewayQueueItem or UnVisitedGateway-QueueItem, insert it into UnVisitedGatewayQueueItem;

(9)If ipRouteType=4, then access ipAddrTable of ipRouteNextHop, get the max ipAdEntAddr. If max ipAdEntAddr is not in RouterQueueItem, insert it into RouterQueueItem;

(10) If ipRouteType=3, and the subnet is not in SubnetQueueItem, then insert it into SubnetQueueItem. Evaluate ipRouteDest to SubnetAddr, evaluate ipRoute-Mask to SubnetMask .Construct a ConnectionQueueItem node, evaluate ipRouteDestIts to ConnectionQueueItem. To, evaluate current route to ConnectionQueueItem. From;

(11) Binding object value returned by (7), execute GetNextRequest operation again，get numerical value in next row，repeat executing (6)-(10) , do loop like this till ipRouteTable accessing is over. If current gateway is not in VisitedGatewayQueueItem, then insert it into VisitedGatewayQueueItem and delete it from UnVisited-GatewayQueueItem;

(12) If there is node still unaccessed in UnVisited-GatewayQueueItem, then get next node as current gateway, repeat (4)-(11);

(13) End.

8.2 Second-Level topology discovery

8.2.1 Discover the status of network devices in subnet

ICMP Ping method is generally used to detect current status of the host in the subnet.But like this paper mentioned before, it is a bit blind that "ping" every IP address available in a net segment by polling way, and it will increase network burden.

We integrate two topologies discovery algorithm to discover the status of network devices in subnet. One is based on ICMP. The other is based on ARP. First we access ipNetToMediaTable in gateway's MIB by ARP. The gateway connects the subnet we detect directly. And ipNetToMediaNetAddress can be used to confirm the active IP in the subnet. Then ping these IP address to discover the status of them. This way can reduce the using times of Ping operation so much that it improve the efficiency of algorithm, lighten the burden of

network. Furthermore, we can take ping operation in asynchronous way to reduce the time of discovery.

8.2.2 Discover the type of network devices in subnet

The type of the network devices can be comfirmed with ipForwarding and sysServices in MIB. ipForwarding can used to predict whether the device has the function that forward the IP message or not. If the device has the function, ipForwarding=1. Or it has no that function, ipForwarding=2. sysServices predicts the services network device offers. We can tell device working on which level of OSI architecture. For instants, ifsysServices=2, this means the device working on second level, it is probably a network bridge. So it is valuable for telling the service level of a network device.

8.2.3 Discover the type of network devices' interface in subnet

The information for network interface is recorded in MIB's ifTable. ifType, a field of ifTable descripts the type of interface as an integer. Table 4 lists the description as follow:

Table1 IfType, value and meaning

IfType Value	Type of Interface
4	Ddn-x.25
6	Ethernet-csma/cd
9	Iso88025-tokenRing
15	Fddi

By accessing the value of ifType, looking up the value on Table1 as dictionary, then we can tell the interface of network easily.

9 Summary

The improved algorithm takes level-division discovery idea. It integrates three popular algorithms, one is based SNMP, one is based ARP, and other is based on ICMP. It is not only can discover the topology of network backbone well, it also can tell the devices' status, type and interface type, in the subnet. Complexity and diversity of network, we ought to take multiple technologies and algorithms to achieve the network topology discovery goal if we want to discover different network s and detect different network devices.

References

[1] WAN G Jianguo, WU Jianp ing, CHEN Xiuhuan, FENG Xiaodong，Conformance Test ing on In ternet IP v4 Ba sed on Protocol In tegra ted Test System[J]. Journal of Software，2000，11(2)：207~212

[2] Carl Mitchell S., Quarterman J S., RFC 1027,1987,Using ARP to Implement Transparent Subnet Gateway[S]

[3] Case J.，RFC 1157, 1990, Simple Network Management Protocal (SNMP) [S]

[4] Rose M., RFC 1213, 1991, Management information base for network management of TCP/IP-based internets: MIB-II[S]

[5] Moy J.RFC 2328, 1998, OSPF Version 2[S]

[6] Xu Dahai, Huang Jianqiang, Wu Kexi and BaiYingcai，OSPF-Base Network Topology Search[J]. Computer Engineering & Science, 1999，21(6)：17~21

[7] Hedrick C.RFC1058, 1998, Routing Information Protocal[S]

[8] Govindan R, Tangm-unarunkit H S. Heuristics for Internet map discovery[C]. IEEE Infocom2000, IEEE, Mar 2000, 1371-1380

[9] R Caceres, N GDuffield. Multicast based inference of network internal characteristics[C]. IEEE Infocom'99 [C]. New York, USA, 1999, 1 :21-25

[10] Shi Zhou, Mondragon R J. The SR Ich-Club Phenomenon in the Internet Topology [J]. IEEE Communications Letters, 2004, 8(3) :180-182

[11] A-L Barabasi, R Albert. Emergence of Scaling in Random Networks [J]. Science, 1999, (10) :509-512

Qiu Jianlin**,** MS, associate professor. He was born in Nantong, Jiangsu, China in 1965. He graduated from Department of CS, Hohai University China and joined the faculty of School of CS, Nantong University in 1985. During 2003-2004, he was a visiting professor at Dept. of Mathematical and Computer Sciences, Colorado School of Mines (CSM), USA. He is a Senior Member of CCF and a member of Information Stored Special Committee of CCF. His research interests include logic synthesis

and optimization and computer-aided VLSI design and network security.

He Peng, MS, lecturer. He was born in Nantong, Jiangsu, China in 1980. He graduated from Department of CS, SuZhou University in 2005.His research interests include network technology and network security.

Gu Xiang, PHD, associate professor. He was born in Nantong, Jiangsu, China in 1973. He graduated from Department of CS, University of Science and Technology of China in 2004. His research interests include network security and network protocol.

Chen Jianping, MS, professor. He was born in Nantong, Jiangsu, China in 1960. He graduated from Department of CS, Nanjin University of science and technology of China in 1982 and joined the faculty of School of CS, Nantong University in 1985. His research interests include algorithm optimization design and network security.

Li Feng, a student of MS in Department of CS, Nantong University. She was born in Xuzhou, Jiangsu, China in 1984. Her research interests include logic synthesis and optimization and computer-aided VLSI design.

An Adaptive Optimistic Total Order Broadcast Algorithm in WAN

Yizheng Chen Jihong Zhu

Department of Computer Science and Technology, Tsinghua University, Beijing, China
Email: chenyz05@mails.tsinghua.edu.cn, jhzhu@tsinghua.edu.cn

Abstract

Total order broadcast is a useful group communication primitive in the construction of many fault-tolerant distributed applications. The high latency of total ordering can be masked by using an optimistic algorithm. A new algorithm has been proposed to enable the usage of optimistic delivery also in WANs.

In this paper, we address the deficiencies in previous optimistic algorithm and propose an adaptive optimistic algorithm in WAN which exploits different behaviors exhibited in different conditions. The simulation results show that the new algorithm outperforms the original algorithm in the numbers of correct optimistic deliveries in the initial stabilization and the transition period when route changes, while its performance does not deteriorate when transmission delays fluctuate. And it is more robust than the original algorithm in resilience to topology variation and transmission delay variability. It turns out that the tradeoff for the improved efficiency is the reduction of the optimistic window.

Keywords: Total Order, Atomic, Broadcast, Adaptive, Optimistic

1 Introduction

Total order broadcast, also known as atomic broadcast, serves as a basis for the construction of many fault-tolerant distributed applications. It is often implemented as a group communication primitive which allows processes to deliver the same set of messages in the same order. It is particularly useful in the fault-tolerant services using replicated state machine approach. By applying this primitive, it is ensured that the state of each replica is kept consistent. A systematical survey and taxonomy of various total order broadcast algorithms can be found in [1].

However, the implementation of such a primitive can be costly for its additional latency. To alleviate this problem, an optimistic total order algorithm has been proposed [2] based on the observation that the spontaneous order of messages received by each process is often the same in local area networks (LANs). Thus, by optimistically delivering the messages on the spontaneous order, the application can perform the computation in parallel with the ordering algorithm. Later, when the total order is established (e.g. receives the sequence number in the sequencer-based algorithms), the process finally delivers the messages. The result of optimistic computation is committed if the order in previous optimistic deliveries is confirmed. If not, the application must roll back and redo the computation with the correct order.

The technique is efficient when a large number of correct optimistic deliveries offset the cost of redoing the computation. Unfortunately, this is not the case in wide area networks (WANs) where the effectiveness of the technique deteriorates by the presence of large transmission delays. To solve the problem, Sousa et al. [3] proposed an algorithm which introduces delay compensation in the message reception to mask the different transmission delays between nodes. It is a variant of the sequencer-based algorithm and in this way each process would mimic the reception order of messages at the sequencer, reducing the mistakes in optimistic deliveries.

In this paper we discuss how to overcome the deficiencies of the algorithm presented in [3]. Then we

propose a novel adaptive approach to reduce the penalty incurred by wrong optimistic deliveries during the initial stabilization and the transition period when route changes, and not to degrade the performance of the algorithm when transmission delays fluctuate. The resulting algorithm is evaluated and compared with the original algorithm.

The rest of the paper is structured as follows. Section 2 introduces the system model and formally defines the total order broadcast. The original optimistic total order algorithm in WAN proposed in [3], as well as its deficiencies and limitations, is described in Section 3. In section 4, an adaptive optimistic total order algorithm is presented. Section 5 evaluates the performance of our algorithm. Finally, we conclude in section 6.

2 System Model and Definitions

We consider a distributed system with a finite set of processes $\Pi = \{p_1, p_2, ..., p_n\}$. We assume that processes communicate by message exchange over reliable (i.e., there are no message creation, alteration, duplication or loss) and FIFO channels. As our goal is to improve the performance of the algorithm in good runs (i.e., processes behave correctly, which is always the case in practice), the issue of failure detection is out of scope of the paper. Mechanisms for fault-tolerance are comprehensively discussed in [1].

2.1 Reliable broadcast

Reliable Broadcast is defined by primitives *R_broadcast* and *R_deliver*, and satisfies the following properties [4]:

- **Validity**: If a correct process *p R_broadcast*s a message *m*, then some correct process eventually *R_deliver*s *m* or no process is correct.
- **Agreement**: If a correct process *p R_deliver*s a message *m*, then all correct processes *R_deliver m* eventually.
- **Integrity**: For any message *m*, every process *R_deliver*s *m* at most once, and only if *m* was previously *R_broadcast* by some process *p*.

2.2 Total order broadcast

Total Order Broadcast is defined by primitives *TO_broadcast* and *TO_deliver*. It satisfies the Validity, Agreement and Integrity properties stated above with additional Total Order property:

- **Total Order**: If two correct processes *p* and *q* both *TO_deliver* messages *m* and *m'*, then *p TO_deliver*s *m* before *m'* if and only if *q TO_deliver*s *m* before *m'*.

3 The Optimistic Total Order Algorithm in WAN

3.1 Overview

In this section, we provide an overview of the original *Optimistic Total Order* (OTO) algorithm in WAN (Figure 2) presented in [3].

This algorithm is an optimistic variant of the traditional sequencer-based total order algorithm. The total order is determined by the spontaneous reception order of messages at the sequencer. When the application invokes procedure *TO_broadcast* (line 10), it is actually done by invoking the underlying primitive *R_broadcast*. When a process *p* receives a message *m* from the network (line 12), it queues *m* for a period of time which was calculated previously to approximate the spontaneous reception order at the sequencer. When that time expires (line 14), *m* is optimistically delivered (line 15) using the primitive *opt_delivery*. If *p* happens to be the sequencer, it computes a sequence number for *m* (line 18) and broadcasts to all processes (line 19). Upon receiving the sequence number for *m* (line 22), the process finally delivers *m* to the application (line 25) using the primitive *fnl_delivery* if *m* is the next message to be definitively delivered. The way processes estimate the delays to be imposed on the received messages will be discussed below.

First, we will describe the estimation of delays on all processes other than the sequencer. When a message and its sequence number are received, both times are recorded

(line 13 and line 23 respectively). After finally delivering the message, both recorded times are used to compare the optimistic and definitive order of two consecutive finally delivered messages (line 27). Then the adjustment to the artificial delays is computed (line 28-31).

This is illustrated in Figure 1 Message m_1 is broadcast by process p_1 and message m_2 is broadcast by another process p_2. The algorithm assumes that transmission delays of seq(m_1) and seq(m_2) from the sequencer to p are the same, so p can use the value of t_{sp} to locally determine t_s. Then the adjustment computed at line 27 would be $\Delta = t_{sp} - t_p$. Note that t_p is negative in Figure 1, indicating the optimistic delivery order of m_1 and m_2 is reversed.

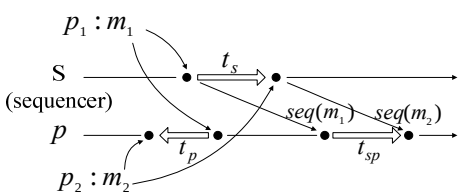


Figure 1 Estimation of delays

The sequencer calculates the delay for its own messages in a different way. Each process suggests a delay based on its local array *delay[]* (line 11). Then the sequencer chooses the maximum value (line 21).

3.2 Delay compensation rules

We formulate the rules to adjust *delay[]* as follows.

Let ω_i^j denote the transmission delay between node *i* and node *j*. The goal of the adjustment is to make *delay[]* at process *k* satisfy the following equation:

$$(\omega_i^k + delay_k[i]) - (\omega_j^k + delay_k[j]) = \omega_i^s - \omega_j^s \quad \forall i, j, k \in 1,2,...,N \text{ and } i, j, k \neq s \tag{1}$$

Eq.(1) ensures each process to mimic the spontaneous reception order at the sequencer.

In order to introduce as few artificial delays as possible and get larger optimistic window, we also want *delay[]* to satisfy:

$$\sum_{i=1}^{N} delay_k[i] \to 0 \quad \forall k \in 1,2,...,N \tag{2}$$

```
1: Initialization :
2:   g ← 0  {Global sequence number}
3:   l ← 0  {Local sequence number}
4:   R ← φ  {Messages received}
5:   S ← φ  {Sequence numbers}
6:   O ← φ  {Messages opt_delivered}
7:   F ← φ  {Messages fnl_delivered}
8:   delay[1...n] ← 0
9:   r_delay[1...n] ← 0  {Delays requested to the sequencer}

10: procedure TO_broadcast(m)
11:     R_broadcast(DATA(m, max(delay[]) − delay[seq]))

12: upon R_deliver(DATA(m,d)) do
13:     R ← R ∪ {(m, d, now + delay[m.sender])}

14: upon ∃(m,d,t) ∈ R : now ≥ t ∧ m ∉ O ∧ m ∉ F do
15:     opt_delivery(m)
16:     O ← O ∪ {m}
17:     if p = seq then
18:         g ← g + 1
19:         R_broadcast(SEQ(m, g))
20:         r_delay[m.sender] ← d
21:         delay[p] ← max(r_delay[])

22: upon R_deliver(SEQ(m,s)) do
23:     S ← S ∪ {(m, s, now)}

24: upon ∃(m,d,o) ∈ R : (m, l+1, t) ∈ S ∧ m ∉ F do
25:     fnl_delivery(m)
26:     if ∃(m',d',o') ∈ R : (m', l, t') ∈ S then
27:         Δ ← (t − t') − (o − o')
28:         if Δ > 0 then
29:             adjust(m'.sender, m.sender, Δ)
30:         else
31:             adjust(m.sender, m'.sender, |Δ|)
32:     l ← l + 1
33:     F ← F ∪ {m}

34: procedure adjust(i, j, d)
35:     v ← (delay[i] × α) + (delay[i] − d) × (1 − α)
36:     if v ≥ 0 then
37:         delay[i] ← v
38:     else
39:         delay[i] ← 0
40:         delay[j] ← delay[j] + |v|
```

Figure 2 Original OTO algorithm

However, note that Eq.(1) should be satisfied with higher priority.

3.3 Deficiencies and limitations

Now we discuss the deficiencies and limitations in the original algorithm.

First, in order to satisfy Eq.(2), procedure *adjust(i, j, d)* always tries to decrease the value for one process in *delay[]* first. This might make it difficult for *delay[]* to satisfy Eq.(1) eventually. Consider a simple scenario in which $N = 3$. Assume that the initial value of *delay[]* is {0, 0, 0} at process p_3 and it should be {0, 5, 3} after the adjustment. If the sender of three consecutive *fnl_delivery* messages m_1, m_2, m_3 is p_1, p_2, p_3 respectively, *delay[]* would be adjusted to {0, 5, 0} after *fnl_delivery*(m_2). Later, after *fnl_delivery*(m_3), process

p_3 learns that it should impose 2 more time units to messages from p_2 than from itself (recall we assume that *delay[]* should be {0, 5, 3}). Then it will decrease *delay[2]* from 5 to 2 instead of increasing *delay[3]* from 0 to 3, making *delay[]* to be {0, 2, 0}. Thus it might be hard to adjust the relative distance between p_2 and p_3 while maintain that between p_1 and p_2.

Second, the way using the *opt_delivery* time recorded at the reception of a message (line 13) as the reference to compute the adjustment (line 27) is somewhat unreasonable. Still consider the above scenario. When process p_3 computes the adjustment of the relative distance between p_2 and p_2 at line 27 after *fnl_delivery*(m_3), *delay[2]* used to compute o' in tuple (m_2, *d*, o') which is recorded at the reception of m_2 (line 13) is 0. However, *delay[2]* has been adjusted to 5 after *fnl_delivery*(m_2). Consequently, this might lead to a wrong adjustment computed at line 27 after *fnl_delivery*(m_3).

Third, an inertia pondering factor α is used to cope with spurious variations on transmission delays (line 35), but α is determined beforehand. In experiments in [3], α is set to be 0.95. This is high enough to resist the fluctuation of transmission delays and not to degrade the estimate. But a lower α is favorable for quick initial stabilization and fast transition when route changes. The extensive measurements in [5] show that routing changes occur over a wide range of time scales from seconds to days because of different reasons (e.g. a link degrades or fails). This leads to our improvement to the original algorithm.

4 The Adaptive Optimistic Total Order Algorithm in WAN

We now discuss how to overcome the deficiencies and limitations mentioned in the previous section.

4.1 Basic ideas

For the first deficiency mentioned in section 3.3, we always choose to increase the value in *delay[]*.

To amend the second deficiency, *delay[]* at the time when the adjustment is calculated is used instead of that at the reception of a message.

In order to work with the desirable inertia factor α in different conditions, we augment the system with an adaptive module. By exploiting different behaviors exhibited in the transition period and when transmission delays fluctuate, the adaptive module could make corresponding adjustment to α.

4.2 Algorithm

The *Adaptive Optimistic Total Order* (AOTO) algorithm is specified in Figure 3. An additional variable *matrix[][]* is used to record the information of the adjustment between each pair of processes.

The algorithm works as follows. The initial value of the inertial factor α is zero (line 10). When process *p* wishes to *TO_broadcast* a message *m*, it *R_broadcast*s *m* (line 13). When process *p R_deliver*s *m*, the reception time *r* and the optimistic delivery time *o* are recorded (line 15). The *upon* statements related to *opt_delivery* and *R_delivery* of the sequence number is the same as those in OTO algorithm. After *fnl_delivery(m)*, process *p* invokes procedure *adapt* to adapt α to the desirable value in different conditions (line 28). Then procedure *adapt*, in turn, invokes procedure *adjust* to make the adjustment.

In procedure *adapt*, as mentioned above, the "fresh" *delay[]* is used to compute the adjustment (line 35). For the sequencer, the calculated adjustment is no longer guaranteed to be zero in this way. Since the sequencer computes the delay for its own messages at line 22-23, it simply returns at line 33. Lines from 36 to 44 are the core of the adaptive mechanism. If the distance between a pair of processes should be adjusted larger, the value of *matix[][]* increases by one(line 38). Otherwise it decreases by one (line 40). If it changes monotonously, i.e., it increases or decreases consecutively for β times, it might indicate a normal transition (e.g. when route changes) in the network. So the inertial factor α will be decreased by Δ_α for fast transition (line 42). Otherwise the computed adjustment

is likely due to the fluctuation of transmission delays. Therefore α should be increased to resist the spurious variations (line 44). The maximum and minimum values of α are not represented in the algorithm to preserve clarity. Then the adjustment is made by invoking procedure *adjust* with adapted α (line 45-48).

```
1: Initialization :
2:   g ← 0  {Global sequence number}
3:   l ← 0  {Local sequence number}
4:   R ← φ  {Messages received}
5:   S ← φ  {Sequence numbers}
6:   O ← φ  {Messages opt_delivered}
7:   F ← φ  {Messages fnl_delivered}
8:   delay[1...n] ← 0
9:   r_delay[1...n] ← 0  {Delays requested to the sequencer}
10:  α ← 0  {inertial pondering factor}
11:  matrix[1...n][1...n] ← 0

12: procedure TO_broadcast(m)
13:     R_broadcast(DATA(m,max(delay[]) − delay[seq]))

14: upon R_deliver(DATA(m,d)) do
15:     R ← R ∪ {(m,d,now,now + delay[m.sender])}

16: upon ∃(m,d,r,o) ∈ R : now ≥ o ∧ m ∉ O ∧ m ∉ F do
17:     opt_delivery(m)
18:     O ← O ∪ {m}
19:     if p = seq then
20:         g ← g + 1
21:         R_broadcast(SEQ(m,g))
22:         r_delay[m.sender] ← d
23:         delay[p] ← max(r_delay[])

24: upon R_deliver(SEQ(m,s)) do
25:     S ← S ∪ {(m,s,now)}

26: upon ∃(m,d,r,o) ∈ R : (m,l + 1,t) ∈ S ∧ m ∉ F do
27:     fnl_delivery(m)
28:     adapt(m)
29:     l ← l + 1
30:     F ← F ∪ {m}

31: procedure adapt(m)
32:     if p = seq then
33:         return
34:     if ∃(m',d',r',o') ∈ R : (m',l,t') ∈ S then
35:         Δ ← (t − t') − ((r + delay[m.sender]) − (r' + delay[m'.sender]))
36:         Let count denote matrix[m.sender][m'.sender]
37:         if Δ > 0 then
38:             count ← count + 1
39:         else if Δ < 0 then
40:                 count ← count − 1
41:         if count increases or decreases consecutively for β times then
42:             α ← α − Δ_α
43:         else
44:             α ← α + Δ_α
45:         if Δ > 0 then
46:             adjust(m.sender, Δ, α)
47:         else
48:             adjust(m'.sender, |Δ|, α)

49: procedure adjust(j, d, α)
50:     delay[j] ← (delay[j] × α) + (delay[j] + d) × (1 − α)
51:     delay_min ← min(delay[ ])
52:     if delay_min > 0 then
53:         for k ← 1 to n
54:             do delay[k] ← delay[k] − delay_min
```

Figure 3 Adaptive OTO (AOTO) algorithm

In procedure *adjust*, the process chooses to increase the value for one in *delay[]* all the time (line 50). This might result in a continuous increase of each value in *delay[]*. As noted in Eq.(1), if the same amount is subtracted from *delay[]* for any two processes simultaneously, Eq.(1) still satisfies. Thus, the process will check the minimum value in *delay[]*. If this value is positive (line 52), it should be taken away from *delay[]* for all processes (line 53-54) to try best to satisfy Eq.(2).

5 Evaluation

In this section we evaluate the performance of the AOTO algorithm with the original OTO algorithm.

5.1 Simulation model and settings

We conducted simulations using the discrete event simulator NS-2 [6]. With a given number of nodes and a probability with which there is an edge between two nodes, GT-ITM [7] is used to generate two random flat graphs as the topology in the test. Each topology consists of 100 nodes and 10 nodes are randomly chosen in the network to *TO_broadcast* messages. The sequencer in each topology is also randomly selected.

The transmission delay in each link is normally distributed with mean generated by GT-ITM and our given standard deviation σ. As observed in [5], the standard deviation of transmission delays is mostly less than 10%. Also, experimental results in [3, 8] show that an optimistic algorithm is useful in networks with less variability and not very high message rates. When this is not the case, the optimistic deliveries are highly inaccurate since the standard deviation is comparable to the message interarrival interval. In this situation, other protocols proposed in [9, 10] are more suitable. Therefore, we performed our test under two conditions: σ =0% and σ =5% and all nodes exhibits a transmission rate of 10 messages per second.

Δ_α is set to be 0.25 and $\alpha_{\max}$ and $\alpha_{\min}$ are 0.95 and 0.05 respectively in the AOTO algorithm. If the adapted α is out of the range, it is set to $\alpha_{\max}$ or $\alpha_{\min}$.

We consider β to be 3 in order to differentiate normal transitions from spurious variations in the network.

5.2 Simulation results

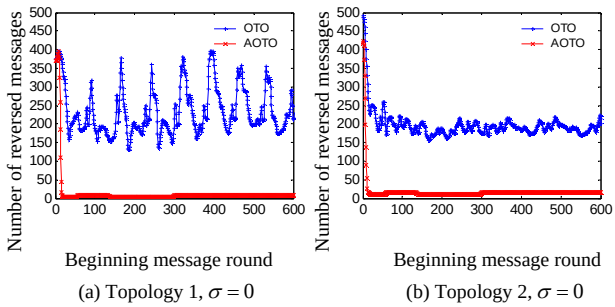
Two scenarios are used to compare the performance between the OTO and AOTO algorithm. Each simulation is run for 120s. Scenario 1 includes the initial stabilization. In Scenario 2, we introduce a route change in the network after stabilization.

First we define the number of reversed messages: if the definitive order of two messages m_1 and m_2 is reversed in the optimistic order, the number of reversed messages increases by one. For example, if the definitive order is $\{m_1, m_2, m_3\}$ while the optimistic order is $\{m_3, m_1, m_2\}$, the number of reversed messages is 2 since the relative order between m_1 and m_3, as well as m_2 and m_3, is reversed.

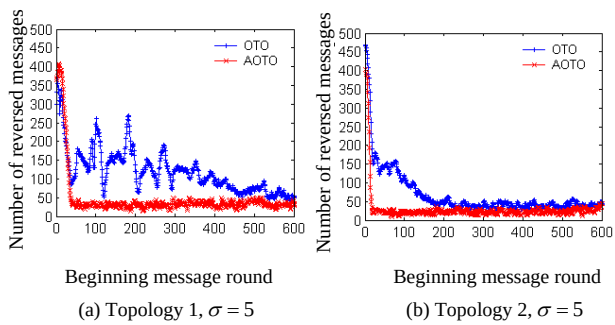
Assume that every message also has a local sequence number given by the sender to count how many messages have been sent by the sender. Then we define the messages in the same round are those with the same local sequence number.

It is observed that messages received in adjacent rounds might interweave during a period. Thus, the reversed number of messages in consecutive *m* rounds beginning at round *k* could be used to gain a perspective of the correctness of optimistic deliveries during that period. The larger the *m*, the more accurately the metric characterizes the correctness and the larger time scale it represents.

The results depicted in Figure 4 and Figure 5 are measured by one node in the network with *m* being 5. This is high enough not to degrade the accuracy of the metric, and to get a view of the algorithm performance during the initialization and transition period in our experiments. Other values for *m* have similar results. In scenario 2, route changes at round 700. As expected, the AOTO converges more quickly in the initial stabilization (Figure 4) and needs less recovery time in the transition period (Figure 5) than the OTO, typically in less than 50 message rounds. In addition, the AOTO exhibits much higher accuracy in optimistic deliveries. The results also show that the accuracy achieved by the AOTO suffers less variation with the topology and transmission delay variability. In this sense, the AOTO is more robust in regard to the original algorithm.

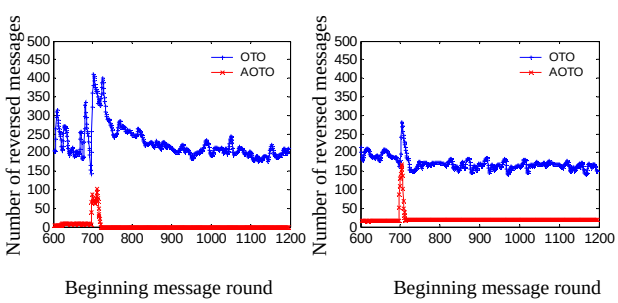


(a) Topology 1, $\sigma = 0$ (b) Topology 2, $\sigma = 0$

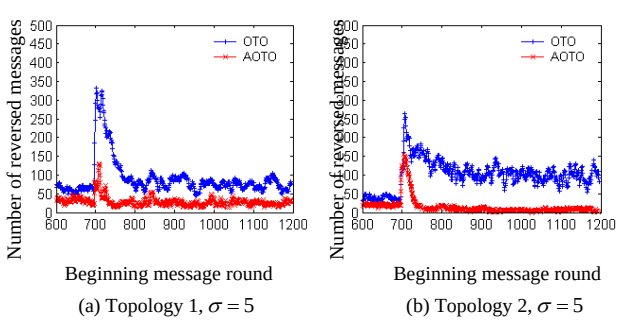


(a) Topology 1, $\sigma = 5$ (b) Topology 2, $\sigma = 5$

Figure 4 Reversed messages number (Scenario 1)



(a) Topology 1, $\sigma = 0$ (b) Topology 2, $\sigma = 0$



(a) Topology 1, $\sigma = 5$ (b) Topology 2, $\sigma = 5$

Figure 5 Reversed messages number (Scenario 2)

The second evaluation criterion is the average latency of the optimistic and definitive delivery. And the time between them for the application to perform some optimistic computation, called the optimistic window [3], is also compared.

The results in Scenario 1 are presented in Table 1. It shows that the optimistic window is reduced dramatically in the AOTO, i.e., approximately 1/4 for Topology 1 and 1/2 for Topology 2 of that in the OTO. However, it is still meaningful for the application to do some optimistic processing. This reduction might be the expense for the improved optimistic order.

Table 1 Latencies and optimistic window sizes (ms)

		Topology1		Topology 2	
		OTO	AOTO	OTO	AOTO
$\sigma = 0$	fnl. latency	1460	1378	401	376
	opt. latency	629	1188	246	315
	opt. window	831	190	155	61
$\sigma = 5$	fnl. latency	1459	1385	400	377
	opt. latency	673	1185	279	313
	opt. window	786	200	121	64

6 Conclusion

The optimistic total order algorithm is useful to mask the high latency of total ordering. A new algorithm [3] has been proposed to offer more correct optimistic deliveries in WANs by using delay compensation.

This paper overcomes the deficiencies in previous optimistic algorithm and presents an adaptive optimistic total order algorithm which is efficient in the initial stabilization and the transition period when route changes. It can also guarantee a high performance when transmission delays fluctuate. Simulation experiments show that the AOTO allows for quick stabilization and fast transition and is more robust than the original OTO. It appears that the tradeoff for the improved efficiency is the reduction of the optimistic window.

References

[1] X. Defago, A. Xchiper, and P.Urban, “Total order broadcast and multicast protocols: Taxonomy and survey”, *ACM Computing Surveys*, Dec. 2004, 6(4):372-421

[2] F. Pedone and A. Schiper, “Optimistic atomic broadcast: a pragmatic viewpoint”, *Theoretical Computer Science*, Jan. 2003, 291(1):79-101

[3] A. Sousa, J. Pereira, F. Moura, and R.Oliveira, “Optimistic total order in wide area networks”, In *Proc. of the 21st IEEE Symp. on Reliable Distributed Systems*, Oct. 2002, pp. 190-199

[4] A. Schiper, “Group Communication: From Practice to Theory”, *SOFSEM 2006: Theory and Practice of Computer Science*, 2006, volume 3831:117-136

[5] V. Paxson, “Measurements and Analysis of End-to-End Internet Dynamics”, *PhD thesis*, Univ. of CA, Berkeley, Apr. 1997

[6] NS2: http://www.isi.edu/nsnam/ns/

[7] GT-ITM: http://www.cc.gatech.edu/projects/gtitm/

[8] L. Rodrigues, J. Mocito, and N. Carvalho, “From spontaneous total order to uniform total order: different degrees of optimistic delivery”, In *Proc. of the 2006 ACM Symp. on Applied Computing*, Apr. 2006, volume 1: 723-727

[9] P. Vicente and L. Rodrigues, “An indulgent uniform total order algorithm with optimistic delivery”, In *Proc. of the 21st IEEE Symp. on Reliable Distributed Systems*, Oct. 2002, pp. 92-101

[10] J. Mocito and L. Rodrigues, “Run-time switching between total order algorithms”, In *Proc. of the Euro-Par 2006*, Aug. 2006, pp. 582-591

The Research of Using Honfyd to Beguile and Detect Network Worm

Senlin Jiang Jibin Wang Tingting Yu

Industrial Training Centre, Wuxi Institute of Technology, Wuxi, 214121, China

Abstract

This paper introduced a new technique of examinning network worm—using honeypot to beguile and detect network worm.The article analyzed the theory and merit of it ,Then analyzed the merit of using HONEYD to do it .Through analyzing the structure of HONEYD，the paper introduced the key technique of HONEYD simulating honeypot— fingerprinting matching and virtual routing topologies.Finally， installed virtual honeypot to beguile and detect “Worm. Blaster”.

Keyword: honeypot, fingerprint , network worm

1 Introduction

The network worm is one kind of automatic program which can spread in the network. Unlike the virus, it can be executed independently. For network worm has a property of “pushing by itself”, a computer which is inffected by network worm can choose hundreds of host to spread, and those computers will inffect others in this way. Therefore the network worm usually directly harm network's normal operation, and causes the network serious jam even to paralyze. It is a calamity that a network lack immunity of network worm[1].

At present, the method of examination network worm are many, this article will introduce one kind of research —— use honeypot to beguile and detect network worm to discover networm in early time.them this page will research using HONEYD to beguile and detect network worm.

2 The Theory of Using Honeypot to Beguile and Detect Network Worm

A honeypot is a resource whose value is being in attacked or compromised. This means, that a honeypot is expected to get probed, attacked and potentially exploited. Honeypots do not fix anything.They provide us with additional, aluable information[2].

So honeypot has such property:

1) The honeypot does not provide any network service, therefore any connection to the honeypot may regard as the illegal connection;

2) A specific disposition tempts measures network worm's honeypot, once attacked, it may determine worm's behavior basically, depends upon the honeypot internal examination mechanism, may carry on the warning to the entire network;

3) The honeypot collects information passively, compared with other examination network worm's method, the honeypot examination worm will not create the extra burden to the network, and it will collect the information quite simplify, the honeypot can effect examination worm more highly[5].

3 Using HONEYD to Beguile and Detect Network Worm

HONEYD is a Low-Involvement Honeypot framework. It can simulates virtual computer systems at the network level. The simulated computer systems appear to run on unallocated network addresses[3].The advantage of using

HONEYD to beguile and detect network worm is That the resource consumption is small[4].

3.1 Logical structure of HONEYD

HONEYD is made up of such parts: Configuration Personality, Packet Dispatcher, Personality Engine, routing, services, protocol handler (Figure 1 Logical structure of HONEYD).

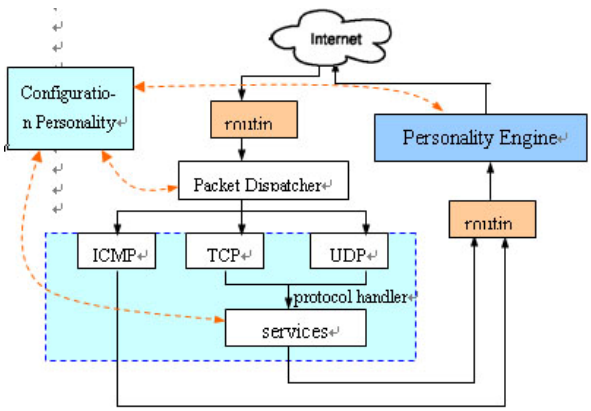


Figure 1 Logical structure of HONEYD

3.2 Analysis of the important technology

All of the honeypot can do many different service based on the Disposition.But attacker has many ways and tools to recognise honeypot. HONEYD could well use the techonology of Fingerprint match and the Virtual network.So face the arrackers probe ,the honeypot Simulated by HONEYD could still do effective camouflage .

1) Stack fingerprint match

HONEYD takes TCP and the UDP behavior reference and use the NMAP fingerprint database; Takes the ICMP behavior with the XPROBE fingerprint database the reference.

We will use the fingerprint information which provides with NMAP which changes the honeypot network stack's characteristic to show:

Fingerprint IRIX 6.5.15m on SGI 02

Tseq(Class=TD%gcd=<104%SI=<AE%IPID=I%TS=2HZ) (test the goal how to initialize ISN)

T1(DF=N%W=EF2A%ACK=S++%Flags=AS%Ops=MNWNNTNNM)

(T1-T7:testing responce of Data packet arriving open/close port)

T2(Resp=Y%DF=N%W=O%ACK=S%Flags=AR%Ops=)

T3(Resp=Y%DF=N%W=EF2A%ACK=O%Flags=A%Ops=NNT)

T4(DF=N%W=O%ACK=O%FlagsR%Ops=)

T5(DF=N%W=O%ACK=S++%Flags=AR%Ops=)

T6(DF=N%W=O%ACK=O%Flags=R%Ops=)

T7(DFN%W=O%ACK=S%Flags=R%Ops=)

PU(Resp=n) (Closing port How to produce ICMP responding package)

2) Virtual network topology

HONEYD can simulate the virtual route topology . Usually, the virtual route topology is a tree which enters by the data packet establishes. What tree's each internal node expression is a router, nearby each contains characteristics and so on data packet detention and loss. The terminal node carries on the response to the network. The HONEYD frame supports the multi-export parallel existence, needs to choose according to the cyberspace cyberspace exports the router

In order to simulate the symmetrical net topology , when a data packet enters and leaves the HONEYD frame time, we must refer to the routing list.

When HONEYD receives to a data packet, it found the correct route tree entrance, and carried on the traversal, started from the root, to know that the terminal which finds the data to wrap the destination address. The data wrapped in each side on the detention and the loss accumulation result had decided whether it could discard in the process which the data packet and the data wrapped in are on everybody's lips to retard how long.

When traversal each router time, the HONEYD frame must consume TTL similarly. If TTL achieves the 0, HONEYD frame to return to an overtime information the ICMP data packet to the source address.

Regarding the virtual network, it may conformity physical system to the virtual network., When HONEYD receives to one for real system's data packet, it traversal the entire topology until to find a router to be able to pay the network which this data packet the real

main engine. In order to find system's hardware address, if necessary,it must transmit ARP to request, then data packet seal, in the ethernet detected. Similarly, when a real system through HONEYD frame corresponding hypothesized route transmission for the honeypot ARP request, HONEYD must respond it(Figure 2 virtual network topology)

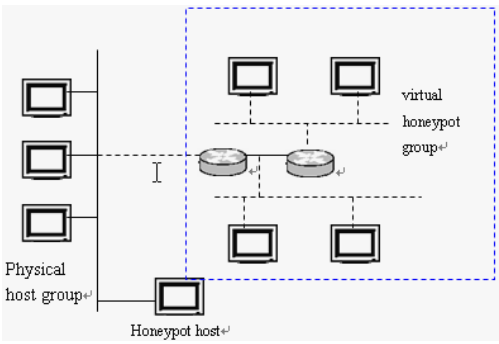


Figure 2 virtual network topology

4 Configuring Virtual Honeypot to Beguile and Detect Network Worm

In order to beguile network worm, it usually use the honeypot to simulate the specific system and the service.using HONEYD beguile and detect network worm need to configure a proper virtual honeypot. The following example is using HONEYD to beguile and detect the network worm “Worm.Blaster”.

According to network worm of “Worm.Blaster” characteristic establishment corresponding characteristic pattern plate, its characteristic includes: The operating system is WINDOWS2000 or WINDOWS XP, here establishes is WINDOWS XP, opens 135 ports and moves RPC the DCOM service.

create windows

create default set default personality "Windows XP Pro"

add default tcp port 135 open

add default tcp port 4444 "/bin/sh scripts/ WormCatcher.sh $ipsrc $ipdst"

set default tcp action block

set default udp action block

When the ARP table has deposited use network IP, it may use these IP address binding pattern plate, thus it needs honeypot, like subnet scope 202.195.151. 224—202.195.151.239 # altogether has 14.

bind 202.195.151.225 windows
bind 202.195.151.226 windows
bind 202.195.151.227 linux
bind 202.195.151.228 windows
bind 202.195.151.229 windows
bind 202.195.151.230 windows
bind 202.195.151.231 linux
bind 202.195.151.232 windows
bind 202.195.151.233 linux
bind 202.195.151.234 windows
bind 202.195.151.235 windows
bind 202.195.151.236 linux
bind 202.195.151.237 windows
bind 202.195.151.238 windows

5 Conclusion

The network worm's harm is serious day by day. Using honeypot to lure and measure the network worms, so as to achieve early warning, is a new approach which can effectively reduce the damage caused by them.

Reference

[1] CNCERT/ CC 2006 report of network security [R/ OL] . 20072022 15. http :/ / www. cert . org. cn/ articles/ docs/ common/ 2007021523214. shtml The development of HONEYD. http://www.HONEYD.org

[2] Lance Spitzner. Honeypots:Tacking Hackers[M]. Addison-Wesley, 2003

[3] The Honeynet Project:KnowYour EnemyI[M]. Addison-Wesley, 2002

[4] The Honeynet Project:KnowYour EnemyII[M]. Addison-Wesley, 2002

[5] Klug D. Honeypots and Intrusion Dection[EB/OL]. http://www.sans.org/rr/intrusion/honeypots.php. 2002-11

The Research of Large-Scale Video Server Cluster *

Qingping Guo[1] Guangyou Zhou[2]

School of Computer Science and Technology, Wuhan University of Technology, Wuhan (Hubei) 430007, China

E-mail:1 qpguo@whut.edu.cn; 2 zhouguangyou007@yahoo.com

Abstract

The Internet develops at a very fast speed and network hosts increase expansibly, which bring an enormous challenge for network bandwidth and servers. As we know from network technology, the increasing speed of network bandwidth is faster than the speed of processor and memory accessing rates. So performance bottleneck appears at server. Owing to the popularization of Broad Band communication, a large number of network applications have been appeared gradually, such as video on demand and video conference, which also need higher capability of server. At the same time, a great many network services couldn't bear the deadweight and couldn't deal with consumer's requests in time because accessing rate rise explodes and it leads to a poor service quality as a long-time wait. By all appearances, any host computer couldn't provide such huge services separately.

How to set up a telescopic network services for load demand is a very important issue. Linux Virtual Server Cluster (LVSC) is a load balance cluster technology which is realized in linux kernel and based on TCP/IP protocol. It has a good scalability and high capability load balance cluster technology, and it especially has characteristics such as strong function, good adaptability, and clarity to users

Keywords: LVSC; High Availability; video on demand; scalability

1 Relative Concept

1.1 Architecture of Linux Virtual Server Cluster

Linux virtual server cluster is a flex network service system structure and has three layers which is composed of load balancer, server pool and shared storage. Linux virtual server cluster takes hiberarchy and applicationless into account, so it is a patulous and high capability load balancing technology. [1]

Let's use Pfister's statement that all clusters should act like 'a single unified computing resource' to describe the architecture of an enterprise cluster. One example of a unified computing resource is a single computer, as shown in Fig1.[10]

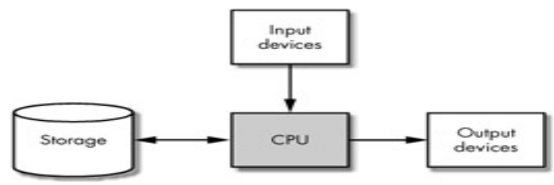


Figure 1 Simplified Architecture of a Single Computer

If we replace the CPU in Figure 1 with 'a collection of interconnected whole computers,' the diagram could be redrawn as shown in Figure 2.

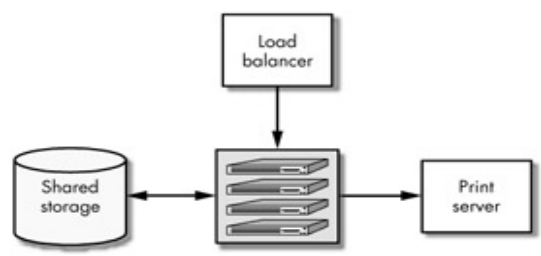


Figure 2 Simplified Architecture of an Enterprise Cluster

The load balancer replaces the input devices. The shared storage device replaces a disk drive, and the print server is just one example of an output device.

The load balancer sits between users of the cluster and the 'whole computers', which are nodes that make

* Supported by a grant from the National Natural Science Foundation of china (No.60403043)

up the cluster. The load balancer decides how best to distribute the incoming workload across all of the nodes.[2]

The shared storage device acts as the single repository for the enterprise cluster's data, just as a disk drive does inside a single computer.

The print server of output device in this example is just one of many possible cluster output devices—a shared print server

1.2 The Working Principle of Linux Virtual Server Cluster

Linux virtual server cluster is composed by a set of servers with interlinkages via high-speed LAN or geography distributing WAN. [3]

At the front of LVS, it has a Load-Balancer which could schedule the network requests to one server of the cluster, and it is transparent to client. The client seems like accessing only one high capability, high availability server when the client accesses cluster system. [3]

The program of client shouldn't be modified, and the retractility of the system is achievable by pellucidly adding or deleting a node simply. The usability of the system is achieved by checking node status or service process.

1.3 The Applications of Linux Virtual Server Cluster Technology [4]

Four applications based on Linux virtual server cluster technology are as follows:

（1）the telescopic WEB services

（2）the telescopic Cache services

（3）the telescopic mailing services

（4）the telescopic stream media.

The load-balancing kernel administrator software IPVS usually takes the VS/DR method to set up a media cluster system. The load balancers sends the user's media service requests to servers of the cluster in balancing and the media server reply to the clients directly., so these make the whole media Cluster very retractile. Media Server could run lots of medium service software.

Now, Linux Virtual Server Cluster has good sustainability for Real Media, Windows Media and Apple Quicktime. Generally speaking, media services usually take a TCP collection (such as RTSP protocol, Real-Time Streaming Protocol) to consult with bandwidth, control the velocity of flow and return to clients directly.[4] Here, IPVS scheduler provides a function considering TCP and UDP together, which sends services to the same media server.[5]

2 Large-Scale Video Server Cluster

2.1 Large-scale video server cluster structure

The demonstration entironment of large-scale video server cluster system is made up of one center management node, one center service node and some verge service nodes (POP service nodes), as following Fig3:[6]

The center administer platform consists of global load balance modules and Web gate-door content-catalog generator modules. Among them, global load balancing modules accomplish consumer redirections according to 'services with the near load balancing' algorithm. According to near services and the content, as well as load balancer and all the POP nodes, requests from clients are redirected to the appropriate node. If the POP near the consumer couldn't provide service, we should consider the near POP, and then reconsiderate the center services. [7]

POP services nodes take the Cluster technology to realize fuctions. The structure is shown in Fig4. It is made up of one front server and lots of video servers. Front server redirects service to one video server according to the load balancing principles. The service data stream flows from video server to consumer immediately.[9] The below part of the Fig3 is RAID-M, which functions as a great capacity storage, and accesses service data provided by the video servers.

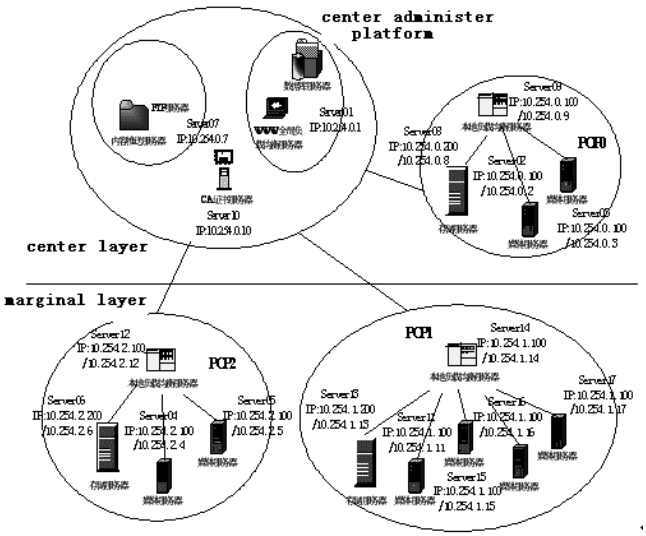


Figure 3 System Chart of Large Scale Video Service Cluster

2.2 Directing Routing (LVS-DR) [10]

In an LVS-DR configuration, the Director forwards all incoming requests to the nodes inside the cluster, but the nodes inside the cluster send their replies directly back to the client computers (the replies do not go back through the Director). As shown in Fig4, a request from the client computer or CIP is sent to the Director's VIP. The Director then forwards the request to a cluster node or real server using the same VIP destination IP address (we'll see how the Director does this). The cluster node then sends a reply packet directly to the client computer, and this reply packet uses the VIP as its source IP address. The client computer is thus fooled into thinking it is talking to a single computer, while in reality it is sending request packets to one computer and receiving reply packets from another.

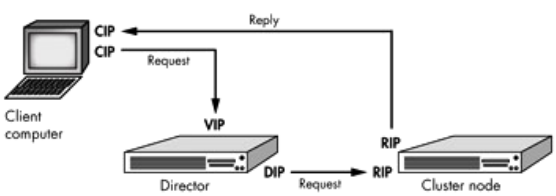


Figure 4 LVS-DR Network Communication

The Linux Virtual Server Direct Routing (LVS-DR) cluster is made possible by configuring all nodes in the cluster and the Director with the same VIP address; despite having this common address, client computers will only send their packets to the Director.

2.3 Service Provider POP Node [8]

The hard disk of video servers keeps the latest films. If the film requested by consumers doesn't exist, it can read from RAID-M. It constructs a three layer storage System. [8] The fist layer is memory of server, which stores playing films. The second layer is hard disk of the server, which stores latest played films. The third layer is RAID-M, which stores all the films and provides services for all video servers. From the first layer to the third layer, the access speed is lower and lower, the storage capability is larger and larger. If you want to improve the performance, you should take full advantage of the first and second layer. One POP node acts as the Fig5, the node takes a VS/DR Schema as follows:

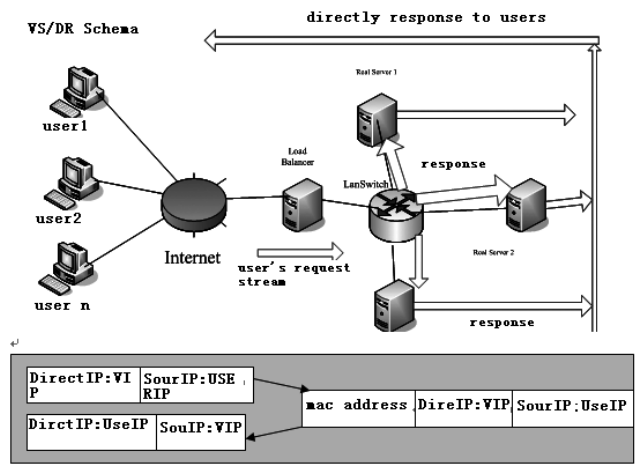


Figure 5 The Structure of service provider POP node

As shown in Fig5, service provider POP node adopts a strategy of direct routing to realize the virtual server. When consumer's requests from the center to POP arrive the director, the director dynamically chooses one server according to the load instance first, and then modifies the MAC address of the data frames with the chosen MAC address of the server, and sends it to the server in the LAN. [9]

Because the MAC address of data frames is the chosen server's MAC address, the chosen server could receive the Data frames surely. When the server finds the VIP of a target address on the server is on a local network device, the server will deal with this message, and then return the response message direct to the client according to the route-table.[9]

3 Schedule Strategy[11]

3.1 The Film Chosen Model

It is a stochastic process about when the consumer enters the system. The clients accessing films is not an average process, but obeys a probability distributing, generally, it obeys Zipf distributing.

$$p_k = \frac{k^{-\theta}}{c} \qquad (1)$$

Here, $c = \sum_{k=1}^{n} i^{-\theta}$, $0.271 \le \theta \le 1$

The expressions above denote probability of visited films. In the formula, θ is a constant, sometimes called depth gene. Bigger θ means the orientation is higher above the average level of movies visited by consumers. Sometimes the depth gene is not same to different consumers and different program concourse. For example, the statistics about video hire industry in 1998 shows that the hire number of $\theta = 0.70$, according with zipf.[11] Many trial data indicate that customers of 27% only locally visit the movies of 71.3%, this trait means that local quality of movie visiting is very good. It is very helpful in large video cluster system to take the strategy of 'services in nearest'.

3.2 LVS Scheduling Methods [10]

3.2.1 Fixed (or NON-Dynamic) Scheduling Method

In the case of fixed or non-dynamic scheduling methods, the Director selects a cluster node to use for the inbound request without checking to see how many of the previously assigned connections are active. Here is a current list of fixed scheduling methods:

Round-robin (RR) When a new request is received, the Director picks the next server on its list of servers, rotating through them in an endless loop

Weighted round-robin (WRR) You assign each cluster node a weight or rank, based on how much processing load it can handle. This weight is then used, along with the round-robin technique, to select the next cluster node to be used when a new request is received, regardless of the number of connections that are still active.

Destination hashing This method always sends requests for the same IP address to the same server in the cluster. Like the locality-based least-connection (LBLC) scheduling method, this method is useful when the servers inside the cluster are really cache or proxy servers.

3.2.2 The main Dynamic Scheduling Methods[10]

Dynamic scheduling methods give you more control over the incoming workload, with little or no penalty, since they only require a small amount of extra processing load on the Director. When dynamic scheduling methods are used, the Director keeps track of the number of active and inactive connections for each cluster node and uses this information to determine which cluster node to use when a new arrives for a cluster service. An active connection is a TCP network session that remains open (in the ESTABLISHED state) while the client computer and cluster node are sending data to each other. As of this writing, the following dynamic scheduling methods are available

Least-Connection (LC) With the least-connection scheduling method, when anew request for a service running on one of the cluster nodes arrives at the Director, the Director looks at the number of active and inactive connections to determine which cluster node should get the request.

Weighted Least-Connection (WLC) The weighted least-connection scheduling method combines the least- connection method and a specified weight or rank for each server to select the cluster node. (This is a default selection method if you do not specify one.) This method was intended for use in clusters with nodes that have differing processing capabilities.

3.3 Load balance Arithmetic [11]

The main idea of the services in nearest is that we lookup the brim node of movies in turn, whether the rest brim nodes and center node could provide service for

task or not, whether the gist of provided services is a copy of service provided by the node. Because the center has all the films, if the task is refused, it must be that the computing capability of the center is on full burthen or does not provide network bandwidth of services. [11] The flow of the arithmetic is as follows:

（1）Find whether there is a film copy in the brim node, if not, go to （3）, and else go to next;

（2）Make sure whether the load of brim node is full or not, if yes, go to （3）, and else go to next;

（3）Search one of the other computing nodes;

（4）Judging whether the computing node has films which the busywork needs, if not, go to （8）, else, go to next;

（5）Judging whether network bandwidth consumer used is full, if yes, go to （8）, else, go to next;

（6）Judging whether the load of the computing node is full, if yes, go to （8）, and else go to next;

（7）If the Load of the computing node is lower than the least load number, then define lowestLoad= Load, clear the lowestLoadQ chain and let it empty, add the computing node into the LowestLoadQ chains. If lowest Load is lower than Load, add the computing node into the chain lowestLoadQ;

（8）If there is a computing node available, return to （3）, else go to next;

（9）If the least load chain lowestACQ is not null, choose one of the computing nodes by chance;

（10）Judging whether the network bandwidth is full, if not, go to （2）, else, go to next;

（11）If there has a leisure in the center node, return to the center node, else, go to next;

（12）Return null.

3.4 Experimental Testing

The collocation of testing environment is: three RS medium servers, one RAID-M medium storage, and one local load balancer. On the RAID-M, there store 100 movies. Via samba server, we can provide services to RS.[11]

The testing result of video cluster is listed in Table 1

Table 1 The testing result of video cluster

Server name	Number of films	Rates of films	Concurrent streams	Flows
Server1	1 film	750Kbps	600	439Mbps
Server2	1 film	375Kbps	1200	439Mbps
Server3	100films	375Kbps/50films 750Kbps /50 films	200	110Mbps
Total	102films	Avarage 62.5Kbps	2000	988Mbps

4 Conclusions

After the director adopting nearest-service load balance algorithm, the Server1 and Server2 service one film alone, and convert other films to Server3. The service capability of the whole cluster can touch 2000 video streams. But the other video systems, running at the same collocation, could reach 800 streams in peak volume at the most for the validity of the load balancing strategy.

References

[1] Zhang Wengsong. The Linux Virtual Server Project, http://www.LinuxVirtualServer.org/

[2] Zhang Wengsong. Linux Virtual Server Scheduling Algorithm, http://www.linuxVirtualServer.org/html

[3] ENGELSHALL RS. Load balancing your web site, WebTechniquesMagazine,Vol.3,No5,July,2005, pp20-23

[4] Alex Vrenios. Linux Cluster Architectur [M]. USA Sam.Vol.3, No 4, July 2006, pp.10-12

[5] Red Hat High Availability Server Project, http://www.ha.redhat.com/

[6] Zhang Wensong,Linux virtual cluster system http://www-900.ibmcom/developerWorks/cn/linux/cluster/lvs/part4/index.shtml.(1-4)

[7] Zhang Wensong, Linux Virtual Server of Telescopic network services, http: //www.ccw.com.cn, 2002,8:(1-3)

[8] Tian Shaoliang, an improved load balance arithmetic based on dynamic feedback. Engineering and design , Vol.2, No.2, July 2005, pp.15-20

[9] Zhang Wensong, Linux virtual server cluster load balancing schedule,Vol,4,No.6,July 2006, pp21-23 http://www-900.ibm.com/developWorks/cn/linux/luster/lvs/part4/index.html 002-05

[10] The Linux Enterprise Cluster. By Karl Kopper No Starch Press 2005,Vol 2,No 10,July 2002, pp.30-36

[11] Liu peng. The storage mechanism of large-scale video service,Vol.10,No.3,July 2002, pp.5-25

A Simple Practical Design Idea of the Campus Network Monitoring System and Its Implementation

Aizeng Qian

Department of Computer Science and Technology Dezhou University, Dezhou, Shandong, 253023,China
Email:qianaizeng@163.com

Abstract

According to the needs of the campus network actual operation and management and aiming at the various defects of the dedicated network management system, this paper presents a simple practical design idea of the campus network monitoring system, according to the idea, a campus network monitoring system is designed and developed, some problems in day-to-day operation of the campus network management needing to be addressed urgently are effectively soluted. this paper dwells on the design principle, the functional modules, the data flow, the database design and part of the core code for the achievement of the functional modules of the system.

Keywords: PING, cron, PHP, C/S, B/S, NMS

1 Introduction

With the development of information construction, the scale of the campus network is increasing, and it has become an important support platform for the work of the colleges and universities[1]. So it becomes an important part of network management to keep the campus network in its normal operation, and the primary task to keep the campus network in its normal operation is to ensure the core campus network switching equipment, convergence layer devices and the core server running normally.

To achieve this purpose, many computer companies have developed the corresponding network management system can be effective analysis, management and monitoring to these equipments[2]. However, there are some problems in these network management system, on the one hand, these NMS is a vendor of equipment for research and development, does not have universal, and campus online operation of equipment from various manufacturers are generally pose, and most of these network management system based on C/S mode operation, inconvenient to use mobile in the campus network[3]; on the other hand these NMS's functions are often too complex, and the requiring of the day-to-day maintaining work is relatively simple[4], the majority of them can be generally determined whether it is normal through observing the status of devices.

Based on the above analysis, with the experience of the author in the campus network management over the years, with the MySQL database for storage platforms, the use of the operating system Linux RedHat cron timing function, PING commands, C and PHP programming language the author develops a simple and practical campus network monitoring system based on B/S mode (hereinafter referred to the Campus Network Monitoring System), and achieves a better management effectiveness.

2 The Campus Network Monitorying System Design Principle

PING is a command that the network operating system provides to determine whether the local host can exchange (send and receive) data with another host. According to the returning information, you can infer whether it is in normal operation the other equipment and its communication condition[5]. Figure 1 is the communication situation from local host to the remote host which has the IP address 211.64.32.1, the time of

two returning packets is less than 1 ms, we can judge the host in the normal operation based on the information to the of the network, if we use PING order to test all the Campus Network Equipment every time, from the results we can judged the campus network operation status, and the purpose of Monitoring the entire campus network achieve.

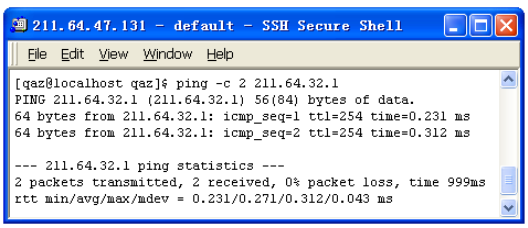


Figure 1 PING test result (normal)

In the system design process using cron timing function cyclically calling PING commands to obtain test results, then use C language analyze the results and store them in the MySQL database, and then PHP cyclically to visit the MySQL database and displaying them in B/S mode, and the purpose of Monitoring the entire campus network achieve[6].

3 The Campus Network Monitoring System Components

The campus network monitoring system is composed of the following functional modules as shown in Figure 2, the module functions are as follows:

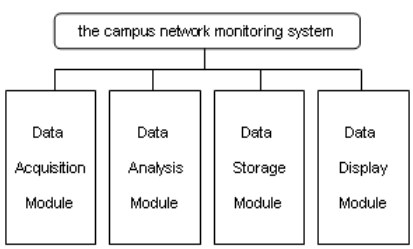


Figure 2 Campus Network Monitoring System components components

3.1 Data Acquisition Module

The main function of data acquisition module is to obtain the test results of the core equipments running on the campus network with the PING order to test them , and the results will be analyzed by the analysis module.

3.2 Data Analysis Module

The main function of the data analysis module is to analyze the test result which obtained in the acquisition module and it will be used in other modules.

3.3 Data Storage Module

The main function of the data storage module is to analyze the status of the equipments obtained in the data analysis module, and then they are stored in the database with a certain format.

3.4 Data Display Module

The main function of the data display module is to cyclically call the state value of the various equipment stored in the database, then the status of the various equipment will be displayed in the browser in a certain way, and the entire campus network is under monitoring.

4 The Implementation of The Campus Network Monitoring System

4.1 Operating environment of the campus network monitoring system

(1) Hardware environment

The campus network monitoring system has a low hardware environment, and the P Ⅳ computer can meet the requirements, a harddisk more than 20 G, a 100 Mbps Ethernet interface card , and can access the campus network..

(2) Software environment

The software platform of the campus network monitoring system is Linux RedHat9.0 network operating

system, the versions may also be other, requires installation of integrated MySQL, PHP and Apache environment, the Linux system cron timing features should be started.

4.2 Data flow diagram of the campus network monitoring system

(1) Linux cyclically calls the operating results of the switch equipments and servers which obtained by the data acquisition module, and then the data analysis module analyzes the above results, finally, the data storage module stores the results to the database in a certain format, as shown in figure 3.

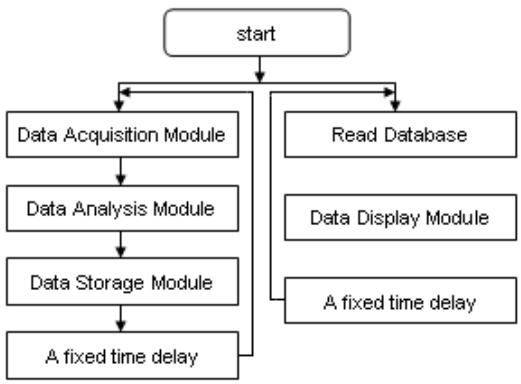


Figure 3 Data flow diagram

(2) The display module cyclically reads the data values that are stored in the database, and displays in a certain way , as shown in Figure 3.

4.3 Implementation of the campus network monitoring system

4.3.1 Database design

This system requires a device table, the data dictionary is shown in table 1:

Table 1 the device table data dictionary

field	type	Null	default value
Id	int(6)	否	
Name	varchar(30)	否	
Ip	varchar(15)	否	
Min	int(7)	否	0
Avg	int(7)	否	0
Max	int(7)	否	0
Time	datetime	否	0000-00-00 00:00:00

The id field marks the record number, the name field marks the device name, the ip field marks the device's ip address, and the min,avg,max field each presents the minimum , the average and the maximum of the equipment feedback time, and is a detailed description of the state of the equipment, the time field marks the record time.

4.3.2 The implementation of the function modules of the campus network monitoring system

(1) the implementation of the data acquisition

Firstly, establishing a order list file ip.sh including all the IP address of all the devices, the content format is: "ping-c number device IP > device IP_result", one line test one device. the aim of plus "- c" parameters is to let Linux system to enable the implementation of the second, otherwise it will have been implementing[7], the order will write each of the test results to a "device IP_result" file which will be analyzed by the data analysis module for analysis.

secondly, let linux network operating system regularly call the order list file, cron[8] is a timing tool in the linux network operating system, can run operations without manual intervention[9]. establishing a text file cron.ip, giving out the period of calling, a test file content that let the linux network operating system to call a list every five minutes is as follows:

5,10,15,20,25,30,35,40,45,50,55 * * * * ./ip.sh

Finally, in "#" prompt administering the order crontab. / Cron.ip, and the timing features can be write into the Linux network system.

(2) the implementation of data analysis

the function of data analysis module is mainly to analyze the device test results according to some related to signature. with practical application we know that there are two test results shown as Figure 1 and Figure 2 with a right network configuration, with two data packets, we can see that the file only has 5 line if the network is right shown as Figure 4, and that the file only has 7 line on the other shown as Figure 1. through the analysis of the test result as Figure 1, we find the minimum time ,the average time and the maximum time can present the network communication conditions in last line, the core code is follows:

if((fp=fopen("21164321_result ","rt"))==NULL)

```
{printf("Cannot open file strike any key exit!");getchar();exit(1);}
    ch=fgetc(fp);j=0;i=0;
    while (ch!=EOF)
    {res[j]=ch;
j++;if(ch=='\n'){res[j]='\0';i++;j=0;line++;}ch=fgetc(fp);}fclose(fp);
    //get minimum, average and maximum time
    if(line==5){min_s=0;avg_s=0;max_s=0;}
//network is abnormal
    if(line==7)        //network is normal
    {for(i=23,j=0;i<=27;i++)if(i!=24){min[j]=res[i];j++;}min[j]='\0';min_s=atoi(min);
    for(i=29,j=0;i<=33;i++)if(i!=30){avg[j]=res[i];j++;}avg[j]='\0';avg_s=atoi(avg);
    for(i=35,j=0;i<=39;i++)if(i!=36){max[j]=res[i];j++;}max[j]='\0';max_s=atoi(max);}}
```

(3) the implementation of data storage

The function of the data storage is store the result and the relelated content into the MySQL database, the core program code is as follows:

```
    mysql_init(&my_connection);//initialize database
    if(mysql_real_connect(&my_connection,"localhost","dbname","db_password", "db_uesr", 0, NULL, 0))
    {printf("Connection success\n");
    sprintf(sql, "insert into device(%s,%s,%d, %d,%d, %s )",name,ip,min,avg,max,time);
    mysql_query(&my_connection, sql); //insert data
    mysql_close(&my_connection); //close connection
    }
    else fprintf(stderr, "Connection failed\n");
```

(4) the implementation of data display

The main function of this module is to cyclically call and display the content stored in database, There are several key steps, firstly, letting IE regularly refresh, this requires to add the code "<meta http-equiv= "REFRESH" content="300; http://192.168.1.252/index.php" />" in php program head[10].

Secondly, connecting the database and getting the data from database, the core code is follows:

```
    $result=$db->query("select * from device order by id");$i=0;
    while($array=$db->fetch_array($result))
    {$ip[$i][0]=$array["id"];
    $ip[$i][1]=$array["name"];

$ip[$i][2]=$array["ip"];$ip[$i][3]=$array["min"];
    $ip[$i][4]=$array["avg"];$ip[$i][5]=$array["max"];
        $ip[$i][6]=$array["time"];    $i++; }
    $rows=$i;
```

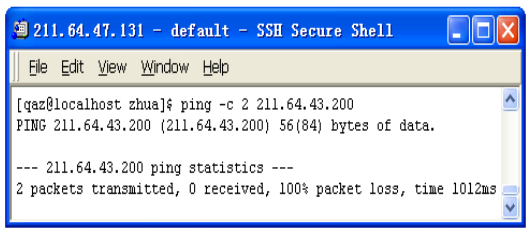


Figure 4 PING test result(abnormal)

Finally, displaying the status of the device in the designated location, the core code is as follows:

```
    for($i=0;$i<$rows;$i++)
    { $aa= ($ip[$i][3]+ $ip[$i][3]+ $ip[$i][3]}/1000;
     if($aa<10) { Calling the picture that presents normal condition, Tip information network communication is very good.}
    if(($aa>=10)&&($aa<100)) { Calling the picture that presents normal condition, suggesting that the network run slower, attention!!.}
    if($aa>=100) { Calling the picture that presents equipment failure, suggesting that there is the virus in the
    network ,handles it immediately.}
```

The running campus network monitoring system is as shown in Figure 5。

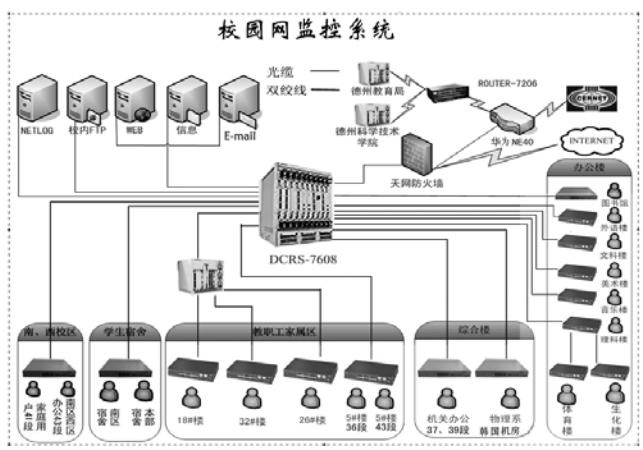


Figure 5 the running campus network monitoring system

5 Concluding Remarks

The campus network monitoring system referred in this article is a network management software which is developped according to the actual network management needs of our university, one year actual operating experiences in our campus network show that this management system can effectively monitor all the network devices running in the campus network, and the the campus network management difficulties are alleviated in greater extent. we are currently revision, and its voice, e-mail and text message alert function will be added we believe that the system will play a greater role in the future campus network management .

References

[1] Liu Rui Chao,Jiang su-bin, Design and implementation of a device-level network management based on the World Wide Web service[J], Network Applications, 2007,12,16~19

[2] Wang Jiangping, Computer network management system and application [J], Nanjing Xiaozhuang university Journal,2002,18(4):88~92

[3] Zheng Xiuying, Chang Gui-ran, Network management data collection on the adaptive algorithm [J], Computer application and software,2007,24(2),159~160

[4] Zheng Xiuying, Chang Gui-ran, Network management data collection on the adaptive algorithm [J], Computer application and software,2007,24(2),159~160

[5] Ma Yan, Zhang Xiao Zhen, A pro-active network management framework of Design and Research [J] , Computer science,2007,34(2),38~42

[6] Xing Junfeng,Fan Tai-hua, Research of 4-layer network management based on linux[J], Micro-computer information, 2007, 23(1), 305~307

[7] Jin-hong,Zhu Ya-bin, the design of IBA network management system and the implementation of the Diagnosis system[J], Computer engineering and design, ,2007，28(3),562~565

[8] use of cron,http://doc.linuxpk.com/2423.html, 2007,12,31

[9] Detailed Explanation of the implementation and examples of linux timing, http://blog.ouc.edu.cn/2007/01/cronlinux.html, January 31 2007

[10] META HTTP-EQUIV="refresh",http://hi.baidu.com/yyjjbb/

A Way of Using Web Service by AJAX

Zhensheng Wu

Information Department, JiangNan University, WuXi, JiangSu, China
Email: jasonwood612@yahoo.com.cn

Abstract

Nowadays using web service has become the most popular way of designing software, and AJAX also plays a very important part of Web2.0. This article gives an example to combine these two hot technologies in web application development. At the beginning of the article, we give the definition of web service and AJAX from the view of a programmer. By comparing with the traditional ways of designing, we wants to show the advantages of both the two technologies, discussing why and how to use them to develop a software/service based on the spirit of SOA structure. At the end of the article, we give a simple example to show how to add a web service into a program by using AJAX and discuss a few parts of it by comparing with the way without AJAX or web service to show their differences and finally make a conclusion.

Keywords: Web Service, SOA, AJAX,Web2.0.

1 Web Service And Ajax

Web Service

Maybe in the next 20 years, all software will become web service or something like that. By using Internet, programmers and customers will find there would be nothing can be strictly called “software”.

Programmers will design their work by just combining different parts of service, most of them can be found on the Internet, it may be free or may be charged, that does not matter. Some parts may not be perfectly matched to the program, so all they have to do is just to make a few changes of the service they get from internet to fix the problems and develop some new functions, finally they sell the program as a web service, not as a software installation. It just like if we want to make a meal, we needn’t to buy a whole market, all we have to do is just picking up something we want and make our own changes to use or sell it [4].

For the customers, they will not buy discs or download an installation of software, they just need to register and use the software as a web service directly and pay with credit card. The way we do now is to buy a lot of copies of software and then we find that 90% of these software are just the same, so maybe in the future if we want to get a new application, we can simply buy the 10% of the software which we don’t have----as a web service.

As a matter of fact, these will not be too far away, even then there will not be a clear difference between designers and customers. Everyone can use the service as a customer, everyone can become a designer by just making a few changes on the service and then publish it. Well, it is also the spirit of web2.0.

Why does web service become so popular? From the view of a programmer, what is web service? I believe there will be millions of articles to give the definition of web service, and they all mentioned the three most important parts of this technology. They are WSDL, SOAP and UDDI.

What is WSDL

WSDL is short for Web Service Description Language which based on XML. Web service will be given and used on net, so there must be a protocol to tell the service publisher and user how to make the match [5].

To the publisher, they have to make this protocol to give some important information about the service, such as transport mode and protocols, methods and interfaces, parameters, and URLs. This should be an instruction of the service, moreover it prescribes how to use the

service. Simply said, WSDL has three tasks to tell the users:

What is this service?----Type and message, containing the methods and interfaces, parameters, messages of request and respond or error

How to use this service?----Protocols, mostly is SOAP based on HTTP

Where is the service?----The URL of the service, it may be local or on the internet.

With these three questions cleared, the user can use the service correctly. So WSDL is an introduction of the service and provided by the publisher, sometimes it can be built automatically by the IDE.

What is SOAP

SOAP is short for Simple Object Access Protocol. It is just a so-called word, because now the objects we use in web service are no longer simple, it gives a standard format of the XML.

Basically saying, it gives a language for both the publisher to output the results and the users to input the parameters and data. As we know, the system and environment we use may totally different with the service's on the server, so an universal language and format between them is required, that is SOAP. So the client may be developed by Delphi, the service on the server may be a java program, no matter the customer use Windows or Linux, with SOAP, the data will be transferred between them in a standard and extensible form. It will absolutely solve the problem of software compatibility [6].

What is UDDI

UDDI is short for Universal Description Discovery and Integration. It tells the publisher how to publish the service on internet and how to find a service on net. It's like a yellow page of service, such as http://uddi.ibm.com.

Actually this part is not necessary for every web service, because in WSDL, we have already given the URL of the service, so the user can simply use it by the URL directly, but if you want to publish it to let more people know your service and sell it, then UDDI is required.

AJAX

Nobody can give an exact definition of web2.0, as an important part of it, AJAX have the same problem. AJAX is short for Asynchronous JavaScript and XML. It is not even a technology but a combination of many technologies; it includes JavaScript, XHTML, CSS, DOM, XML, XSTL and XMLHttpRequest. Each of the words above can be discussed in a new article and they are all well used by us for years.

In this article, AJAX is a technology to use JavaScript and XML to transfer data between browser and server asynchronously [7].

To make a brief article, we just want to talk about the most important part of AJAX, that is XMLHttpRequest, JavaScript and DOM, you can also find articles directly about them.

XMLHTTPRequest

XMLHTTPRequest provides a protocol to communicate with the HTTP server asynchronously. It makes the transmission into the way we feel in a desktop program instead of waiting the pages to refresh again and again. By clicking the button "submit", in the traditional pages without AJAX, we just handle the data or a form of data to the server and wait until the server responds and give the output back. In AJAX, we don't give all the works to the server all the times, sometimes the client does part of the job when waiting the necessary response of the server, not just waits. So it liberates the server somehow and decrease the time we wait [9].

Here is an example of XMLHTTPrequest object based on JavaScript:

```
var xmlHttp // a variable to the XMLHTTPRequest object
function createXMLHttprequest(){
if (window.ActiveXObject){
xmlHttp = new ActiveXObject("Microsoft.XMLHTTP");
    }
else if (window.XMLHTTPRequest){
xmlHttp = new XMLHttpRequest();
}
}
```

After creating this object, we can use its methods and properties in JavaScript to implement AJAX.

JavaScript

We use JavaScript to handle the XMLHTTPRequest

object to fix the transmission to the server asynchronously. To make the article short, we don't talk too much about it, there are lots of books about JavaScript---- in AJAX or not.

DOM

Document Object Model is a few APIs to use HTML and XML. It provides structural form of documents and defines how to access the document structure by JavaScript. With DOM, we can edit the dynamic pages, every element in the document is a part of DOM, and we can access the property and method of these elements by JavaScript. Using the APIs of DOM is the kernel part of AJAX [3]. Here is an example of DOM.

```
<html>
 <head>
<title>DOM document example</title>
 </head>
 <body>
<table>
    <tr>
        <td>name</td>
        <td>password</td>
    </tr>
    <tr>
        <td>admin</td>
        <td>admin</td>
    </tr>
 </table>
</body>
 </html>
```

2 Traditonal Ways And Soa

Traditional Ways

In traditional ways we design software into many modules, each module has a function. So when we design another software, even they are nearly the same, we have to divide the software into the same modules. Although we can use some of the codes we have written before, it's a meaningless and absolutely a tough work. So a lot of companies start to develop something called middleware and structural framework like CORBA to make it easier to reuse the code. It somehow works, but everyone knows how complex CORBA is and when you are making a small application, we will find the time we spend on the structure sometimes is even a lot more than just copying the codes of the modules we wrote before.

SOA

SOA is short for Service Oriented Architeture. As we mentioned above, each module of the traditional software has a particular function, so each of them can be built as a service, all we need is a few protocols to make the input and output canonical, after this work, developing software becomes just combination and standardization. It makes the programming easier and able to build larger programs.

SOA is a concept, it is also a platform of services. The services can be local or web ones. So the kernel of it is a platform of management of the services we use [2].

Now the most popular way to build a SOA platform is:

Use Struts as the MVC architecture.

Use session bean or spring to manage the actions and operations.

Use hibernate or CMP to manage the link with database and the data object.

Use web service to provide standard interfaces.

Use BEA as the service bus.

3 Using Web Service By Ajax

As discussed above, web service and AJAX are both very popular and well grown. The famous companies and websites such as IBM and google.com are the best examples of these two technologies. We will provide a small application to show the basic elements of how a web service is built and used by AJAX.

In this example, we provide a web page that has a very simple function. This page has is a blog with many articles and music. It also provides the users to use the Yahoo search for new songs, by clicking the submit button, the browser will show the results according to the key words we entered as a search engine without

starting a Yahoo.com page. This application is commonly used to build web sites and blogs.

We can make this application in three ways. In the chapter below, we will discuss how they are designed and the advantages/disadvantages of each way briefly. Some codes of the kernel part are given as well to show the technologies we discussed at the beginning of the article.

Build the page with just struts

Commonly, we built every application by a framework such as struts. Struts is a widely used framework of MVC structure. To develop a search engine page, we need to build a jsp page and an action form, and configure the two important XML documents----struts-config.xml and web.xml just like we develop every application. By clicking the submit button, the action form we built will record the key word we just input and start the corresponding action to compute the results, with some search algorithm, at last the action form put the results back onto the page and refresh [1].

That is the brief data flow of a struts application and here comes a big problem. In the world of algorithm, the algorithm of search engine may be the most difficult one, the database we need (if we build it ourselves) could be quite huge, facing to such a big algorithm problem in such a small application, if we insist to build it, that would be a very huge work and million lines of codes----for just one page [10].

That is also a commonly faced problem, when we are developing some small applications, some of them are quite simply and we just write the operations and functions and put them into a framework, some of them may be nearly the same as the application we developed may be just a weak ago, mostly we just copy the codes to this new application and then find it does not work, so we get back and analysis the tiny difference between the two applications and fix the problems, the time we lost may be enough to rebuild a new one.

So many people put their codes into modules, that is a good way and mostly it works well, but in this application, how can we find a module of a big search engine like Yahoo's? That may be the top secret of them and absolutely huge, if we build it in this way----if Yahoo sells the algorithm to us-----we will find the application has a quite huge body with a tiny heart.

Struts is an excellence framework to build J2EE applications, but we find it's hard to develop this one with just struts.

Build the page with web service

The Yahoo search can be found on UDDI as a web service, it provides the same search engine to the users, with this web service we can do the same search job without enter www.yahoo.com. So we can build the application with Struts framework and add this web service to avoid the problem we mentioned above. We have already discussed how to add a web service to an application. By doing this, the application will be quite small; we leave the most complex part to the web service without knowing how the results come out.

But there is still a problem. This is a blog page with lots of articles and multi-media elements such as songs and pictures, so every time we start a search, we have to wait until it refresh, the most part may not change after every refresh, but these elements which don't change at all will cost us a great deal of time.

When we are waiting the refresh, we can not read the articles, we can not hear the songs or watch the pictures. We all have the same experience----the page is good, but the time we wait is too long to see a new element, if we have to wait all the element refreshed just to get one of them new, mostly we will stop the tour on this page.

So using web service in this application is a smart way, but just using it directly may not be the best way.

Build the page with web service and AJAX

We use web service to make the application small but we have to face some other problems, in this one, if we use AJAX to make the search, only the search part of this page will be refreshed. And because it will access the web service on the server asynchronously, so during the time we wait, we can still keep reading the articles and listen to the music of this page [8].

We have given the code to create a XMLHTTPRequest object, the codes below is the AJAX part of the application. It is not integrate.

//here is a function to get the search web service by ajax

function search(){

```
if (document.all.query.value = =""){
   alert("please enter the key word");
   return;
}
//get the key word
context = document.all.query.value;
//open the image of search waiting
codument.all.wait.style.display = "inline";
//create a XMLHTTPRequest
creat XMLHttpRequest();
xmlHttp.onreadystatechange = processor;
//create the access to the YahooSearch
xmlHttp.open("GET","YahooSearch?query="+co
ntext+"&results="+rssize);
//send the request
xmlHttp.send(null);
}
//handle the xml from the server
function processor(){
if   (xmlHttp.readystate = = 4){   //if respond
     if (xmlHttp.status = = 200){   //if receive the xml
          //close the image of wait
          document.all.wait.style.display = "none";
         //get the child node of the result from the xml
          result=xmlHttp.responseXML.
getElementByTagName("Result");
               parseResult();
          }
else{
     document.all.wait.stlye.display = "none";
     alert("sorry, search failed);
}
}
}
//the function to display the result
function parseResule(){
     //here are the codes to manage the display of
the results and the search part of the page
}
```

We passed over some of the unimportant codes such as the codes of displaying the result. With AJAX the refresh will affect just a part of the page, which will change after the search. This will absolutely make the users more comfortable.

4 Conclusion

We have discussed the concepts of web service and AJAX, we also introduce some technologies of them, they are all widely used nowadays in the field of developing software and web services. We compared the traditional ways of developing with SOA structure; we also talked about something on web2.0.

AJAX and web service are the most popular way to develop web application and they will absolutely become a big part of the web2.0. By giving a simple example we want to show why we should use them and how to make the start. There are lots of such examples on the Internet and they are perfectly giving us a new view to the web, also a new view to the way of designing and programming.

References

[1] Bing Liu, The conformity of Java web developing with JSP+AJAX+Struts+Hibernate, Bei Jing, Publishing house of electronics industry, 2008.

[2] Ai Hu Liang, The conformity of application develop based on SOA and Sruts+EJB+Web service, Bei Jing, Publishing house of electronics industry, 2008.

[3] Cameron Adams, Mark Boulton, Andy Clarke, Simon Collison, Jeff CroftWeb Standards Creativity: Innovations in Web Design with XHTML, CSS, and DOM Scripting, Friends od ED, 2008.

[4] Eric Newcomer, Greg Lomow , Understanding SOA with Web Services, Addison Wesley, 2006.

[5] Steve Graham, Building Web Services With Java: Making Sense of XML,SOAP WSDL, and UDDI, SAMS, 2003.

[6] James Snell, Doug Tidwell, Pavel Kulchenko, Programming Web Services with SOAP, O'Reilly, 2002.

[7] Dave Crane, Eric Pascarello, Darren James, AJAX in Action, Manning Publications, 2006.

[8] Ryan Asleson, Nathaniel T.Schutta, Foundations of Ajax, Apress, 2006.

[9] Gross C, Ajax Patterns and Best Practices, Apress, 2007.

[10] Arnold Doray, Beginning Apache Struts: From Novice to Professional, Apress, 2007.

Architecture of a Decision Support System for Macro-Economy

Guohua Chen[1, 2] Guoqiang Han[1]

1 School of Computer Science and Engineering, South China University of Technology, Guangzhou ,510640, China

2 Dept. of Computer Science, Guangdong Polytechnic Normal Univ., Guangzhou, 510665, China

Email: ghchen2007@yahoo.com.cn

Abstract

This paper mainly introduced a Decision Support System for Macro-Economy, including its infrastructure, related contents and implementing techniques. It had pointed out the hardships encountered by the members in the planning departments of the government and related scientific institutes in the process of predicting, monitoring, statistic and analyzing the developing trend of macro-economy. It had also classified the predicting, monitoring, statistic methods and summarized the implementing techniques for an input-output system used in the development of a decision supporting system for macro-economy. In the meantime, it had analyzed the effect of such a system by investigating a real example in use. As a conclusion, it had pointed out some insufficiency in the current design and directions to pursue.

Keywords: macro-economy; decision supporting system; statistic; prediction; monitoring; input-output system

1 Introduction

As information technology is used more widely in various sectors, the industry had witnessed its rapid and diversified developments in China. Its applications in e-government are also very attractive. E-Government applications had gone through four stages, i.e. the beginning stage, the promoting stage, the developing stage, and the fast developing stage. Now the application of information technology in government, or electronic government administration, had become quite extensive and intensive. Its ultimate goal was to improve the working efficiency and decision-making ability of various government departments by facilitating their economic decision making process with software, which stands out for a higher application level.

The decision supporting system for macro-economy was set forth to help the government members to monitor the overall developing tendency of the macro-economy and to predict its underlying patterns and problems. It could also facilitate the related decision-making process. Firstly, it was necessary to build an index system to enable governmental departments to consecutively accumulate various index data, which was essential for the decision-making process. What was also needed was a set of convenient analyzing and inquiring tools that could easily show data gathered from various sources onto one panel, ready for further analysis and process. At least, a variety of visual analyzing diagrams, such as bar-hart, line-graph, pie-chart, radar-graph, should be available. Besides, data format converters should be provided to enable data exchanges among various types of economy analyzing software.

Nowadays, both the governmental departments and related research institutes have faced the following problems in the process of monitoring, predicting, and analyzing the macro-economy:

- Diversified data: a large number of economic data and literal materials were stored in different departments and offices in various formats. The absence of unified management and efficient organization, not only made data inquiry difficult, but also lead to a possible lost of data.
- Absence of a matured predicting and analyzing tool and related monitoring means: As the quantitative analyzing method was difficult to use, most economy analyzing reports adopted qualitative analyses and obtained superficial findings, whose authenticity also suffered from the interference of the human factor.
- Most of the traditional analyzing, predicting and calculating tools were statistical software developed for expert users, which were too sophisticated to be widely accepted.

2 Methodology

The major prospective users of the our system include the planning departments, the statistic departments, the decision-making departments of governments at all levels and related research departments in the universities and research institutes, especially personnel and specialists working in the development and reform commissions, whose major concerns were macro-economy prediction and monitoring. Besides usefulness, they also required the system to be user-friendly and highly specialized. An analysis of the characteristics of their work showed that a workable decision support system should involve various aspects as statistics, mathematic economics, database technology, software design technology, and network technology.

In order to predict, analyze, and monitor the macro-economy, it is necessary to gather various kinds of relative information with network technology, then composite and classify the data in the view-point of macro-economics, and re-store it in an unified database, and then we can analyze and process the data with statistic method, visualize it or turn it into the demanded formats with the help of computer programs and presented the results on a simple user interface, so that end users can easily predict, analyze, and monitor the macro-economy with ease.

Based on the findings of a long-term investigation of the user's needs of the above-mentioned prospective user group, we divided the system into the following aspects. See Figure 1.

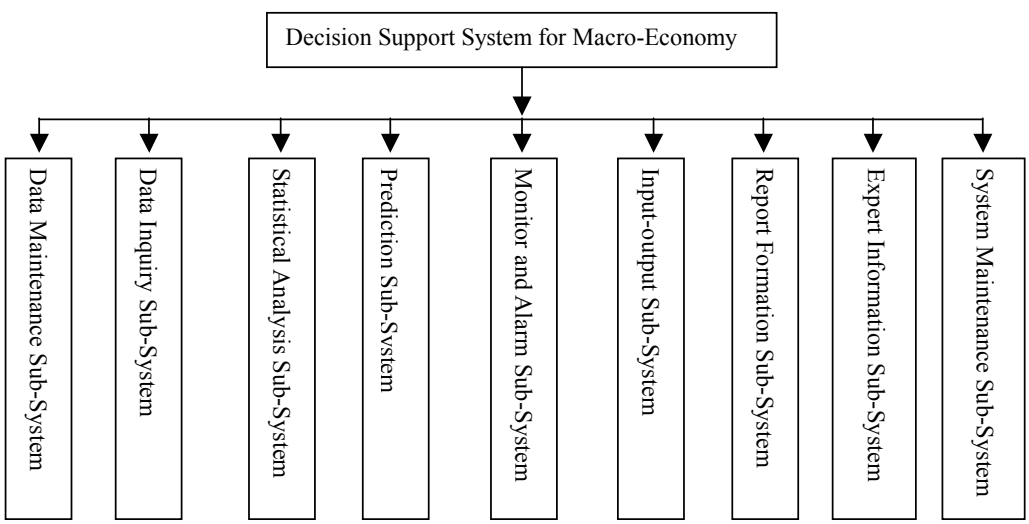


Figure 1 The system architecture Functions of the various sub-systems were as follows

2.1 Data Maintenance Module

It mainly provide the macro-economy database with various data and related data maintenance service, such as governmental sectional indices: national accounts, industry indices, agriculture indices, investment indices, export indices, consumption indices, banking indices, finance indices, etc.

2.2 Data Inquiry Module

It enables users to inquire data in a flexible manner. Besides the usual inquiry method, it also provided:

- Simple Inquiry: via the setting of inquiry formats established from the frequently used indices and graphs, users could get the results immediately by selecting the previously set indices, so the inquiry process was simplified.
- Expert Inquiry: simple calculating and analyzing functions, such as arithmetic or average calculations for composite indices , were added to the usual inquiry.

2.3 Statistical Analysis Module

In order to execute in-depth analyses on the running status, total volume and structure of the macro-economy, this module provided reports and tools for the following statistic analyses:

- Fixed statistical analysis: a large number of fixed analysis reports were established, which not only made the work of compiling statistical yearbook easier, but also met the specific requirements of the user group. New reports could also be added according to users' practical needs. These fixed reports include national income allocation analysis, consumption structure adjustment analysis, investment demand analysis, industrial structure adjustment analysis, cash flow analysis, inflation analysis, economic cycle analysis, and economic growth analysis, etc.
- Specialized statistical analysis: it integrated various specialized statistic analyses.

2.4 Prediction Module

This module has not only integrated the usual predict models such as regression-predict, time series smooth-predict, Grey-predict and auto regression-predict, but also provided the composite predict method and optimal model recommendation method. Users could either choose suitable predict methods according to the prediction objects and goals, or use the optimal model recommended by the system, and a third choice would be using the composite predict to combine several models so as to get the best results.

2.5 Monitor Module

This module had used certain exponent-based indices to signify the macro-economy cycle. Users could diagnose the economy operation, and predict its developing directions and variation range. Alarm could be started if the economic indices showed signs of overheated development or recession, so that relative departments could make adjustments accordingly.

2.6 Input-output Module

This module analyzed the interrelated quantitative relationships among different economic sectors, based on the historical input-output data reports. The present module could

- work out the direct consumption coefficient and indirect consumption coefficient;
- predict demands for overall output and intermediate products based on the given variation data of the finished products;
- analyze how the alteration in the overall output of one economic sector influenced other sectors.

2.7 Report Formation Module

This module mainly included:

- fixed reports: personnel at the planning and statistic departments usually needed to make a lot of reports, and the present module established various types of fixed reports beforehand that were easily accessible;

Customer-defined analyzing reports: users can form their individual analyzing reports by using the customer-defining functions.

Table 1 Situation before and after using the system

TASK	TIME (MINUTES)		EFFECT	
	BEFORE	AFTER	BEFORE	AFTER
looking for macro-economic data	5	0.5	useing statistical yearbook, inconvenient, results presented only in number.	combination of graphs and tables, presented the desired data clearly, convenient, needless to refer to other materials
analyze the macro-economic data	indefinite	1	time and effort consuming, unworkable sometimes, users need to be good at mathematics and arithmetic	could show current status of the economy with figures and tables, etc. after desired indices were selected
predicting macro-economic data	indefinite	1	results were difficult to get by manual operation, and were not satisfactory in terms of accuracy and validity	many predict methods provided quick and accurate results, which were compared with the unprocessed data
forming reports	15	1	useing personal reporting tools resulted in unrelated and nonstandard reports	easily formed standard reports that could be stored and printed
monitoring economic movements	indefinite	2	took a lot of calculation and professional economic analysis	feasible and scientific calculation enabled users to monitor any indices
Input-output analysis	Indefinite	1.5	same as above.	easily determined the quantitative relationships among various economic sectors

2.8 Expert Information Module

This module collected basic information of the relative experts, such as their photos, resumes, specialties, work places, and research areas, and could perform sequential or fuzzy inquiry.

Before using the present system, users may have established their own internal network systems for electronic government administration based on certain types of database and related operating systems. It was necessary to build the present system on an open platform so as to support the existing servers and operating systems. Therefore, it should be independent from the operating system, the database server and the database container. Taking all these factors into consideration, we implemented all the algorithms related to statistics, prediction, monitoring and input-output into a JAVA module, thus established a set of software components

2.9 A Set of Statistics, Perdition, Monitoring, and Input-output Calculation Software Component

The components include the following JAVA algorithms:

- Smooth prediction component: including
- algorithms such as moving average, moving weighted average, moving difference average, first-order exponential smoothing, second-order exponential smoothing, cubic exponential smoothing;

- Regression analysis component: including single variable linear regression, parabolic regression, cubic spline curvilinear regression, etc

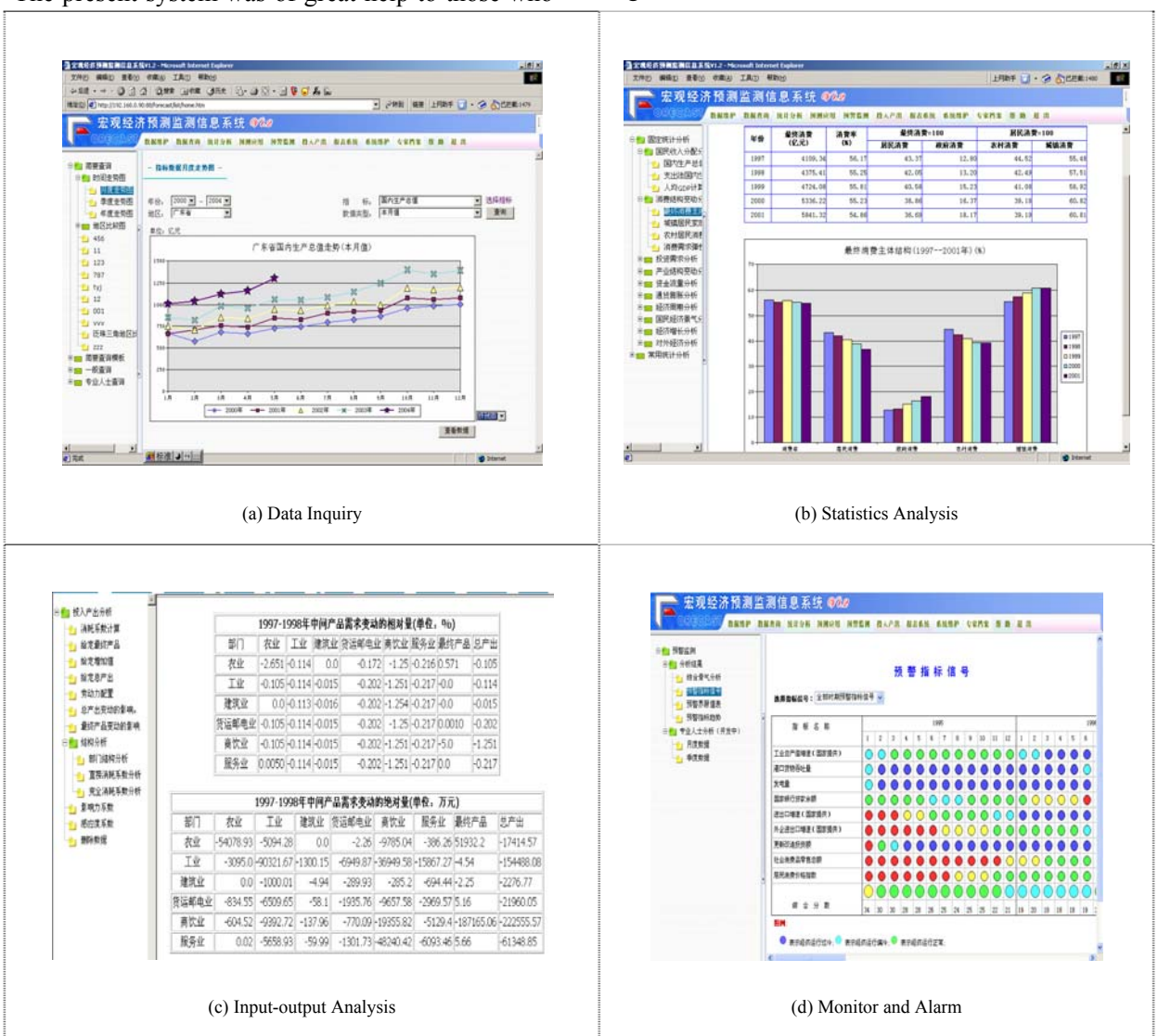
Besides, there are two-stage least squares components, Trend prediction components, Seasonal prediction components, Grey prediction component:, Statistic analysis component, Input-output analysis component and elastic predict component, Box-Jenkins predict component etc

3 Effective analysis

The present system was of great help to those who spent a lot of time and effort looking for macro-economical data and analyzing it. The information center of Guangdong development and reform commission, had adopted the present system, and the table 1 compared the situation before and after their using of the system.

4 Example

Some results of the inquiry, statistics analysis, monitoring analysis, and input-output analyses conducted by the present systems were showed as Figure 2.



(a) Data Inquiry (b) Statistics Analysis

(c) Input-output Analysis (d) Monitor and Alarm

Figure 2 (a).Data Inquiry, (b).Statistics Analysis, (c). Input-output Analysis, (d).Monitor

5 Conclusion

Nowadays, the presented system was not only wildly used in the planning and statistic departments in Guangdong province, but also was known around in China. However, it should be noted that it still had many insufficiencies, and the following directions may be pursued to improve the current design:

- The future system should further strengthen support for the non-linear prediction method, and many recent theory advancements, such as chaos theory and catastrophe theory, had not been integrated into the system.
- Mathematic formulas for prediction were not user-friendly enough, and MathML language should be adopted to improve the user interface.
- User controlled report formation tools needed to be strengthened.

Information technology had not been fully developed in the field of macro-economy prediction and analysis. There were still many important area, for example the catastrophe analysis, calling for further investigation. On the other hand, the policy formulated by the governmental planning departments had great influence on the economic development of the country. So greater efforts should be conducted in this area and more people were expected to contribute their ideas.

References

[1] Bowerman B L, O'Connell R T. Foresting and Time Series: An Applied Approach.(Third Edition). Beijing: China Machine Press. 2003

[2] Feng Z Q. Economic Predicting and Policy-Making. Beijing: China Financial and Economic Press. 1995

[3] Yi D H.. Statistic Estimate, Methods and Application. Beijing: Statistic Press of China(in Chinese). 2001

[4] Wang J M. Practical Model of Non-Linear Movement Input-output. National Defense Industry Press(in Chinese) 2003

[5] Hegger R, Kantz H, Schreiber T..Practical implementation of nonlinear time series methods: The TISEAN package, CHAOS 9, 1999, pp413-420

[6] Schreiber T, Schmitz A. Surrogate time series, Physica D 2000 pp142 -346

[7] Song H, Jiang F.. Program Design for Java Server, Tsinghua University Press, (in Chinese) 1999

[8] Liu Y, Wei F.. Proficiency in JBoss-EJB and Web Service: Development and Illustration, Publishing House of Electronic Industry(in Chinese) 2004

[9] Abraham B, Ledolter J. Statistical methods for Forecasting: New York: Wiley 1983

[10] Box G E P, Cox D R. An Analysis of Transformations, Journal of Royal Statistical Society, B 26 1964 :pp322-243

[11] Box G E P, Jenkins G M.. Time Series Analysis: Forecasting and Control, 2d ed. San Francisco: Holden-Day 1976

[12] Chatfield C ,Prothero D L. Box-Jenkins Seasonal Forecasting: Problems in a Case Study, Journal of Royal Statistical Society, A 136 1973

Study of Supply Chain Inventory Management with Fuzzy Optimization Theory

Jizi Li Zhijun Wu Zhiping Zuo

School of Economic & Management, Wuhan University of Science & Engineering ,Wuhan, 430073, China

Email: jisonli@yahoo.com.cn

Abstract

One inventory optimization model for supply chain was built based on fuzzy logic to properly model the uncertainty associated with both market demand and inventory level. The fuzzy process was represented and the new inventory control approaches were applied in cases study. The simulation and analysis as to both single stage and multi-stage stock with fuzzy controller or classical controllers were separately made. By comparative simulation, the results prove that the new inventory method can obtain better performance such as lower total cost, more stable inventory level and so on.

Keywords: Inventory management; Optimization Model; Fuzzy Logic; Supply chain management

1 Introduction

Inventory usually represents from 20% to 60% of the total assets of manufacturing firms [1]. Therefore, inventory management policies prove critical in determining the profit of such firms. Inventory management is to a greater extent relevant when a whole supply chain (SC), namely a network of procurement, transformation, and delivering firms, is considered，since competition does not occur between single firms than supply chains[2] . Inventory management is indeed a major issue in supply chain management (SCM), i.e. an approach that addresses SC issues under an integrated perspective. The core objective of inventory management is to minimize the total expected inventory costs per unit time while satisfying the customer demand on time. Under the influence of the supply chain management, traditional inventory control theories and methods are no longer adapted to the new environment[3]. So it will have great practical implication to find new methods for inventory control.

In this paper, we analyze and discuss an approach to use fuzzy control applied in inventory control under supply chain management based two cases, which are the single stage inventory and the multiple stage inventory in supply chain, where the demand process is uniform distribution. Firstly, the classic inventory control methodologies are discussed, where the classic inventory model cost equation is given. Meanwhile, a fuzzy control system based on the (s, S) framework is introduced. The core advantages of a fuzzy controller are robustness under uncertainty, expert experiments considered, and inaccurate information. Therefore, the fuzzy controller in inventory control makes it easier to design the system. Next, we apply two inventory control approaches to the two cases, the fuzzy control system will be compared with the classical inventory control system for each case. Then, both systems were simulated over a span of the demand periods for each case using Matlab. The design procedure and fuzzy control effectiveness are illustrated. We compared the fuzzy control system with the classical control system for the different aspect. Finally, some conclusions are drawn out based on the simulation.

2 The inventory control approaches

Inventory is considered as an accumulation of materials or product that will be used to satisfy future demand for materials or product.

Inventory requires the policy to mange the

inventory, which may involve the quantitative problems dealing with controlling material are:

The amount raw material ordered

Lead time to order raw material

Right method used

A minimum cost.

Obviously, the total inventory cost involves all the costs of the related to the previous actions, i.e. the following four factors contribute to inventory costs:

（1）Holding Cost

Cost of storage and space

Capital cost

Service costs incurred from taxes, insurance and obsolescence

Cost of incurred from risks of perishing, damaging

（2）Shortage Cost

Cost of lost sales and good will

（3）Procurement Cost

Cost of preparing, placing orders

Cost of shipping and handling

（4）Average setup cost of placing an order.

2.1 The Classic inventory control approach

As mentioned above, the choice of the policy will be done with the objective of the minimum costs. An inventory policy is the review and ordering discipline used to control inventory.

The most common polices are: periodic review, which need to review a new order of the amount specified by the order quantity at equal interval of time. Inventory levels are observed at equal intervals of time T. Policy defined by:

$$\begin{cases} Q_i = 0 & L_i > s \\ Q_i = S_T - L_i & L_i < s \end{cases} \qquad (1)$$

Where S_T=target inventory level when order placed; s=alarming level; L_i=inventory level at end of period I; Q_i=order quantity at period i.

Where no order is placed when the inventory level is greater than the alarming level (the reorder level); and an

order quantity of (S_T - L_i) is placed when inventory level is at or below the alarming level.

Three parameters define this policy: S_T, R, and T. the optimal values for S_T, R, and T are determined to minimize total inventory costs. We can calculate the general inventory quantity for periodic review as the following equation:

Invtoday (t+1)=Invtoday(t)-Demand (t)+Order (t)

Where *Invtoday (t+1)*: Next inventory quantity; *Invtoday (t)*: Current inventory quantity; *Order (t):* Current order quantity.

We can summarize the total cost of a general inventory model for periodic review as a function of its principal component in the following manner:

(Total inventory cost)=(purchasing cost)+(setup cost)+(holding cost)+(shortage cost)

In practice, however, that an inventory model need not include all four types of costs, either because some of the costs are negligible or will render the total cost function too complex for mathematical analysis, so we can delete a cost component only if its effect on the total cst model is negligible. For this paper, we just calculate the total inventory costs using the following cost equation:

$$C_T = C_h + C_o + C_s \qquad (2)$$

Where C_T: Total inventory cost; C_h: Total holding cost; C_o: Total ordering cost; C_s: Total shortage cost

2.2 Fuzzy methodology in a supply chain

A fuzzy set is characterized by fuzzy boundaries: unlike crisp sets in which a given element does or does not belong to a given set, in fuzzy sets each element belongs to a set with a certain membership degree. The function that returns the membership degree of each fuzzy set element is called membership function. In this paper, triangular membership functions have been adopted because they are considered the most suitable form to model market demand, inventory level, and ordering quantity. According to Zadeh [4], the membership function should be designed by the expert, i.e. managers, that gives the estimate. This is quite simpler than asking managers to design a probability function. To model any variable, say market demand, by a triangular membership function managers only need to estimates the values that do or do not belong to its domain (fuzzy set).

As stated previously, a classical (s, S) policy is: at the moment when the inventory drops below level s, order units up to level S. this implies there is always a fixed number Q=S-s units ordered. But for desecrate cases, if the inventory level is lower than s, place an order. It' s possible that the inventory level drops to zero before the new order arrives. We present an improvement approach inventory control for the supply chain using fuzzy control system. With this approach, several aspects of the system were handled in the same manner as in the crisp runs. The number of orders, number of stockouts, and average inventory were all calculated in the same way. The fuzzification occurred when an order was being placed.

When the inventory level was dropped below a point (inventory level), the inventory level and the current period's demand were given membership function values. The membership values were based on the logic described below. To maintain flexibility in the model, all the parameters indicated below are in terms of the model's inputs. This allows the model to adapt to different cases. The core advantages of a fuzzy controller are robustness under uncertainty, expert experiments considered, and inaccurate information [5-7].

Fuzzification. We classify the demand, inventory level and order quantity into the five degrees between the least and the highest: L1(Least), L2, L3, L4 and L5(Highest). The corresponding membership function is shown in Figure 1.

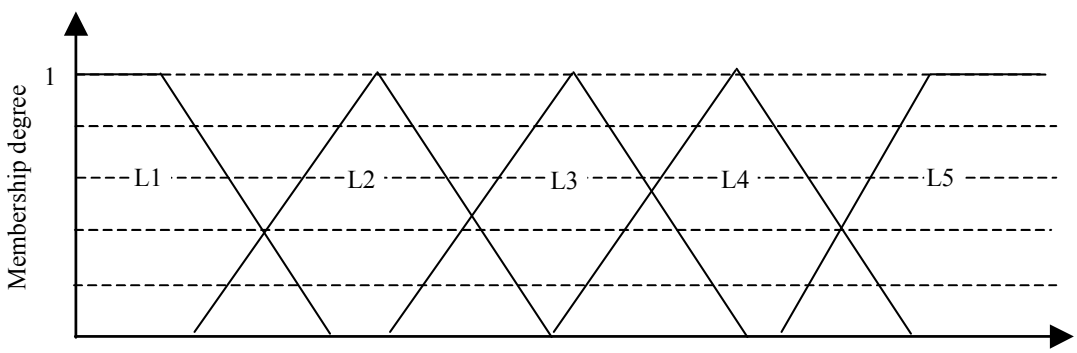


Figure 1 Membership function of demand, inventory and ordering quantity

The values on the x-axis represent the different values for different variables. The scalar factor could be changed easily. Varying the value of this scaling unite can be tuning the membership function to make the performance better.

Fuzzy Control Rules. The fuzzy control rules are based on the experiences on inventory ordering policy. The relationship between demand amounts, current inventory level and the ordering quantity to be ordered is summarized in Table 1. For example, this table should be read "If demand is L5 (the Highest), and the inventory level is L3 (Medium), then order High amount of items."

Fuzzy Operators. Selection of the intersection and union operators was constrained by the need to reduce the frequency of ordering. If the operators were compensatory, that could cause the resulting purchase amounts to be further from the extremes, eventually causing an increase in the frequency of orders. This would increase the cost of the system. Another factor of interest was computing cost; complex operators drastically increase the number of computations necessary to run the system. For these reasons, the maximum and minimum operations were selected as the union and intersection operators.

Table 1 Demand, inventory level and order quantity relations

Order Demand / Quantity / InventoryLevel	L1(Least)	L2	L3	L4	L5(Highest)
L1(Least)	L3	L4	L4	L5	L5
L2	L2	L3	L4	L5	L5
L3	L2	L3	L4	L4	L5
L4	L1	L2	L3	L4	L5
L5(Highest)	L1	L1	L2	L3	L4

3 Case study

To compare the performance and implementation of the fuzzy and classical inventory control systems. We apply two inventory control approaches to two cases, which are the single stage inventory and the multiple stages inventory in supply chain. Our supply chain looks like the following:

We are a distributor for a certain product. The demand for our product comes form a single source, a retailer (for case 1) or multiple sources, three retailers (for case 2). We get supplied the product from a manufacturer, whom we order from whenever our inventory is at a certain level. For the multiple stage inventory problem, the manufacturer in turn gets supplied from a vendor.

We make the assumptions for the cases:

The demand has a uniform distribution. Demand is a function of t, which is Demand (t).

Average daily demand (D) from past data.

The manufacturer (for case 1) or the vendor (for case 2) is a reliable supplier.

About delay time: For the singer stage inventory, the delay time for us to receive an order after the order had been placed with the manufacturer is some days (delay). For the multiple stages inventory, the delay time to receive an order after the order had been placed is some days (delay).

Initialize inventory: For the singer stage inventory, to cover the first delay days, initialize inventory by: inv=D*delay. For the multiple stages inventory, initialize inventory based on D, delay, and number of stages from reliable source.

Simulation

Because the order is placed in the beginning of the period, we don't know the exact demand, we create the demand based on its distribution functions. In Matlab, we use rand() to create random values for demand at every period and store the values in an array. The simulation could be based on the cost evaluation equation used the following data:

$C_T=C_h+C_o+C_s=h \times Q+K \times N+p \times$ (numbers of Stockouts).

Where:

N=number of times of ordering

Q=average inventory quantity

K=setup cost for placing an order *K*=8000$

h=holding cost per unit inventory per unit time, *h*=0.35$/unit/period

p=shortage, or stockout cost, *p*=4$/unit stockout

Table 2 Simulation for singer stage inventory

series	Fuzzy Inventory Controller			Classical Inventory controller		
	Total cost	Order times	Average inventory	Total cost	Order times	Average inventory
1	4.1430e+005	105	3.7048e+003	4.6553e+005	215	3.4380e+003
2	4.0155e+005	93	3.8863e+003	4.5836e+005	210	3.3296e+003
3	3.9354e+004	88	3.9568e+003	4.4934e+005	185	3.5642e+003

Table 3 Simulation for multiple stage inventory

series	Fuzzy Inventory Controller			Classical Inventory controller		
	Total cost	Order times	Average inventory	Total cost	Order times	Average inventory
1	5.5840e+006	810	80.2435	5.8703e+006	873	55.4867
2	5.3821e+006	706	78.6943	5.6440e+006	824	54.3846
3	5.0632e+006	687	79.8654	5.4846e+006	807	55.6789

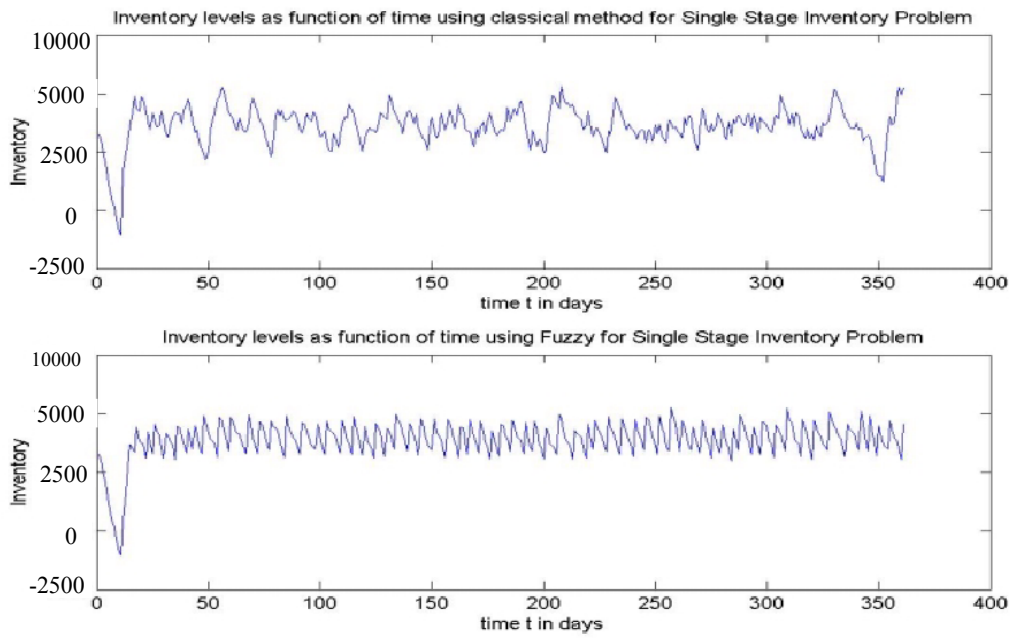


Figure 2 Case 1: Inventory curve for using fuzzy and classical approach

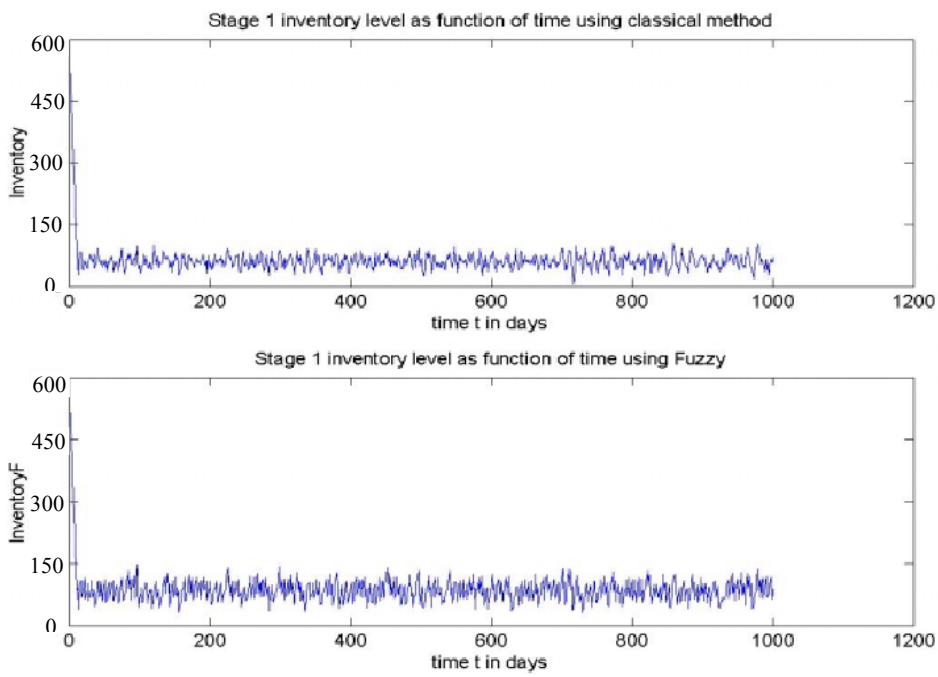


Figure 3 Case 2: Inventory curve of stage1 using fuzzy and classical approach

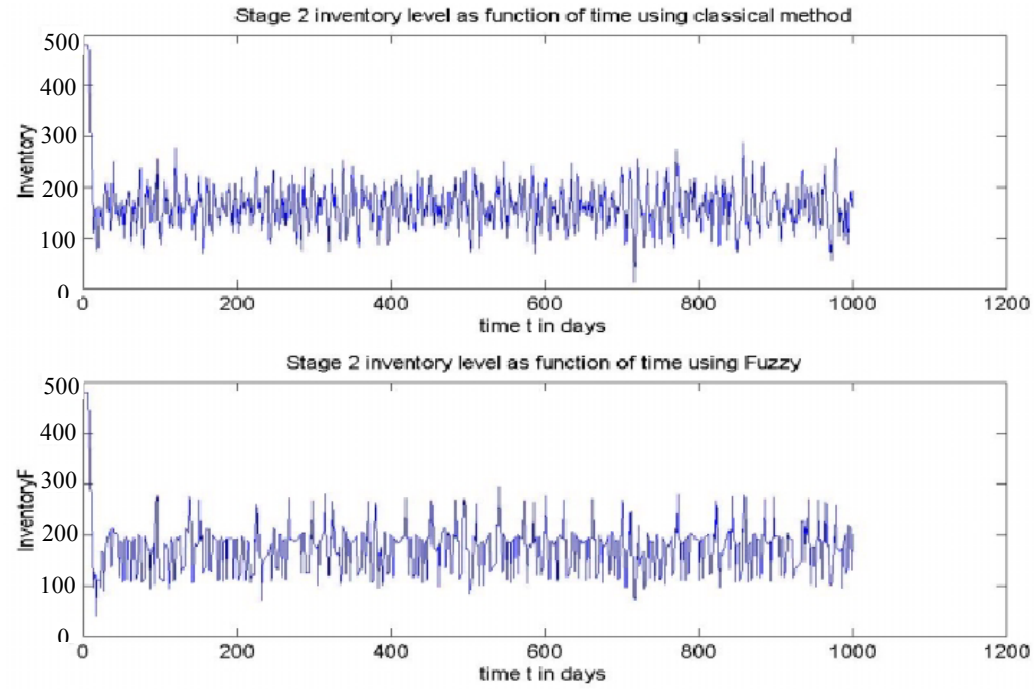


Figure 4 Case 2: Inventory curve of stage2 using fuzzy and classical approach

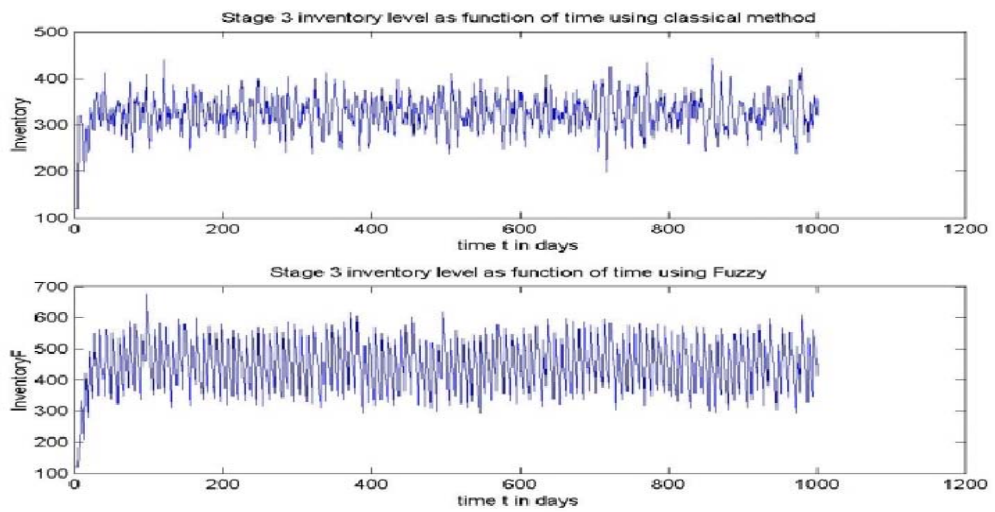


Figure 5 Case 2: Inventory curve of stage 3 using fuzzy and classical approach

To compare the fuzzy and classical inventory control systems based the above cases. This was done by comparing the total cost, excluding ordering cost, of each system using identical demand streams for some periods for two cases. Based on the uniform demand distribution, we ran simulation of 360 periods to evaluate the performance for the two cases. We started with the control rules as the Table 1 and fuzzy operator as above. See Figure 2-5, Tables 2 and 3 for explicit results. Table 2 and Table 3 compare the simulation between the fuzzy inventory system and the classical inventory system. Clearly, for the fuzzy inventory system, but its total cost is much lower, for instance, compared with 465530, 458360 and 449340 of total cost in classical inventory system, there are 414300, 401550 and 39354 for fuzzy inventory system in 3 series

respectively, because its number of times of order is much less than using the classical inventory system even though its average inventory is higher than the classical approach. The detail is shown in Table 2 that order times in classical way are all almost 200, while for fuzzy approach nearly below 100, the same things occur in multiple stage inventory. Additionally, its inventory level is much stable than using the classical inventory system. The simulation has demonstrated the fuzzy logical approach can successfully tune a fuzzy controller using tuning membership function. The fuzzy controller can cope with the imprecise information, its matrix can be improved by the experts based on the experiences.

4 Conclusion

In this paper we have provided a simple fuzzy logic approach for inventory control for the supply chain environment subject to uniform distribution demand. Our approach explicitly accounts for the differences from the classical approach in design based on our simulation studies we see that the approach yields benefits for different levels of variability in the in a supply chain environment, and fuzzy logical shows a powerful tool for inventory control n supply chain. However, fuzzy rules must be chosen carefully. Improperly chosen rules and parameters can result in worse performance than the classical systems. The proper-tuned fuzzy control system is superior to its classical control system when intelligently implemented. Therefore, the fuzzy logic approach is fairly robust. This is a significant attribute of our approach as it increases the viability of its application to inventory problems.

References

[1] H. L. Lee, C. Billington, Managing supply chain inventory: Pitfalls and Opportunities. Sloan Management Review, vol. 33, 1992, pp.65-73

[2] M. Christopher, Logistic and Supply Chain Management. Pitman Pub., London, 1992

[3] F. Chen, Stationary policies in multi-echelon inventory systems with deterministic demand and backlogging. Operations Research,vol.46, 1998, pp.S26-S34

[4] L. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes. IEEE Transactions on Systems, Man, and Cybernetics 15, 1975, pp. 28-44

[5] H. Giannaoccaro, P. Pontrandolfo, B. Scozzi, A fuzzy echelon approach for inventory management in supply chains. European Journal of Operational Research, vol.149, 2003, pp.185-196

[6] H. J. Zimmermann, Fuzzy Sets Theory--and Its Applications. Kluwer-Nijo. Pub. Boston, 1988

[7] G.Y. Xong, H. Koivisto, Research on Fuzzy Inventory Control under Supply Chain Management Environment. P.M.A. Sloot et al. (Eds.): ICCS,2003, pp.907-916

[8] R. Brown, Smoothing, forecasting, and prediction of discrete time series, Prentice-Hall, 1962

[9] X. Zhao, J. Xie, Improving the supply chain performance: use of forecasting models versus early order commitments, Int J Prod Res, 2001, 39, pp.3923-3939

[10] S. Axsăter, Modelling emergency lateral transshipments in inventory system, Management Science, 1990, 36（11）, pp.1329-1338

A Studying the Conceptual Model and Critical Technology Based on Enrterprises Finance System *

Ming Ni

College of Economics and Management,East China Jiaotong University, Nanchang, Jiangxi, 330013, China

Email: xyznm96@163.com

Abstract

The tradition financial accounting pattern has played an important roles in enterprises financial field last century, the tradition accounting pattern has faced great challenge , with the development of computer, network technology and the change of manufacturing pattern, etc.. Constructing a practical and feasible financial system (FS) conceptual model is the premise of improving the competence of FS successfully. In view of such two important platforms as technology platform and corporate culture platform that are provided by Enterprises Informatization (EI), the conceptual model of FS for SDN node enterprise, which includes two platforms and a core part, a financial information operation circulation, is designed in the article. The whole operation process of FS conceptual model is described in terms of technology, tool and corporate culture. SDN node enterprise can refer to the model when running into the FS plight.

Keywords：Financial Accounting Pattern; Conceptual Model; Critical Technology; SDN

1 Introduction

Enterprises generally measure and control accounting information by the tradition financial accounting pattern when exert the functions of financial system (FS).The tradition financial accounting pattern has faced a lot of challenges, though it has played an important roles in the past years.

1) Finance-oriented and post event information which were measured by a single monetary unit, which can't provide multi-monetary, global, non-financial and predictable information.

2) The information which has not been graded can't provide different information to all kinds of users. Large of the same financial reports make non-financial stuffs confused.

3) The way to dealt with accounting data (keep an account of borrow and lend) is different from other ways of enterprise's systems. It causes accounting information can't used by other systems, even creates data redundancy and difference.

4) The tradition financial accounting neglects the differences of various information users' demands and applications. The common format of tradition financial accounting report can't satisfy the individual demands of SDN node enterprise, and makes it worse to discriminate information.

5) The tradition financial accounting can't satisfy the timely demands of various cost information in SDN node enterprise. It is hard to supply timely financial information to node enterprise since the tradition accounting report are generally published every month, season, half year or year.

6) The tradition accounting report just supplies a kind of monetary information, being lack of non-monetary decisive information, such as the time of events occurred, participants.

* The research is supported by such projects: the National Natural Science Foundation, China supports the research project (70472075), the project of science and technology for the department of education of Jiangxi Province (Grant No. 2007-183,GJJ08244), the project of the Jiangxi Province Natural Science Foundation(2007GZS0898) , and the National High-Tech. R&D Program for CIMS, China (2002AA414310).

7) The diversification demands of products bring more challenges to measure the cost, and the tradition accounting can't correctly measure the cost of diversified products. However, the tradition accounting can't break through because of the enterprise isn't open enough.

2 The Crucial Technology of FS for SDN Node Enterprise

SDN node enterprise has a unique advantage of Finance informatization because of its high level of enterprise informatization. A technology platform, one of the results of the success of enterprise informatization, is produced. It includes: computer, network technology and intelligence deduce and decisive technology. Finance informatization must rely on the two technologies. The processes of enterprise informatization include Finance Information Acquisition, FIA; Finance Information Settling, FIS; Finance Information Storage, FIS; Finance Information Reengineering, FIR; Finance Information Sharing, FIS; Finance Information Application, FIA. Informatization technologies and tools which are needed in the six processes are as follows.

1) Technologies and tools in Finance Information Acquisition, FIA. This process needs Internet Technology、Search Engine、Enterprises Finance Information Portal(EFIP) and Finance Information Map(FIM). Internet and Web Technology makes enterprises acquire abundant finance information. Search Engine provides a more convenient way to find required finance information. EFIP will combine two functions of Enterprises Information Portal, EIP and Enterprises Collaboration Portal, ECP. EIP is a portal of data and information of enterprise structuring. It will provide an unified system to extract business intelligence and finance information from databases and information bases. It will realize a pattern which converts general information to finance information by only one window frame, and build up longitudinal relationship between users and information. ECP build a finance information solution of REA pattern by the horizontal contacts among terminals. It facilitates the contacts and collaborations among different terminals by means of meeting, chatting, discussion, work flow and notice, etc. This kind of finance information tool based on EFIP includes Plumtree, Intraspect, Viador E-portal, etc. Users can arrange personal desktop according their habit to acquire finance information conveniently. FIM provides finance information users a technology to find the masters of these finance information, which plays an important role in FIA.

2) Technologies and tools in Finance Information Settling, FIS. This process needs Files Conversation Technology (FCT), Integration and Conversation Technology (ICT), Finance Information Expression (FIE), etc. By the communication among enterprise staff and experts, ICT applies ABC method to analyze and reengineer operation processes, which makes non-finance information and substandard files be normal, explicit to store and search them conveniently. Some non-finance explicit information can normalize the relative cases or scenarios, which match the technology of process analysis. FCT will convert external finance information to finance information based on REA pattern, which satisfies different users' demands, and then standardize them to store and searche conveniently by a unified standard. This technology directly converts finance information to finance information files, but not as same as the technology which converts non-finance information, so it matches REA conversation technology. FIE is a method to structure finance information. At present, there are nine kinds of finance information express technologies: logical expression, characteristic expression, generation rule expression, semantic web expression, frame expression, object-oriented expression, status space expression, process expression, and qualitative physical model expression, etc. These method of expression will be inserted in software tools, such as LivelinF of Opentext, Infoplace Document Organizer of Infoplace, Ishare of Ibridge Software.

3) Technologies and tools in Finance Information Storage, FIS. This process needs Finance Information Warehouse (FIW), Secondary Storage (SS), Operation Finance Information Storage (OFIS), Finance Information

Mart (FIM), etc. FIW is a subject-oriented, integrated technology, and reflecting the historical changes and relative steady sets of finance information and detailed finance information. FIW and SS are main storages to store larges of finance information in SDN node enterprise. SS aims to enlarge the capacity of finance information warehouse. There are original finance information, image warehouse, data warehouse, instance warehouse, situation warehouse, rule warehouse, script warehouse, object warehouse, process warehouse and model warehouse, etc. OFIS is a subject-oriented、integrated technology which reflects present and variable detailed finance information. The difference between OFIS and FIW is that OFIS supports finance information user immediately. FIM is designed for different users which is sub-technology of FIW. Now, main FIW manufacturers used to develop database, such as Oracle, Sybase IBM Essbase/DB2, and Informix, etc.

4) Technologies and tools in Finance Information Application, FIA. FIA is performed in two fields, which includes forming all kinds of finance reports, evaluating operation processes and providing decisive information, and improving enterprise culture. Every field needs different technologies and tools. Technologies and tool to form all kinds of finance reports, evaluate operation processes and provide decisive information include finance software, DDS, BI, Simulation Technology (ST), Finance Information Discovery (FID), Application Based on Work Flow, etc. Simulating work flows by ST will reengineer process in every node enterprises. Application Based on Work Flow not only provides a routine which arranges files, but also confirms files paths by analyzing the files contents, and applying Yellow Page system, which will change work flow application software to satisfy finance information users demands. Both DSS and BI are technologies and tools to form scientific decision. FID is a process technology of exploring new finance information from finance information warehouse. The technologies and tools of improving enterprise's culture is communication platform. Communication platform is a virtual space to foster harmonious and tacit atmosphere among staffs. It provides a chance to communicate as well as collaborate with other staffs, where non-finance information are used. The common communication platform includes BBS, Netmeeting, and e-learning, etc. The manufacturers which provide FIA include Lotus Notes Domino, and Microsoft Exchange Server, etc.

5) Technologies and tools in Finance Information Reengineering, FIR. This process needs Selection and Filtration Technology (SFT), Meta Finance Information Technology (MFIT), and Files Management Technology (FMT), etc. SFT aims to improve safety of FIA in enterprise. Different grade users have different degrees of acquiring finance information in finance information warehouse, which is a mapped condition from FW to FM. MFIT manages finance information and files in finance information warehouse. Meta finance information, a experienced summary graph, which denotes the data of finance information. According to meta finance information standard, much application software will share the finance information in finance information warehouse while facilitating maintenance of finance information warehouse. FMT will deliver some unvisited finance information for a long time to secondary storage. At present, manufacturers which provide technologies and tools in FIR include Smar Team provided by Smar Team and Dassault, which automatically integrates with OA of Microsoft.

6) Technologies and tools in Finance Information Sharing, FIS. This process needs Training System(TS), Learning System(LS), Virtual Work Team(VWT), Media and Electronic Publish(MEP), Community Of Practice(COF), Finance Information Delivery(FID), and E-mail, etc. TS has been a main method to share finance information since it develops from tutor guidance, video teaching and "training based on Computer" by CD in the classroom to the stages of "Network training" and "Network performance support system" by TV/Telephone Meeting technology, and virtual university, etc. VWT provides a collaborative situation to share finance information. FID forwardly delivers finance information to users. MEP, LS and E-mail are the tools of sharing finance information. COP is brought

out for the research of Xerox community's members by Brown and Duguid. COP also provides a virtual space for non-finance information share in enterprise. At present, manufacturers which provide technologies and tools in FIS include Campus of CourseInfo and BlacF Board.

3 The Conceptual Model of FS and the Operation Process For SDN Node Enterprise

Enterprises which want to improve the capability of FS in virtue of the above critical techniques, first of all, should build the conceptual model to guild enterprises to implement the technique transform of FS（see Fig.1.）

In Fig.1, the two platforms produced by enterprises informatization (EI), exist the interaction and influence of relationship for each other, expressed by in Fig.1. The six processes are the whole process of EI, and exist dynamic relationship among them, just correspond the six regions in Fig.1. VI region, Finance Information Application (FIA) as the center, from the I region of FIA (Acquisition) to V region of FIS(Sharing) is a circle, and then access to I region to start a new circle, namely, I FIA (Acquisition) →Ⅱ FIS(Settling)→ⅢFIS(Storage) →Ⅳ FIR →Ⅴ FIS(Sharing)→ⅥFIA… just like this to cycle, denotes the transition among the six processes by symbol, which ensures the information of supply to Ⅵ FIA(Application) being up to date and. And then, illustrating the operation process of the finance informatization solve project at the angle of the finance informatization whole process.

1) FIA (Acquisition) Ⅰ. In field Ⅰ, SDN node enterprise can get various kinds of original finance information as well as non-finance information through learning and communication by means of finance information acquisition technology and tool (EIP,ECP),. Through individual learning and team learning, enterprise learning can raise partners learning each other. Whichever hierarchy learning, node enterprise should have foundation corporate culture supported by sign ② in Fig.1. Various kinds of learning participant who need finance information, can acquire non-finance information from document materials, practice, station in turn, training, team corporation and custom communication, and so on. Beside technology platform, partners learning also need reasonable non-finance information value interest mechanism and non-finance information sharing mechanism, etc. culture platform to support. Enterprises should settle the getting materials, namely access to Ⅱ region.

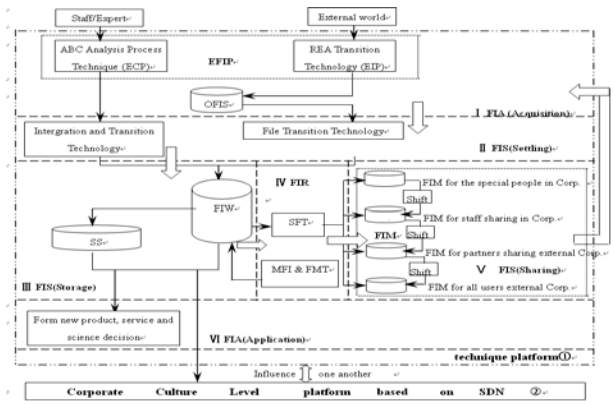


Figure1 The conceptual model of finance system for SDN node enterprlse

2) FIS (Settling) Ⅱ. In field Ⅱ, SDN node enterprise mainly settles finance information from field Ⅰ. SDN node enterprise can make the finance information coding and standardization and then store the information in finance information warehouse by means of handcraft or FIS technology and tool. It is an important basic work which affects finance information acquisition directly. After settled, the finance information can enter to Ⅲ region.

3) FIS (Storage) Ⅲ. In field Ⅲ, SDN node enterprise puts the settled finance information to finance information storage. It is easy to carry out finance information storage. But the non-finance information should transform case or scenario to store in FIW in order to the financial users to extract the non-finance information by reading and analysis. During the whole finance information storage process, node enterprise should pay attention to the security, such as the storage site, the mature and the dependability of the storage technology and so on. The finance information in FIW should be maintenance, namely enter to Ⅳ region.

4) FIR Ⅳ. In field Ⅳ, SDN node enterprise reengineers the finance information of FIW by means of the FIR technology and tool. This process includes confirm the sharing level of FIW among customers and finance information maintenance. There are four kinds of FIW users, namely the special people in enterprises, staff sharing in enterprises, partners sharing external enterprises, all users external enterprises. The four kinds of users should have different level to share FIW to improve its security. In addition, there is little access finance information in FIW, which should be transformed to SS to improve the usage ratio of the finance information of FIW. The reengineered information can be shared by customers, namely enter to Ⅴ region.

5) FIS (Sharing) Ⅴ. In field Ⅴ, the reengineered finance information has been confirmed the sharing levels, which enters four kinds of FIM. First of all, the FIM for the special people in enterprises, this kind of finance information has the highest secrecy, mainly relates the fate and the development of the enterprises, and this kind of information can not be shared generally, but after a period time, can be transformed to second FIM. The second kind, the FIM for staff sharing in enterprises, this kind of finance information has higher secrecy, but is opened for staff in enterprises. The enterprises should improve the sharing range of the finance information, mainly including overcoming sharing obstacle factors, supplying stimulation measure and the effective performance measure system and so on. When this kind of finance information reaches some period time, it can enter third FIM. The third kind, the FIM for partners sharing external enterprises has a general secrecy, but just only be opened among partners in order to improve the competence of partners in SDN situation. When this kind of finance information reaches some period time, it can enter forth FIM. The forth kind, the FIM for all users external enterprises is opened completely, in order to let more people know by means of Internet technology and push technology. After the cycle of finance information operation finishes, will enter the next cycle, namely enter field again. However, during each finance information operation cycle, finance information application is the core position, and the only purpose of the operation cycle, which runs through the five processes and forms interact relationships among processes.

6) FIA (Application) Ⅵ. In field Ⅵ, the application field includes two aspects, namely forming varied kinds of financial reports, valuating each work process and supplying information for decision-making and improving corporate culture. The first aspect mainly is finance information or non-information application, and the second aspect is non-finance information application. Because applying finance information to form various financial reports or estimating each work process or forming more science decision, which is to face SDN node enterprises and bring “Spill-over Effect” in the finance information sharing process. Such non-finance information as knowledge capability, human recourse, etc. can improve corporate culture by the means of node enterprises communications, which are the reliable warrant for forming enterprises core competence and blurring the enterprises boundary. Therefore, enterprises should keep a balance between finance information application and non-finance information application. Enterprises can not give up the non-finance information application just because the finance information has a rapid effective. In the finance information application process, it has a relationship with other five finance information operation processes, which indicates that the finance information application is in the core position of the finance information operation.

4 Discussion and conclusion

Enterprises integrate the six processes of the finance system operation by means of the above key technology, forming a cycle which is composed by the finance information application as the core, and other five finance information operation processes, and the finance information application forms a interplay relationship which helps to improve enterprises finance information system capability. The improving process is little by little. When SDN node enterprise starts to

develop just, it is not suited to solve the finance system plight, but when it reaches some informatization level, the improving process can be carried out.

References

[1] Wei Peiwen.Informatization and accountant mode reform. Beijing: China finance & economy press,2002

[2] Xiao Hanzhong.Compared with Varied kinds of accountant mode. Beijing business university transaction, No.2, 1996, pp.22-24

[3] Xu Fuxuan.The elementary investigation of the Multifunctional and open enterprises SDN. Forecast, Vol.21, No.6, 2001, pp.19-22

[4] Davinel A.M, Robert B.S.et al.Functional interdependence and product similarity based on customer needs. Engineering Design, No.11, 1999, pp. 1-19

[5] William E. McCarthy. The REA Accounting Model: A Generalized Framework for Accounting System in a Shared Data Environment . The Accounting Review. Vol.57,No.3, 1982, pp. 554-578

[6] Li P.G.. Model and analysis of manufacturing system performance---theory and method.Wuhan: Press of University of Middle China for Science and Technology,1998

[7] Yuan H.B., Huang X.Y.. A Conceptual Framework for Manufacturing Flexibility and Its Measurement. HIGH TECHNOLOGY LETTERS, No.5,2000,pp.50-53

[8] Ni Ming; Xu Fu-yuan. Extended AHP Method in Virtue of Rough Sets Theory Based on GA. Journal of Applied Sciences, Vol.23,No.5, 2005,pp.517-521

[9] Danna E. J, Jasperson J. Modeling Conversion Process Event . Journal of Information System. Vol.8 No.1. 1994,pp.43-54

[10] James A. Hall. Accounting Information System (3rd Ed.)[EB/OL]. http://www.swcollege.com/acct/hall01/hall01.htm,1998-09-27

Research on the Personalized Retrieval Model for Knowledge Management System Based on Multi-agent

Ziyu Liu[1, 2] Lei Huang[1] Dongyun Xu[1]

1 School of Economics and Management, Beijing Jiaotong University Beijing 100044, China
Email: purpleyuliu@sohu.com

2 Institute of Information Science & Engineering, Hebei University of Science & Technology
Shijiazhuang, 050018, China

Abstract

There are several challenges in supporting users' retrieval activities in knowledge management systems. For example, users can not receive immediate notification when the new knowledge is adding into knowledge management system. To solve these problems we proposed a personalized retrieval model for knowledge management system. On the basis of sufficiently analyzing the necessity and feasibility for using the multi-agent technology to develop the personalized retrieval model for knowledge management system, this paper proposes a three-layer structure of personalized retrieval model for knowledge management system based on multi-agent which includes client layer, personalized logic layer and data service layer with user feature base, method base and knowledge base. The prototype application is implemented using Eclipse technology, and agents are deployed on the FIPA-OS, we used J2EE to develop the knowledge based on agents' reasoning.

Keywords: Multi-agent; Knowledge management System; Personalized Retrieval Model; Three-layer Structure

1 Introduction

At the present more and more units and individuals has recognized the significance of knowledge management, and begun to build a platform for achieving knowledge management-knowledge management system. In order to retrieval and access the knowledge in the knowledge management system, we need a search engine to accomplish this function.

However, there are several challenges in supporting users' retrieval activities in knowledge management systems. First, the collection of a knowledge management system constantly changes with knowledge being added, edited, or removed by teachers. The knowledge of knowledge management system will be increasing with the pass of time, and the existing way of knowledge management has its shortcoming because that it gives barriers for users to obtain and use knowledge. Second, the users can not receive immediate notification when the new knowledge is adding into knowledge management system [1].

Against the foregoing it need introduce agent technology into the model design of knowledge management system [2] [3]. That can achieve the rapid transmission of knowledge, and can also handle and provide personalized information services for users. And that also can better provide knowledge support and service for knowledge acquisition. The multi-agent technology's rapid development provides the possibility of solving these problems.

Therefore, combining multi-agent technology with knowledge retrieval model will bring a new idea for the personalized retrieval of knowledge management system. On the basis of these, this paper proposes a three-layer structure of the personalized retrieval model for knowledge management system based on multi-agent. This system is designed to run on the web.

The paper is organized as follows. Section 2 is devoted to introduce agent and MAS. In Section 3, we

establish the personalized retrieval model frame for knowledge management system based on multi-agent technology. Section 4 gives the prototype implementation method of the personalized retrieval system and introduces the information flow among agents of the system. Finally conclusions are given in Section 5.

2 Agent And MAS Introduction

The concept of agent was started from John McCarthy in the mid of 1950's and established by Oliver G.Selfridge several years later. In the early days, many researchers have been studied about agent in boundary of AI. Since 80's the agent have widely applied [4]. Although lack of a universal definition of agents, there is a general agreement that an agent is a reusable component that exhibits a combination of the six characteristics: Autonomous, Adaptable, Mobile, Knowledgeable, Collaborative and Persistence [5] [6]. So in software system, agent means software component that has inference capability, and can interacts autonomously as a surrogate for its user with its environment and other agents to achieve the predefined goal, and reacts to changes in the environment.

Agents can perform tasks individually or work cooperatively in teams. Agents may also migrate between machines by virtue of being a complete unit of execution that can make its own decisions. Agents communicate with each other and with other machines via messages. Agents can be enhanced to learn from their past executions and adapt to perform better in the similar situations.

Inspired by distributed artificial intelligence, a MAS (Multi-Agent System) consists of autonomous, generally heterogeneous and potentially independent agents which work together to solve special problems. As described by Brennan, autonomous, cooperative, and scalable are the typical characteristics of MAS that has the following capabilities [7] [8]:

(1) Independent decision-making by individual agent based on its domain knowledge, local and global conditions.

(2) Interacting with other agents and humans for effective negotiation, cooperation and coordination.

(3) Perceiving changes in their environment and acting as a consequence.

(4) Taking initiative to reach certain objectives.

These distinctive characteristics of MAS can undoubtedly facilitate the personalized retrieval systems. It is appropriate to adopt the MAS technology in the personalized retrieval model of knowledge management system.

3 Proposed Solution

3.1 Framework

The framework extends the personalized retrieval model for knowledge management system by incorporating multi-agents entities as shown in Figure 1. The framework is composed of:

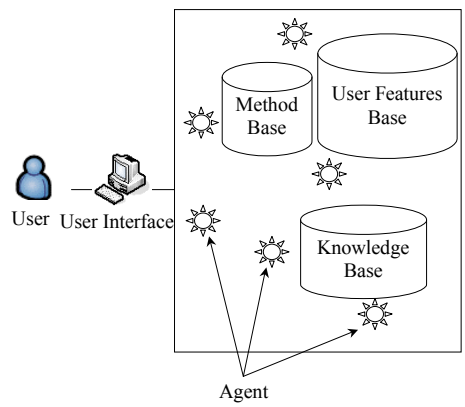


Figure 1 The personalized retrieval framework

(1) Multi-agents: as we want to build the personalized retrieval model in knowledge management system, we need an agent-based system. At the same time, the knowledge of the knowledge management system has the feature of on-line and dynamic. The system should inform users immediately when the new knowledge is adding into knowledge management system. Various agent types are used to realize the system goals. The agents' roles, capabilities, intelligence, autonomy, cooperation, communication language and protocol as well as shared ontology are considered.

(2) Knowledge based module: for this module an agent system approach is chosen to implement the mapping of the knowledge. This module stores knowledge and incorporates mathematical solvers.

(3) Method Base System: this module stores algorithms, rules and reasoning model etc. that are used in knowledge retrieval.

(4) User features base: It main stores the users' information of interest feature etc., and its content is dynamic.

3.2 System architecture

We designed a three-layer structure of the personalized retrieval model for knowledge management system based on multi-agent, which is made up of client layer, personalized logic layer and data service layer, shown in Figure 2. Client layer is responsible for user interaction and the result of retrieval presentation. It submits request from user to middle model for knowledge layer which executes personalized logic and shows result to user or finishes user's operation. Middle layer receives request from client layer, then submits retrieval task to retrieval agent to be dealt with, finally returns result to applicant. Knowledge building agent is responsible for the building of knowledge base of the middle layer. In this way, the personalized retrieval system with three layer architecture can separate personalized and retrieval logic from client to one or more middle layers. It adopts load balance, dynamic scale and standard interface technology to integrate client and server tightly. It is characterized by thin client, maintenance and extension with ease. B/S three-tire architecture improves the ability to cooperate with others.

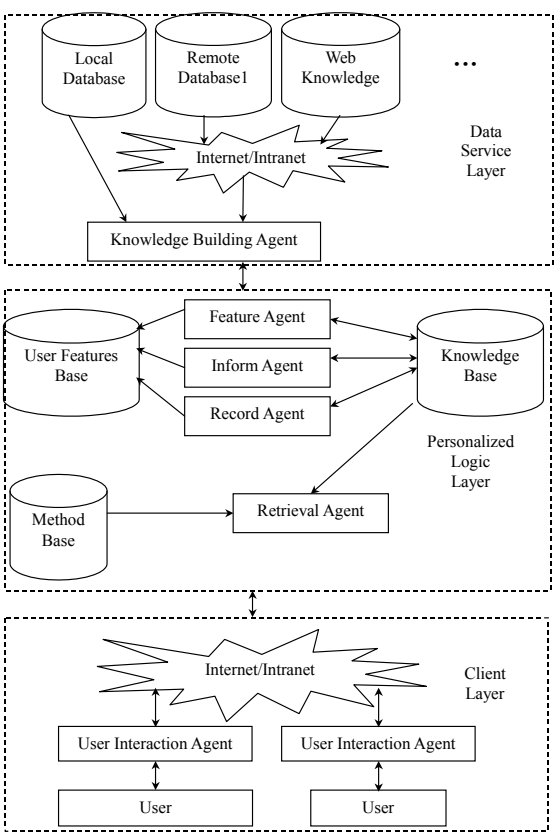


Figure 2 Three-layer structure of the personalized retrieval management system based on multi-agent

3.3 Application of agent in the personalized retrieval model

As for the personalized retrieval model of knowledge management system, agent is a key technology to improve its intelligence. The different assignments brought forward by different users, distributed to different Agents, are accomplished by reasoning. In the personalized retrieval model there are six agents including knowledge building agent, feature agent, inform agent, record agent, retrieval agent and user interaction agent cooperating to accomplish work. The communications between various agents need a language that has semantic feature, so here we use OWL that is a kind of description language for ontology to achieve communication and collaboration between agents.

(1) Knowledge building agent

Knowledge building agent is responsible for the mapping from knowledge sources to knowledge base, and organizes the knowledge of knowledge management system.

(2) Feature agent, inform agent and record agent

When the new knowledge is added into knowledge base record agent will automatically classify and organize the new knowledge, and then inform agent will be activated, and it will submit the new knowledge to appropriate users through e-mail according to the

demand of users in user feature agents. Feature agent is used to record the interest and feature of users. Inform agent is used to trigger and notify users when new knowledge are adding into the system, and the users who are interested in it can access such knowledge [1].

(3) Retrieval agent

Retrieval agent executes search tasks. The structure of retrieval agent is shown in Figure 3 [9]. The retrieval agent has knowledge about the task domain, as well as the capabilities of other agents. It obtains the necessary methods from method base according to the different retrieval task, and it also may seek the services of a group of agents that work cooperatively and synthesize the integrated result. Then it searches the corresponding knowledge from the knowledge base. Finally it submits the results to user interaction agent.

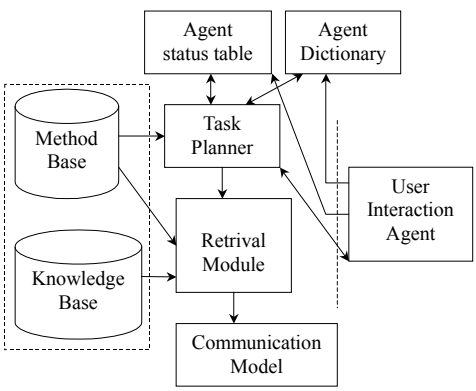


Figure 3 Retrieval agent structure

(4) User interaction agent

User interaction agent interacts with the user in assisting him/her performing the retrieval activities. The user can provide a general description of the problem at hand in terms of high level goals and objectives, and retrieval to be performed. The user interaction agent is responsible for receiving user specifications and delivering results back.

4 System Prototype Implementation And Information Flow Among Agents

4.1 System prototype implementation

The prototype application is based on high-speed railway knowledge. The system was deployed using Eclipse technology [10]. At its core Eclipse is a platform of common tool infrastructure, including: an XML-based plug-in architecture for interoperability; frameworks for structured text and graphic editing; a common debugging infrastructure, and a complete Java development environment [11]. Furthermore, because the Eclipse core is open source, it can be easily modified, and freely downloaded. The use of Eclipse infrastructure significantly reduced the amount of effort required for the development of the system.

The agent building platform used for this project is FIPA-OS [12]. FIPA-OS is an application domain independent agent platform which provides agent services, i.e. message brokering and agent registration, and building blocks for implementation of the agents. According to the FIPA-OS architecture an agent application is implemented by defining application specific agent, task, conversation and message classes based on the generic ones in the platform.

The Reason for choosing FIPA-OS platform is because agents do certainly reside on multiple platforms and exhibit different behaviors. FIPA-OS can interoperate with other heterogeneous FIPA complaint platforms (i.e.: JADE) [13]. By making the agents of the system follow FIPA-OS standards we permit a large span for future enhancement of the application.

The architecture of an agent in our system is based on application specific grouping of the agent building blocks in the FIPA-OS. The resulting overall template of each agent consists of three modules:

(1) Communication module which provides the interface mechanisms for inter-agent communication and collaboration, this interface module is made visible to the other agents of the system and the users.

(2) Operation module which contains methods and computations that implements the process of solving a particular problem, or respond to a request from other agents or users.

(3) Knowledge module that comprehends a collection of domain-specific and domain independent knowledge appropriate to the problem solving.

Both the operation and knowledge module are private to the agent, thus, direct manipulation of the

content of these modules required access privilege.

We used J2EE to develop the knowledge based of the agents' reasoning. J2EE is an expert system shell and scripting language written entirely in Java language. Using J2EE, we were able to build agents that have the capacity to "reason" using knowledge we supply in the form of declarative rules.

We utilized the Ontology Web Language (OWL) for agent communication. When the peer's sender agent sends message(s) to another peer (receiving agent), this message is in a defined OWL format and is transported using the Simple Object Access Protocol (SOAP). The receiver peer's agent parses the request message, processes its detail, and may return to the sender a replay also in OWL format.

4.2 Information flow among agents

Since the cooperative jobs between agents are implemented through the information transfer between each other, it is very important to make clear the content and direction of the information flow among the agents of the system, so that the whole system's working process will be fully understood. The system's main information flow of the personalized retrieval model is shown in Figure 4.

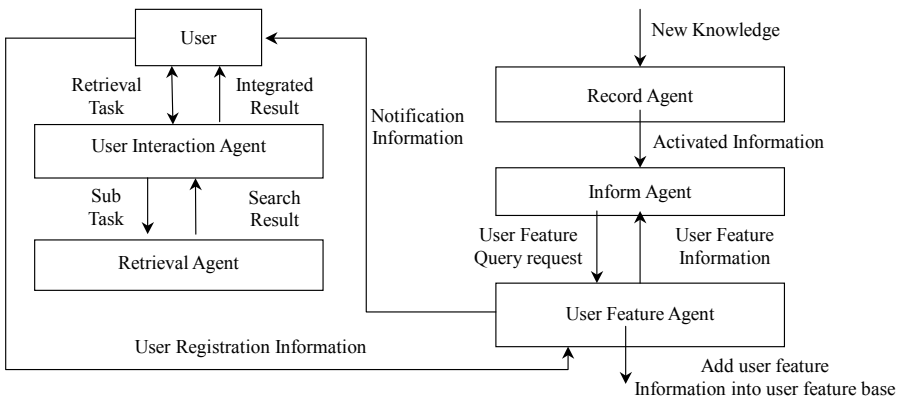


Figure 4 The system's main information flow

(1) User feature agent will add user feature information into user feature base when a new user finished registration in the system.

(2) Record agent will generate activated information when the new knowledge is added into knowledge management system, and then sends it to inform agent.

(3) Inform agent submits the request of querying for user feature to user feature agent.

(4) User feature agent implements query in the user feature base according to the request of inform agent, and returns the results of user feature information to inform agent.

(5) Inform agent notifies the corresponding users by e-mail according to the results of the user feature agent returned.

(6) The user interaction agent accepts the user's retrieval task, and the user interaction agent first decomposes the task into some sub-tasks, second interacts with the user and identifies these sub-tasks, and the last hands them to the retrieval agent.

(7) According to different retrieval sub-task, retrieval agent implements retrieval task. After finishing that, retrieval agent sends the corresponding result to user interaction agent.

(8) User interaction agent incorporates the result and generates the integrated results in a more comprehensible and natural way, then submits it to the user.

5 Conclusions

With the knowledge management system, the agent technology's rapid development, it is possible to use the multi-agent technology to build the personalized retrieval model in knowledge management system. In

this paper, we propose a three-layer structure of the personalized retrieval model for knowledge management system based on multi-agent. The model has three-layer system structure, which includes client layer, personalized logic layer and data service layer with knowledge base, method base and user feature base. The application of implemented prototype system indicates that the proposed solution is a flexible and efficient tool with enormous capabilities for supporting personalized retrieval.

References

[1] Liu Gaoyong, Wang Huiling, "The Support of Agent Technology for Cooperative-learning in Community of Knowledge and Its Realization", Information Studies: Theory & Application, Vol.29, No.3, 2006, pp.365~367

[2] M.Claypool, P.Le, M.Waseda, D.Brown, "Implicit Interest Indicators", Proceedings of the 6th International Conference on Intelligent User Interfaces (ACM), 2001

[3] C C Chen, M C Chen. Y.Sun, "PVA: a sell-adaptive personal view agent", In: proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining 2001, San Francisco, California, 2001

[4] Minjeong Kim, Seungyun Lee, Injae Park, Jintae Kim, and Sooyong Park, "Agent-oriented software modeling", In: Proceedings of Software Engineering Conference (APSEC) Sixth Asia Pacific, Dec, 1999, pp. 318~325

[5] Gilda Pour, "Integrating agent-oriented enterprise software engineering into software engineering curriculum", Frontiers in Education, 32nd Annual, Vol.3, No.11, 2002, pp. S2G-8-S2G-12

[6] Griss, M.L. and Pour, G, "Accelerating Development with Agent Components", Computer, Vol.34, No.5, 2001, pp. 37~43

[7] Shaohong Wu, and Kotak, D, "Agent-based collaborative project management system for distributed manufacturing Systems", Man and Cybernetics of IEEE International Conference, Vol.2, No.10, 2003, pp. 1223~1228

[8] Bin Yu, Munidar P. Singh, "An Agent-Based Approach to Knowledge Management", CIKM'02, Nov. 4-9, McLean, Virginia, USA, 2002, pp. 642~644

[9] Nora Houari and Behrouz Homayoun Far, "Application of Intelligent Agent Technology for Knowledge Management Integration", Proceedings of the Third IEEE International Conference on Cognitive Informatics (ICCI'04), 2004, pp. 1~10

[10] R. Barquin and H. Edelstein, editors, "Building, Using, and managing the Data Warehouse", The Data Warehousing institute, Prentice Hall, 1997

[11] D.Rayside and M. litoiu, "Visualizing Flow Diagrams in WebSphere Studio Using SHriMP Views", Kluwer, 2003

[12] http: //fipa-os.sourceforge.net/

[13] T. Peddireddy, and J.Vidal, "Multiagent Network Security System using FIPA-OS", Proceedings IEEE Southeast Con, 2002

The Research of Students Management Information System Based on J2EE

Wei Dai[1,2] Shengjun Xue[2]

1 Department of economics and management, Huangshi Institution of Technology

2 Department of computer science and technology, Wuhan University of Technology
Wuhan, Hubei Province 430063, China
Email: dweisky@163.com

Abstract

This paper introduces three kinds of popular framework technology of opening source: Struts, spring, Hibernate, we integrate these three kinds of framework technology and apply these technologies to our real web project development. Doing this can reuse code and make your project more maintainable. At the same time it can improve the development efficiency evidently.

Keywords: MVC; Struts; Spring; Hibernate; IoC

1 Introduction

In recent years, with the rapid development of network and Internet, B/S system architecture has become the mainstream of web application system for its convenience and nice expansibility, the architecture has logically divided application into 4 layers: Presentation, Business Logic, Permanence and Date. This multiple system architecture provide the developer with a modularization method which is based on component design and web application developing, the realization technology of each layer is variable, each technology has its advantages, disadvantages and applying scope, how to choose technology of each layer and how to combine these technology to implement a application system is a researching direction that we should lucubrate. In this essay, I adopted Struts architecture to implement Presentation layer, Spring to implement Business Logic layer, Hibernate to implement Permanence layer, construct a web application integration architecture combing the advantages, and analyze its application in the students management system.

2 The Introduction of Structs Architecture

2.1 MVC Mode

MVC mode is the separation of viewer, controller and model, it present nice combination of system arrangement, configuration, function, reduce coupling and make developing and maintenance more flexible.

Mode has been divided into three core parts: model, viewer, controller, each has its own different function.

Model: the main part of application, it presents business date and business logic.

View: The part is associated with the user interface in application, it's the interface that user can see and alternate with.

Controller: Controlling the data display and updating condition of user interface according to the user's input.

2.2 Work Flow

Struts is an architecture based on MVC mode, it is divided into model, view, controller according to MVC, among which controller is implemented by ActionServlet, mode is implemented by Actionand ActionForm,view is implemented by JSPand taglib, when the server starts up, ActionServlet is initialized, read configuration file (Struts-config.xml), when the controller accept the http request, controller compare it

with Action class configured by Struts-config.xml according to the URL address, find the class that deal with this HTTP request, put in the parameter in HTTP to ActionForm, Action performes its method of execute(),transfers JavaBean which deal with business logic to operate date object, after it return a ActionForm object, it encapsulates information of page redirection, act as a point sign. ActionServlet accept this ActionForword object, looking up in Struts-config.xml, gain the JSP address corresponding to this point sign, sent worked HTTP request to this JSP page, and read data from HTTP and display,this is the general work flow of Struts.[1]

3 The Introduction Of Spring Architecture

Spring is an architecture which has settled many problems familiar in the development of J2EE. Spring support a consistent method to manage business object, and encourage the good habit of coding to interface instead of class. The basic of Spring architecture is the Inversion of Control of JavaBean attribute. However, it is only a integrated part: Spring use IoC container as a component, Spring offers the exclusive data accessing abstraction, including simple and effective JDBC architecture, it has improved the efficiency infinitely and reduced the possibility of error. The data accessing component of spring has integrated the solution of Hibernate and O/R Mapping. Spring has also supported the exclusive affair management abstraction, it can manage the technology in each bottom affair, such as JTA or JDBC affair which offer a coherent coding mode. Spring provides an AOP architecture written by Java language, it provides POJOS with announced and other affair management and it can also realize its own aspects according to needs. This architecture is so strong that it makes the application throw away the complexity of EJB, at the same time enjoy the key services interrelated with traditional EJB. Spring also provides the strong and flexible MVC Web architecture which can integrate IoC container. [2][4] The ultimate merit of it is that it can replace business object easily, if only it add dependency with JavaBean attribute and configuration file, then it can replace cooperation object resemble interface easily when in need.

4 The Introduction Of Hibernate Architecture

The aim of putting forward ORM is to solve the problem of how to use an OO way to control and manage relationship database in OO developing flat. Usind O/R Mapping technology can realize a relatively absolute object permanence layer, developer can encapsulate relationship data into normal object using this layer, and further use OO method to manipulate these data. After a long time, in Java developing field, design based on application of databases is relationship oriented, that is the dealing process associated with databases doesn't realize true object oriented.

The solution of EJB, JDBC permanence has such problems, developers try to use Hibernate to envelop a layer of object oriented on databases, so that developers can release from database coding which is complex, repeatable, low technology content. Hibernate appear because of requirement, it is an open source architecture based on object oriented, and encapsulate JDBC with light object, Java application can access databases through Hibernate API or with the help of JDBC, Java developer can manipulate databases with the thought of object oriented as he wanted. [3] Hibernate not only provides mapping from Java class to databases, but also provides query and resume mechanism of data. Prepared with manual manipulation of databases through JDBC and SQL, Hibernate can reduce the load of operating on databases enormously. In addition, Hibernate can use proxy mode to predigest the process of loading class, which will reduce the coding workload of picking up data from databases through Hibernate, and consequently improve the developing efficiency. And the Hibernate Architecture is as follow:

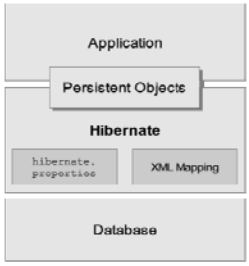


Figure 1 Hibernate Architecture

5 To Implement a Students Management System Based on Structs Hibernate and Spring

5.1 The Choice of System Mode and Design of Configurations

Struts is good at viewer, Spring is good at business layer, Hibernate is good at data permanence, to combine the three parts can achieve good system modularization, reduce coupling between system layers, and make system transplant convenient. When it is used in the software configuration design of students' management system, I employ struts, hibernate and spring as the technology developing platform of students management system. I confirm five large modules of this system according to the facts in schools, that is, students' status and dossier management, record management, diathesis measuring and judging management, teaching material management and course management. Students' status and dossier management module mainly includes the basic information of enrollment and behavior information in school. We can accomplish the following functions, recording, adding and deleting, modification and query of students' basic information. Besides achieving basic grade record, record management can base on the name, class, specialty and students' number to inquire about one subject or the general score, statistics of one subject, statistics of general score, distributing curve of students' score, test paper analyzing, score reports forms and so on. Diathesis measuring and judging management includes information after students entering college, such as social practice, part-work and part-study system, time register of various competition activity, activity content register, activity summarization, diathesis information of hortation and evaluation. Teaching material management includes such information: the name, publishing company, publishing time and price of teaching material. Course management includes three modules: the setting and management of each basic subject information, the setting and management of basic specialty and the setting and management of specialty subject. In this module we involve enrollment, modification and deleting function of students' information, other module design thought is the same.

5.2 Design of Resource Layer

1) Definition of permanence class:

Hibernate act on ordinary java object to make it a permanent class. Persistent object is a pure Java object accord with Java Bean, it contains attribute and method accord with consolidate standard. POJO is a simplex function Java object, its attribute can only accessed by its own get and set method, in this way it hides the interior implementation detail, makes the data manipulation corresponding each attribute canonical. To predigest it, I define three attribute in Student class, which is id, cardId, name, its attribute is corresponding to the field in Student table, and type is identical with it.

```
Public class Student{
private String id;
private String cardId;
private String name;
//omitting get() and set() method
}
```

2) Compiling mapping file of object-relationship:

The mapping file Student.hbm.xml corresponding to databases includes metadata that object/relationship mapping needed. Metadata includes the statement of persistent object, and the mapping relationship from each attribute of class to each field in databases. Attributes can exist as a common value or a relationship pointing to other entities, they are subsiding keywords in tables which belonged to relationship databases. Each table of the databases has a Hibernate mapping file, it is

used to create data object, Student.java correspond to Student.hbm.xml here, describe as follows:

```
<hibernate-mapping>
<class name="model.Student" table="Student" >
< id name= " id"><  generator class= "identity"/>
< / id>
<property name="cardId" type="string“/>
<property name="name" type="string“/>
</class>
</hibernate-mapping>
```

3) Hibernate allocation:

For carrying on application procedure with the Hibernate of hold out for long time turn, need to be place in the application the procedure of the pack an allocation document, for the purpose of Hibernate ability with accuracy completion allocation and beginning start to turn.The allocation document contain two kinds of format hibernate.properties and hibernate.cfg.xml, two kind allocation the item be all similar, general usage the latter.Underneath allocation hibernate.cfg.xml document.This document deposit a Mysql database a conjunction to drive procedure and register a database of user's name/password, counterfeit entity allocation document of position etc.

```
<hibernate-configuration>
<session-factory>
    <property
name="show_sql">true</property>
    <property  name="connection.driver_class">
  com.mysql.jdbc.Driver</property>
    <property name="connection.url">
    jdbc: mysql: //localhost: 3306/schoolProject?
useUnicode=true&amp;characterEncoding=GBK
</property>
    <property
name="connection.username">root</property>
    <property
name="connection.password">pwd</property>
    <property name="dialect">org.hibernate.dialect.
MySQLDialect</property>
    <mapping resource="Student.hbm.xml" />
</session-factory>
</hibernate-configuration>
```

5.3 DAO Interview Layer Design

1) Write to connect

DAO interview layer the data which be responsible for to pack first floor interview detail, can not only make concept clear, and can exaltation development efficiency. First demand write to connect, connect of the essence lie in can pass it to adjust to use whichever realization this connect of object, but need not go to understanding concrete realization type how realization of, system other mold piece all with this connect to make contact with, and usage pick up a people the benefit is at the expand of system or change, change a concrete of database interview technique or changed a database product, demand make to change to move in the realization the type, but not the other code that need to be change application be also easy to test and system transplant. The IBaseDAO most basically connect, have already establish among them, delete, the method of modification object, other all DAO's connecting all is inherit in IBaseDAO, for example IstudentDAO, if the system need to be expand, return can write IteamDAO, IcourseDAO etc..IStudentDAO source code is as follows:

```
public interface IStudentDAO extends IBaseDAO{
    public Student findById(String id) ;
    public List findByName(String name) ;
    public List getAllStu() ;
}
```

2) The realization interface

The BaseDAO inheritted IBaseDAO, realization establish among them, delete, modification method, want to use Spring HibernateDaoSupport among them, it be a convenience realization HibernateDAO super type, can make use of it to acquire HibernateDAO perhaps SessionFactory.Most convenience of the method be a getHibernateTemplate(), it return to a HibernateTemplate object, this template pair of examination type abnormality packing become solid modern style abnormality, this make you of DAO's connecting have never need to throw a Hibernate abnormality.For example in BaseDAO establish a method as follows:

```
public void createObj(Object o) {
try {getHibernateTemplate().save(o);}
catch (HibernateException e) { System.out.println
("The solid example arrive a database failure ", e);}
}
```

Other all DAO concrete type all will direct or indirectly inherit BaseDAO, and realization concrete type rightness should of connect.

StudentDAO source code is as follows:

```
public class StudentDAO extends BaseDAO
implements IStudentDAO{
    public Student findById(String id) {
Student stu = null;
try {
stu = (Student) getHibernateTemplate().get(Student.
class, id);
}
catch (HibernateException e) {
        System.out.println("check to seek id for"+
id+" of object failure", e)
    }
    return stu;
}
```

5.4 Business Logic Layer Design

At business logic layer demand earnest thinking each business logic ability use of hold out for long time layer object and DAO, want completion allocation Spring frame. In actually of the item the development, each realm will have own special business logic, just because of so, cause the code in the item height the Ou match, original probably drive heavy use of code or function, because of with concrete of the business logic bind settle at a cake of but cause very difficult drive heavy use. Therefore we realization these are concrete logic of the code sample to is divided into alone of 1 F, its purpose is hope pass a layer, lower it with system other part of the Ou match a degree.

1) Write the business logic:

DAO layer on is a business logic layer, DAO can have a lot of, but the business logic only have 1, can adjust to use in the business logic each DAO carry on an operation and write business first logic of connect, in Iservice only have the validate (int age) of a method body, certainly also can join other method, business realization the type inherit to connect gossip now among them method. We connect in our business object what the setter method of usage accept, these connect allow object of lax definition of realization, these object will drive constitution perhaps infusion. Business realization type ServiceImp code as follows:

```
public class ServiceImp implements IService{
private IStudentDAO stuDAO;
public IStudentDAO getStuDAO()
{return stuDAO;}
public void setStuDAO(IStudentDAO stuDAO)
{ this.stuDAO = stuDAO; }
public boolean validate(int age)
  { boolean flag=true;
      if(age>30)
      flag=false;
      return flag;
  }
}
```

Used Spring IoC container function here, in ServiceImp the quantity stuDAO of change of IStudentDAO be dependence Spring of infusion. Usually, application code container or frame that need to be tell, let them find out an oneself type for need, then establish object of solid example, therefore apply the code to the usage solid before the example demand establish object solid example. However, the IoC will establish the mission of the object solid example to hand over to IoC container or frame, by so doing, resources and procedure separate, the Spring pass resources to Bean form infusion procedure, be demand change resources, as long as change to move Spring of allocation document then, but need not change to move a procedure code, lowered the Ou of module to match a degree thus, exaltation type of heavy use sex, and easy to test.

2) Allocation resources that need to be infuse into:

Write bean.xml an allocation resource that needs to be infused into, among them some source code as follows:

```
<beans>
<bean id="StudentDAO" class="dao.StudentDAO">
```

```
<property name="sessionFactory">
<ref local="sessionFactory" /></property>
</bean>
<bean id="ServiceImp" class="service.ServiceImp">
<property name="stuDAO">
<ref local="StudentDAO" /></property>
</bean>
</beans>
```

The allocation of the Spring document function is a resources that install demand's infusion, use Spring management DAO and its dependence and imply to be like business management machine, object factory, include business logic of service object, with data access object these object of quote from. First, we pass to establish a value infusion method constitution the data source related parameter, then, we data source solid example infusion give data interview type, end, we for each concrete business infusion correspond interview machine, can see a Spring studentDAO concrete business infusion correspond of interview machine ServiceImp, pass interview machine ServiceImp to adjust to use the method in the studentDAO business thus.

6 Conclusions

Struts+Spring+Hibernate is the main stream application technology used by various of famous software enterprise abroad, as well as the trend of the future's J2EE development. As a front-stage control frame, Struts simplified the development of the software, effectively separated the web designer and the programmer's work, and largely increased the expansion of the project, arase the productivity, reduced the cost of maintance. As a compound frame applied for all levels, Spring has powerful applications and flexible, of which it is very suitable for some bottom platform of large scale software project. As a persistent frame level light component used for the back-stage O/R mapping, Hibernate lightly encapsulated the persistent level, reduced the complex of the program. Thus, it is easier for debugging, reduced the burden of the programmer, as well as capable of strong expansion, open API, so as to modify the Hibernate source code by ourselves, expanded the function of it. In a word, the union framework of Stuts+ Spring+Hibernate's application will be much prospected.

References

[1] ZhiGuo Shi, Weimin Xue, Jie Dong. *JSP Application lectures*. Beijing, Tsinghua University publisher, 2005

[2] Jie Meng. *Master in Spring-Java lightweight structure a development practice*. Beijing, Post and Telecom Press, 2006

[3] Weiqin Sun. *Master in Hibernate: The java object hold out for long time to turn a technique detailed solution*. Beijing: Publishing House of Electronics Industry, 2005.36~38

[4] Shifei Luo. *Master in Spring.* Publishing House of Electronics Industry, 2005

[5] Lihong Zhang, "*A Reserch on an Improved Communication Mechanism of Mobile Agent*", Lan Zhou University, May 2007

[6] Jizeng Wang, Zibin Man, Jun Zhou, "*Informaiton-table-based scheme for mobile agent communication*", Computer Engineering and Design,China, Vol.28 No.11, June 2007, pp. 2566-2568

Profile

Wei Dai, a postgraduate of Wuhan University of Technology. His research interests are Calculator network, being in conjunction with of calculator support work, artificial intelligence.

ShengJun Xue, medium total party member, Ph.D. degree. Professor, Doctor Supervisor of Wuhan University of Technology. Incumbent science and engineering university periodical agency in Wuhan pair president, part-time Hubei province the highway conveyance of the transportation committee cent Wei would pair director, Ministry of Education Doctor point the fund review an argument expert, the calculator science and technology of the Ministry of Education university and college teaching instruction committee member of committee, science and engineering university college journal(transportation science and engineering version) in Wuhan plait Wei. Attending the nation class including the fund item of the nation natural science item is 4 and manage with participate province department class and business enterprise research item 16.

A Task-and-Role-Based Access Control Model for Workflow System

Xu yi

School of information technology, Jiangnan University, Wuxi, Jiangsu, 214122, China
Email:yixu58@msn.com

Abstract

An access control model called T&RBAC (task-and-role-based access control) is proposed, which based on RBAC and TBAC and models from the tasks in workflow and dynamically manages the permissions through tasks instances and task context. In the T&RBAC model, permissions are assigned to tasks, tasks are assigned to role, and the relationships between tasks and task instances are defined. The basic concepts of T&RBAC are explained, the formalization descriptions are given, the related methods which use task context to dynamically control security permission are be put forward, and the strategies used to assign users to role are established. At last, the supported security principles are be analyzed.

Keywords: Access Control, Task, Role, Workflow, User Assignment Strategy

1 Introduction

The workflow management technology is one of emerging technologies in computer application domain in recent years, which had achieved the enhancement production organization level and the efficiency goal, through the process integration, and was widely applied to the electronic commerce, the electronic government affairs, and logistics management and so on, but its security problem day by day is also prominent. In particular, a robust secure workflow model is needed to allow controlled access of data objects, secure execution of tasks, and efficient management and administration of security (Joshi et al. 2001, Kang et al. 2001, and Thuraisingham et al. 2001). The basic security service for workflow system [1] include: Authentication and Authorization, Access Control and Audit, Data Privacy and Data Integrity, Non Repudiation and Security Management & Administration, One of most important service is the access control. Access control is the ability to permit or deny the use of a particular resource by a particular entity.

There are several types of access control model which were used in recent years, for example, Discretionary Access Control (DAC), and Mandatory Access Control (MAC) and Role-Based Access Control (RBAC) [2,5]. These models all are embarked the protection resources from the point of view of system (control environment are static state), which all are the passive security model. The basic concept of RBAC is that users are assigned to roles, permissions are assigned to roles, and users acquire permissions by being members of roles. But in RBAC, it cannot record the process which subject access objects, and cannot restrict usage of authorization permission by timeliness. Subject could execute the permission countless times which is assigned to the subject.

But the security of the workflow system has its special requirements: (1) the authorization must synchronize with current executing task, i.e. with a task beginning, the user was granted the permission by principle of least privilege, and the permission must be revoked so long as the task finished. (2) To grant permission and revoke permission are driven by the event. To execute next task usually is depended on the result of carrying out of previous task in that process, in the same way, to grant permission to that task also is depended on the result of carrying out of previous task,

so access control management have to fit together event occurrence order of sequence. (3) The authorization management is the context sensitive, which depend on the historic record of process implementation. (4) The permission that was granted is only effective in the stipulated interval.

In 1997, Professor Mason University's Ravi Sandhu proposed one kind of new security model: Task-Based Access Control (TBAC) [3,4,10], which establishes the security model and realizes the security mechanism according the task (activity), and is a activate security model because the access control for object along with the task context. The basic conception of TBRC includes authorization step, authorization unit, task and dependency. But TBAC isn't suitable for a large business enterprise system; because TBAC only simply introduces the trustee-set to express task performer, and does not clearly separates the role and the duty, does not support role hierarchy and passive control access.

In this paper, an access control model called T&RBAC (task and role based access control) is proposed, which based on RBAC and TBAC[3,4,9]. We reflected the trustee-set in the TBAC to map the role in the RBAC, and introduced the conception of task, task instance and task context, then changed the authorization frame from user-role-permission for RBAC to user-role-task-permission for T&RBAC, thus enhanced the security of the workflow system.

2 Task and Role Based Access Control (T&Rbac) Model

T&RBAC selects the role and the task as two basic characteristics, and places the task and the role in the equal important position. The primary concept of T&RBAC is that users are assigned to roles, tasks are assigned to roles, permissions are assigned to tasks, user acquires tasks by being members of roles, and user owns the permission for task while carrying out it. In this model, the task can be categorized into not-workflow task (NWF_T) and workflow task (WF_T) from the point of view of enterprise and application layer. A NWF_T is independent and hasn't logical order, such as look into awaited affair. Access control of NWF_T makes use of RBAC, and statically assigns privilege; on the other hand, access control of WF_T makes use of TBAC, and dynamically assigns privilege. The privilege changes along with the task execution, the role has the privilege only while carrying out task, the privilege is revoked when the role does not execute the task. Figure 1. of the elements.

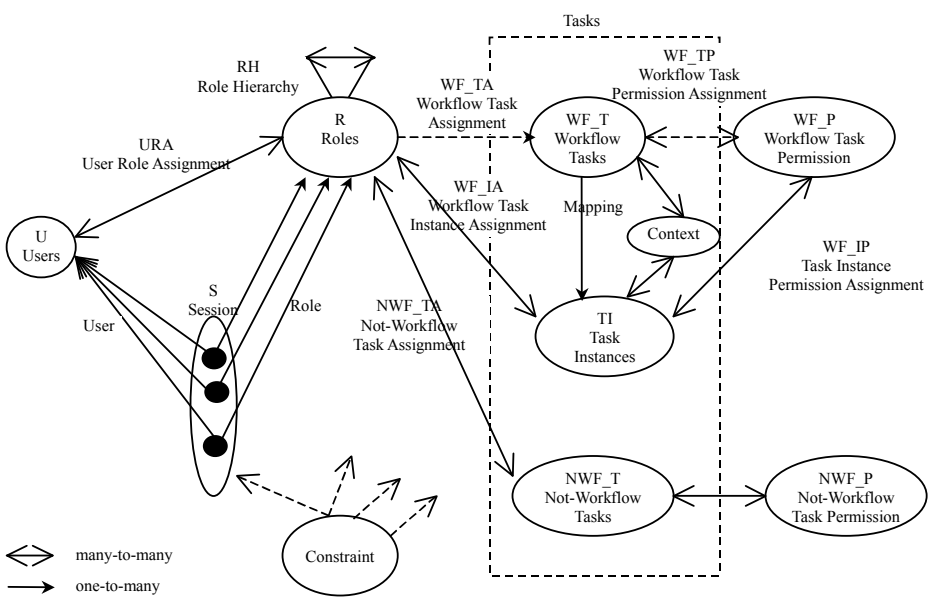


Figure 1 T&RBAC Model

2.1 Basic conception

(1) User (U): Any person who interacts directly with a computer system. U={u_1,u_2,,,u_n}.

(2) Role (R): A job function with the organization that describes the authority and responsibility conferred on a user assigned to the role. R={ r_1,r_2,,,r_n }.

(3) Session (S): A mapping between a user and an activated subset of the set of roles the user is assigned to. It could be expressed: U:S→U, R:S→2^R.

(4) Task (T): A task is an indivisible logical unit in workflow process; it could accomplish one special purpose. It is composed of start condition, end condition, application data and restrictive condition. The task is a static concept and predefined by process creator; it can be categorized into not-workflow task (NWF_T) and workflow task (WF_T) according to the task type.

(5) Permission (P): The permission is an abstraction set that has some operations to the some object. The permission of Non-workflow task (NWF_P) is the set of straight executive permission task need; the permission of Workflow task (WF_P) is set of active permission which trigged by task needed. The active permission set of current task is the executive permission set of next task. Non-workflow task permissions are disjoint from workflow task permission, i.e. NWF_P∩WF_P =φ.

(6) Task Instance (TI): A task instance is an instance of the task, or in other words it is a copy of the task that is made to run an instance of it.

(7) Context (C)[7]: The context is the restrictive condition for restricting whether the task can be executed. It includes: user identifier, time, input condition, process relevant data and so on.

2.2 Formal descriptions

Definition 1: T&RBAC= (U, R, T, TI, P, C, WF).

(1) Workflow(WF) is composed by a series of Task(T). $WF \subseteq T \times T \subseteq 2^D$, D={Order Dependency, Defeat Dependency, Divided Permission Dependency, Agent Dependency }[3] .

(2) User Role Assignment (URA): a many-to-many user-to-role assignment relation. $URA \subseteq U \times R$.

(3) Task Role Assignment (TRA): a one-to-many task-to-role assignment relation. $TRA \subseteq T \times R$.

(4) Task Permission Assignment (TPA): A many-to-many permission-to-task relation. $TPA \subseteq T \times P$.

(5) Role Hierarchy (RH): The $RH \subseteq R \times R$, is an partial ordering relation, expressed as "≤", it means grade relation between roles within organizes.

(6) Task Instance Mapping(TIM) ζ: T→2^{TI}, a function mapping each task template to many task instance, such that ζ(a) ∩ ζ(b) =φ, if a≠b and a,b∈TT.

(7) Workflow Task Instance Role Assignment (WF_IRA): $WF\_IRA \subseteq R \times TI$, it is determined by ζ and TRA.

(8) Workflow Task Instance Permission Assignment (WF_IPA): $WF\_IPA \subseteq TI \times P$, it is determined by ζ and TPA.

(9) Context (C):C=(userID,UA,RH,TRA,TP,time, input condition, process relevant data).

Definition 2: Related method

(1) session_user: S→U, a function mapping each session s_i to the single user session_user(s_i).

(2) session_roles:S→2^R, a function mapping each session s_i to a set of roles.

$$session\_rols(s_i) \subseteq \{r \mid (session\_user(s_i), r) \in URA\}$$

(3) $tasks(r) = \{tt \mid (\exists r' \le r)[(r', tt) \in TRA]\}$, a function mapping tasks can be assigned to role r.

(4) $taskIns\tan ce(C) = \{ti \mid (\exists r' \le r)(\exists tt \in tasks(r'))$
$[ti \in \zeta(tt)] \wedge [canExecute(C, tt) = true]\}$, a function mapping task instances to a user authenticated in task context.

(5) $tpermmisions(C, ti) = \{p \mid \exists tt[(ti \in \zeta(tt))$
$\wedge (ti \in taskIns\tan ces(C)) \wedge ((tt, p) \in TPR)]\}$, a function mapping permissions for a user authenticated in task context.

2.3 User assignment strategy

After the user was authenticated and then function taskInstances(C) was invoked, user could obtain task instance. It is important that how to select a user from session_user(s_i) to execute this task instance. If selection is improper, it will be able arose the successive task to be unable to execute, then suspend or abort the

workflow process. So, in this paper, we proposed relevant user assignment strategy [8].

(1) Set user-to-role strategy while assigning roles to users. The user-role strategy include: static participate, competitive participate, ordinal participate and agent participate [9].

(2) Bring forth special skill of each user sufficiently, let one user to do similar task continuously.

(3) During the period of executing workflow process, once overload of one user appearance, then the system should assign task to other user which has the same privilege and load less than that.

(4) A pair of conflict tasks cannot be executed by the users which have benefits with each other.

(5) During the period of executing workflow process, if the number of executing users is greater than the maximum users, this task could not assign to other users to execute again.

2.4 T&RBAC security analysis

T&RBAC supports two well-known security principles:

(1) Least privilege: While executing the task, the basic permission is assigned to the role which is executing task; on the other side, if the task has not been executing or aborted, the assigned permission would be revoked from the role when not execute task or abort task. Thus, the correspondence permission only can be assigned to the user when the task is being executing and the task is assigned to the role and also the user is assigned to the role.

(2) Separation of duties: Invocation of mutually exclusive roles can be required to complete a sensitive task, such as requiring an accounting clerk and an account manager to participate in issuing a check.

3 Conclusions

Safely access control is very important for the large scale and complex workflow system. In this paper, we proposed a new access control model called task-and-role-based access control model (T&RBAC). This model can integrate the actual workflow and safety access control strategy together, combine advantage of RBAC and TBAC. It can express control mechanism of complicated workflow clearly.

References

[1] WFMC.TC0021019: Workflow Management Coalition Workflow Security Considerations White Paper, 1998

[2] Sandhu,R.S,Coyne,E.J.,Feinstein,H.L.and Youman,C.E. Role-based access control models. IEEE Computer, 1996, 29(2), pp 38-47

[3] Thomas RK, Sandhu RS. Task-Based authentication controls (TABC): a family of models for active and enterprise-oriented authentication management. Proceedings of the IFIP WG11.3 Workshop on Database Security, 1997

[4] Roshan K. Thomas and Ravi S.Sandhu. Task-based authorization: A paradigm for flexible and adaptable access control in distributed applications.Proc.16th NIST-NCSC National Computer Security Conference,Baltimore,MD, 1993;pp 409-415

[5] Huang, W.-K. and Atluri, V. SecureFlow: A Secure Web-enabled Workflow Management System, in 'Proceedings of the 4th ACM Workshop on Role-Based Access Control',1999, pp. 83-94

[6] HUANG Fujan.Task-Role Based Access Control Model and its realization technology. Journal of Jining Teachers College,2006,27(6);pp.26-28

[7] FU Song-ling，TAN Qing-pi. Security Task & Role-based Distributed Workflow Model. Journal of national university of defense technology,2004,26(3),pp.57-62

[8] ZHANG Hai-juan , FU Zheng-fang , ZHANG Hong-lin. User assigned policy of security workflow. Application Research of Computers,2008,25(1),pp 238-240

[9] SHANG Qiu-ming, CAO Bao-xiang. Analysis of Workflow Drived by Role and Application of UML-uc Modeling Technology. Application Research of Computers,2006,23(7).pp 126-128

[10] Cui-xiao ZHANG, Ying-xin HU,Guo-bing ZHANG. Task-Role Based Dual System Access Control Model. IJCSNS International Journal of Computer Science and Network Security, 2006,6(7B).pp,211-215

The Application of Network Technique in High-school Laboratory Management

Min Zhang Shuangchen Ruan Rong Yang Shucai Cai Wenxiao Huang

College of Electronic Science and Technology, Shenzhen University, Shenzhen, Guangdong, 518060, China
Email: zhangmin@szu.edu.cn

Abstract

The experiment teaching was very important for undergraduates, as it can help them to have the makings of a fine scientist or engineer. With an example of a laboratory booking management web site based on ASP.NET and ADO.NET, the paper analyses the actuality and problem of laboratory management in high school, and introduce the application of network technique in high-school laboratory management. The practice shows that the management web site can make the management of open laboratory more science, can make the use of instrument more availability, and is very useful in other side in our life with little change.

Keywords: Network technique, Management System of Laboratory, ASP.NET, ADO.NET, C#

1 Introduction

High-school laboratories have very important effect on improving undergraduates' makings of a fine scientist or engineer. How to open laboratories for undergraduates is a hot issue of every high-school laboratories' work [1]. Shenzhen University is located in the Shenzhen Special Economic Zone, who had a lot of researching into management of opening in the laboratory [2-3].

2 The Actuality of the Laboratory Management

On the one hand, our country's high education develops rapidly, and the number of undergraduates grows heavily in recent years. Because of the practical character on the experimental teaching, the venue, the per capita pieces (sets) and safety are a few specific requirements or restrictions. And it is not appropriate to expand the number of students of the experimental class. In this case, mostly high schools increase the schoolteachings and experimental teaching assignment to be adapted well to the improvement of the number of undergraduates. On the other hand, because of universal attention to manipulative ability, the schools have increased experimental courses and projects. These measures have also greatly increased the workload of the experimental teaching. Thus, driven by the two major factors -- the development of higher education and social needs, experimental teaching tasks show a sudden increase trend.

When the open laboratory construction becomes a hot issue in laboratory work again, we aware that the experimental time and experimental content must be open [4]. This experimental teaching management bring a bigger challenge: The contents of opening up require more innovative teaching experiment; experimental time of opening up require the experimental arrangement more flexible. Under the traditional manage method of the laboratory, we only need to finish the experiment course and experimental process before the beginning of each semester, and would not need to make too many changes. And the open laboratory because of the opening of the experiment content, making the arrangements for experimental time would not follow the traditional method, and often in actual operation will see the following two conditions: 1.some period of opening time, no students come to do experiment; 2. another laboratory open sessions, the number of students to experiment greatly exceeding the laboratory's support limit.

To resolve these problems, an efficient, and modern experimental management measure must be in dint of to go with actual demands. At present, major colleges and universities have explored new experiment management artifices, such as telephone booking experiment, spot registration, and other management methods. Telephone booking needs cumbersome procedures of specialized duty, modify, change; spot registration management is simple, but it can not guarantee that the greatest use of laboratory resources. And under the two ways, if sudden incident such as large area viruses' inbreak or power cut occur, it is difficult to notify each of those appointments.

MIS (Management Information Systems) is a modern management artifice now generally recognized by the well-known company. Shenzhen University in the exploration of development of laboratory management, with the idea of MIS system, using ASP.NET and ADO.NET technology, developed an on-line booking function of laboratory management site. Such open laboratory information management system is an important part of the laboratory to promote the construction of the benign development of laboratory management, improve their self-management ability and self-restraint, and it is truly the necessary management foundation towards open.

3 The Introduction of the System

The main task of "Booking open-experiment management system" is experimental arrangements and laboratory equipment management, including the introduction and propaganda of the actuality of the laboratory management, and discussions of the experimental technology. Experimental arrangements stressed change time and content of the relatively fixed arrangement fashion into the on-line booking fashion; experimental device management including the procurement and use of equipment, mainly relate to the experimental or laboratory planning content. In addition, the introduction and propaganda of the actuality of the laboratory management, and technical discussions etc is the postulate to strengthen the daily laboratory management. System structural shows in Figure 1.

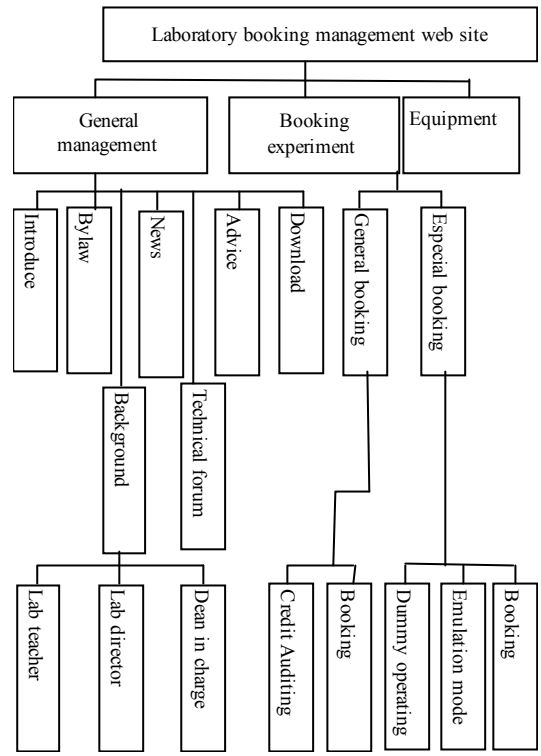


Figure 1 Systematic Structural Sketch of laboratory booking management web site

4 The Technical Foundation

ASP.NET is a powerful, very flexible server-side technology, used to create dynamic WEB pages [5-10]. Web site absorb server-side dynamic web page (ASP. NET technology), which will be convenient to capture the underlying database access functions, liberating from the traditional repetitive design of static pages, and can be updated in real time, real-time processing. NET Framework is a huge toolbox used to create all applications (especially Web applications). As a component of the technology of Framework, ASP.NET compared to the past, server-based dynamic web page technology has following several revolutionary breakthroughs:

1) Including the introduction of object-oriented technology, to develop object-based server controls, ASP.NET Web developers can direct control the HTML

tags object appeared on the browser though server-side's program code.

2) ASP.NET itself against the state of their own maintenance, web design state maintenance personnel do not deal with the relevant details.

3) ASP.NET support multi-language programming model, it can be used to prepare ASP.NET dynamic Web content as long as it is the .NET-supported programming language, including the C # and VB.NET. And this system use VB.NET to develop. ADO.NET is a group of database objects developed from Microsoft with. NET platform, specifically designed to deal with the application's data processing problems, and to provide Application Programming Interface that the database functions Construction needed, respectively, corresponding to the operation of the database the necessary capabilities. In addition, ADO.NET also includes a database connection used to establish the connection object Connection.

5 The System Design

Appointment of open laboratory management system base on administrator's setup automatically generated the booking experiments and experimental time for each user. Internet users can readily know that the latest using circs of laboratory, and appoint experiment according to their actual. The system also supports user feedback, information dissemination etc. website basic functions at the same time. According to the actual laboratory management, system use ACCESS database based on the two-dimensional table relationship to manage system data, including the main table and the table is as follows:

1) The user of personal information table

School, student names, passwords, date of birth, city, address, telephone number, e-Mail, gender, whether the administrator, whether the authority opening of groups.

2) Information Table groups

Group number, group description.

3) Laboratory Information Table

Number laboratories, laboratory name, capacity, guidance teachers, laboratory location, laboratory telephone, the state laboratory.

4) Experimental information table

The number of names, capacity, the number of bookings, guidance teachers, laboratory location, the name of the experimental condition, the experimental time, experimental date, laboratory telephone.

5) Table booking information table

Booking number, school, student name, experiment # experimental name, date of experiments.

6) Information Table

Information number, title information, release dates, publication, published content.

7) Message Table

Message code, Title, message content, message, your name, e-Mail messages, message time, the number of hits, the number of replied information, whether the back posts.

6 The Development

According to the design requirements, the system is divided into five modules for design; the various modules connect with each other. Security management module is on top of four modules (user information management, laboratory information management, information dissemination management, Message Management), and to limit the authority of different users. Administrators can use all the system function while other users can only use part of it. Modular design is conducive to the improvement of each parts function, adding space for new modules as well as, conducive to system's expansion.

1) Security management module

System uses dual restrictions to ensure its safe operation. First, the website added limiting functions on user rights in the management; secondly, in each page operated on management page user's privileges are restricted. At the same time, the paper also provides functions of preparation and updates of the database files, which has played a fundamental protection.

2) It separate different plate through the management pages when design, and use different groups' number to distinguish different users, and systems based on different groups, show different pages. Management pages use Dreamweaver custom labels DataSet to connect with the database, and hold the landing information (including group).

3) User information module

User information Modules includes the general user information modules and administrator information module. User information Modules is personal information management module while Information Manager modules include all the user information management and it also has set up user privileges and blacklist management function (add blacklist users will be the cessation of all rights).

4) Experimental information management module

This part is the core of this system, including the laboratory information table, experimental information table, booking information table. Administrators can manage the information of 3 tables (add, modify and delete). Ordinary users can only manage their own booking information table, or check for other 2 information Table.

Each step of this operation corresponds to a database operated a number of steps. These processes are transparent to users, so that will be greatly reduced users operational complexity, design of humanity to improve the use of the system efficiency.

7 The Features of the System

Booking function On-line is a new type of experimental arrangement, it will experiment for the initiative from students to teachers to increase the autonomy of the students, reduce teacher's Course Scheduling tasks. Compared to the traditional method of laboratory management, the laboratory management system of on-line booking function has the following characteristics:

1) The introduction of a modern enterprise management of MIS management concept is conducive to laboratory management efficiency and maximizes the effectiveness of the resources.

2) The introduction of modern communications technology, make the management tools advancing with the times. For example: the function of the mass-mailing or SMS notification appointment user. When there are special events need to give notice of the appointment users, automatic mail lists or messaging will play an effective role;

3) Use network technology to increase the teacher-student interaction platform. System technical discussions with the message, "according to tutor teachers, students passive learning" to "take the initiative to ask students and teachers for teaching, and students learn each other" model, and raise their interest in learning and learning initiative;

4) The use of virtual and simulation technology, has reduce the loss of precious equipment, and effectively prevent accidents occurred. System specially increases the virtual operation examination for the valuable equipment. Operators must pass online examination; only after obtain the competency of valuables laboratory equipment operation can booking the use of expensive laboratory equipment. Managers can through adjusting the difficulty of questions to limit the operation lever of the users, to reduce misuse possibility of valuable equipment, he improving the safety factor of operating valuable equipment.

5) It established a perfect evaluation methods of Integrity, improved the user's self-management awareness. Most of the appointments are the basic judge on students' integrity according to whether come to experiment to. This evaluation method is crude, and for the late, or leave early, experimental content (experiments, or in the name of experiment online play games) etc. situation not to deal with. System further standardizes the evaluation methods of Integrity, and to fully take into account various factors, making the integrity management more scientific.

8 Conclusion

Based on a booking management system of opening laboratory, the paper introduced the application of network technique in high-school laboratory management. The practice shows that network techniques enable teachers and undergraduates to take part in the general management of the laboratory, reduce Laboratory manager's workload and debase the difficulty of laboratory management. It can resolve teacher shortage, equipment lack, venues inadequate, and can mostly provide innovated space for undergraduates. It redounds to improve undergraduate skills. It has practicability in the use of laboratory and management. At the same time, with modifications of the system, it can also be used in social production and other aspects of life, such as booking tickets, booking cards.

Acknowledgements

The authors warmly thank Geng Yanxia and Cao Guangzhong for his involvement in the work of experiments and useful discussion.

This work was from Project BKYBJG20060270 supported by Guangdong Province Higher Education and Teaching Reform Project.

References

[1] Yin Xin, Liu Zijian, Liu Hongxia. "Exploration and practice of laboratory opening", Experimental Technology and Management, Vol.23, No.8, 2006, pp.101~103

[2] Xu Ming, Cai Zhenxiang. "Design and application of the web-based web programming experiment system", Experimental Technology and Management, Vol.23, No.8, 2006, pp.69~71

[3] Zou Yongdong, Zheng Yizhi, "To Open the Key Laboratory for the Undergraduates to Culture their Practical Ability and Creativity", Experimental Technology and Management, Vol.23, No.3, 2006, pp.109~111

[4] Pan Hui-wen, "Discussion on the Style of Laboratory Opening", Journal of Guangzhou Physical Education Institute, Vol.25, No.3, 2005, pp.125~129

[5] Chris Hart, John Kauffman, Beginning ASP.NET 2.0, Wiley Publishing Inc., 2006

[6] Dino Esposito, Programming Microsoft ASP.NET 2.0 applications: advanced topics, Microsoft Press, 2006.

[7] Jiang Pei, Wang Xiaomei, Programming ASP.NET Web, Tsinghua University Press, 2007

[8] Li Deqi, Programming ASPNET, Post & Telecom Press, 2007

[9] Gao Yixin, Internet Application Development With Asp, Post & Telecom Press,2008

[10] Damon Armstrong, Pro ASP.NET 2.0 Website Programming, Post & Telecom Press,2008

Application on Filtration Mechanism of CORBA Notification Service in Network Management System *

Xiaohong Wang Jingyang Wang Min Huang Huiyong Wang Liyan Zhang

Hebei University of Science and Technology, Shijiazhuang , Hebei , 050054, China
Email: ever211@163.com

Abstract

The event service in CORBA has offered an asynchronous communication mechanism between the event supplier and the consumer, but this mechanism has some disadvantages in practical application, such as the connections between the event supplier and the consumer can not be preserved permanently, the event data that has not been sent also can not be preserved permanently, and the QoS is not supported. CORBA notification service expands the event service, has some new functions such as structured event, filtration mechanism and QoS support, it can be more adaptable in various kinds of environments. This paper has illustrated some new characteristics such as structured event, filtration mechanism and QoS in CORBA notification service. A filtration mechanism is proposed to implement the event report while applying CORBA notification service based on push model in the network management system. This mechanism is more efficient in distinguishing different users and different events. Therefore, the transmission of event is more efficient, costs in processing are greatly decreased.

Keywords: CORBA notification service, Push model, Filtration mechanism, Network management system

1 Introduction

In order to solve the interconnection problems between hardware and software system in the distributed processing environment, OMG (Object Management Group) has put forward a kind of new solution, which is CORBA (Common Object Request Broker Architecture) system architecture. CORBA is not dependent on the programming languages, software and hardware platform and network protocol[1][2][3].

The event service in CORBA has offered a communication mechanism between event supplier and event consumer. This mechanism, which has loose and asynchronous characteristics, is used to set up a common distributed event model. With the help of event service, CORBA has the same characteristics as a middle component primarily. However, it still has obvious limitations as follows:

- Event filtering is not supported in no-type event service. It means all events are transmitted to all event consumers whether the consumer needs these events or not; to typed event, the interfaces need to be set between the event supplier and event consumer so that specified events can be transmitted.
- The connections between the event supplier and the event consumer can not be preserved in event channel permanently; the event data that has not been sent also can not be preserved permanently. This method of discarding registration and event data will cause that this information can not be reused whenever the system restarts.
- The event service does not support QoS (Quality of Service), so the reliability of communication can not be guaranteed.

CORBA notification service is the expansion of the event service. With some new functions of notification service such as structured event, filtration mechanism and QoS etc. it can be more adaptable in various kinds

* This paper is supported by CSC to Xiaohong Wang and the fund of Shijiazhuang (Grand NO. 07113431A)

of environments, for example, managed devices request communication with higher level.

In network management system, data information that need to be report actively by managed system are mainly the real-time state information of the device and fault information of the device. For managed device, the event of report is different; for management level, the event that network manager required is also different. In order to get the real-time event efficiently, this paper adopts CORBA notification service based on push model to carry out the event report function in each module of network management system [4],[5],[9].

Meanwhile, in the event transmission of CORBA, if the event channel sends the events to all event consumers that connect to it without any judgment to the event data, costs in processing are great. In order to solve this problem, the filtration mechanism is necessary when CORBA notification service in the network management system is used, so that a certain event can be ensured to send to the consumer who cares about it accurately.

2 Characteristics of Notification Service

The traditional event service consists of three parts: event supplier, event consumer and event channel. The event supplier sends all events to the event channel; the event consumer obtains all events from the event channel. In the notification service, three new concepts are introduced: event channel factory, event management object and event proxy object. The event channel factory manages event management objects; event management object manages event proxy object. The notification service structure is shown as Figure 1.

When CORBA notification service is applied in the network management system, the push model is adopted to perform the transmission of the events. The event supplier pushes events to the event consumer proxy in the event channel factory; the event consumer proxy communicates with the event supplier proxy; the event factory pushes the events to the event consumer. The push model is better than the pull model. In push model, circular enquiry is avoided and the real-time transmission of the event is fully guaranteed.

The CORBA notification service has some new characteristics such as structured event, filtration mechanism and QoS etc.

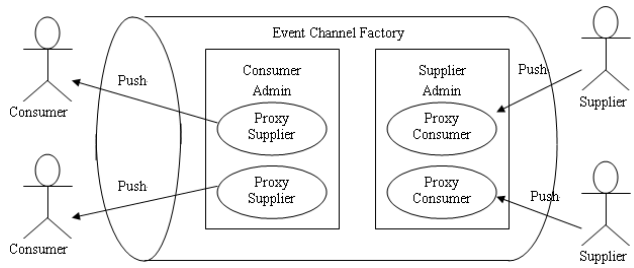


Figure 1 Notification service structure

2.1 Structured event

A structured event is defined as the data structure of event, which can be used to represent and store all kinds of events. The structured event consists of two parts: Event head and event body. The event head is divided into two parts: regular length part and variable length part. The regular length part is used to define domain name, type name and event name of an event; the variable length part is mainly used to define QoS attributes of an event. The event body is divided into filtered event body and retained event body. The event body mainly stores events that can be filtered based on a certain policy which user defines. In addition, based on structured event, the notification service has also defined the sequence event type. This type is an array of a structured type event. While events of this type are transmitted, an event supplier can transmit a lot of events once a CORBA method is activated, event consumers can receive a lot of events once a CORBA method is activated too.

2.2 Filtration mechanism

Contrast to event service, the most improvement at performance in notification service is the introduction of filter object. Through this mechanism, Users can accurately locate events that they needs. The notification service has defined a group of interfaces in CosNotifyFilter module to support event filtering. There are two kinds of filter object: forwarding filter and

mapping filter.

- The forwarding filter is used to confirm whether event should be transmitted forward. If the event content corresponds to the restrictions that are defined in the forwarding filter, this event would send forward; otherwise, it would be discarded.
- The mapping filter is used to change the attributes of the priority and the life cycle of an event. Because of the introduction of the mapping filter, the priority and the life cycle of an event can be changed to influence the transmission and other processing.

This paper introduces CORBA notification service to the network management system. Mainly three aspects should be considered. How to create the filter object? How to carry out the filtration mechanism in the network management system? Finally how to implement the efficient transmission of events?

3 Implementation of the Filtration Mechanism

The system architecture of the network management system based on CORBA notification service is shown as Figure 2.

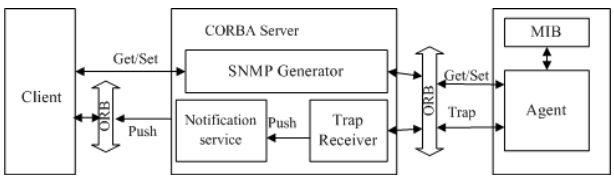


Figure 2 System architecture of network management system based on CORBA notification service

The CORBA service module is the key part of the whole network management system. The event report function adopts push model, which is more propitious to transmit real-time data.

In the network management system, CORBA notification service can implement that the client obtains the real-time state information and the fault information of the device. However, different client demands for different event. For example, different users care about different device type information and different kinds of fault information, but the event channel can not distinguish these events effectively. If there is no filtering, the event channel will send all events to all event consumers that have been connected to the event channel. This will cause great costs in processing. So, the filtration mechanism needs to be implemented in CORBA notification service, thus the client can only obtain the event that this client needs. The filtration mechanism belongs to CORBA notification service [10][14].

In the implementation of the filtration mechanism of CORBA notification service, function module in event filtering consists of four kinds of objects: forwarding filter, mapping filter, filter factory and filter manager.

- Forwarding filter: it is used to confirm whether event should be transmitted forward. If the event content corresponds to the restrictions that are defined in the forwarding filter, this event would send forward; otherwise, it would be discarded..
- Mapping filter: it is used to change the attributes of the priority and the life cycle of an event. Because of the introduction of the mapping filter, the priority and the life cycle of an event can be changed to influence the transmission and other processing.
- Filter factory: it is used to create forwarding filter and mapping filter.
- Filter manager: it is used to add a filter object to the target object or delete a filter object from the target object.

In order to introduce filtration mechanism to the notification service, the forwarding filter and the mapping filter should be created and added to the admin objects and the proxy objects in the event channel. The notification service structure of adding filter objects is shown as Figure 3.

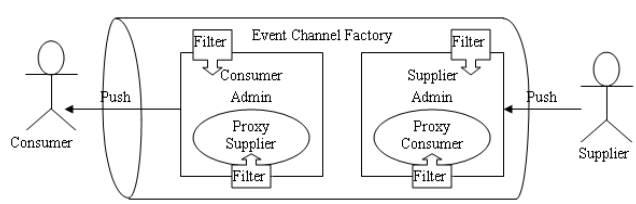


Figure 3 The notification service structure of adding filter objects

The filtering function is mainly implemented by forwarding filter and the mapping filter. The forwarding filter mainly implements three kinds of function: Dealing with the restrictions; matching the restriction with the event; registering objects who interested in present event. The mapping filter mainly implements two kinds of the function: Dealing with the restriction and matching the restriction with the event.

When the filtration mechanism of CORBA notification service is used in the network management system based on push model, a filter object is first added to the supplier admin object; the proxy consumer object managed by the admin object also adds its filter objects. When an event arrives, the filter in proxy object combines with the filter in admin object, and then the present restrictions will be generated at the proxy object jointly. The restrictions will match with the structured event, and finally the events that correspond to the restrictions will be added to the event transmission queue. In the consumer admin object, because of the symmetry of the notification service in structure, there will be the same process. In the process, a mapping filter can only be attached to the consumer admin object and the proxy supplier object [15].

In the network management system, when transmitting an event, it is important to distinguish that whether this event represents the `state information or the fault information of a device, and which device sends this event. Therefore it is clear that which client the event should be transmit to. So, when a forwarding filter is created, the restrictions must include the device identification, the basic type (state information or fault information) of an event and the general type (the type of state information and fault information) of an event. Meanwhile, because there are some clients care about the state of all devices, and when managed devices send the real-time state information or the fault information, the priority and life cycle of the event is needed. So, when a mapping filter is created, the time when an event start and the priority of an event's general type should be considered. The transmission of an event can be confirmed through the time attribute and the priority of the event.

The general process of event filtering is shown as Figure 4.

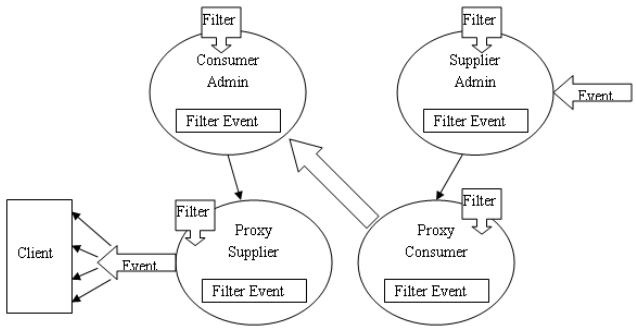


Figure 4 General process of event filtering

4 Conclusions

This paper puts forward to a filtration mechanism applied in CORBA notification service based on push model, which is used in the network management system to implement the event report. The application of the filtration mechanism is very efficient in distinguishing different users and different events, make the transmission of event more efficient. The filtration mechanism based on CORBA notification service is not only applied in network management system, but also widely applied in other fields of CORBA notification service.

References

[1] Object Management Group, "The Common Object Request Broker Architecture and Specification, " Editorial Revision : CORBA3.0.2 , 2002 ,12

[2] Object Management Group, "Event Service Specification," Version1. 2 , 2004 ,10

[3] Object Management Group, "Notification Service Specification," Version1. 1, 2004 ,10

[4] H Farooq Ahmad, "Multi2Agent Systems: Overview of a New Paradigm for Distributed Systems," The 7th IEEE International Symposiumon High Assurance Systems Engineering, 2002. 11

[5] Michi Henning and Steve Vinoski, "High Level Programming Based On C++ CORBA," Tsinghua press, 2000.11, pp.660-688

[6] N. n Hu and P. Steenkiste, "Evaluation and Characterization

of Available Bandwidth Probing Techniques," IEEE Journal on Selected Areas in Communications, vol. 6, 2003, pp. 879-894

[7] Jiang Ye,Zhao Yunhe, and Chai Zhi, "The Designation and Realization of the Communication Network Management System," Computer Engineering, vol.7, 2006, pp. 54-58

[8] A. A. Alhussein and R. Sumit, "Modeling Random Early Detection in A Differentiated Services Network," Computer Networks, vol.5, 2002, pp. 537-556

[9] B. Allcock, J. Bester and J. Bresnahan, "Data Management and Transfer in High Performance Computational Grid Environments," Parallel Computing Journal, vol.5, 2002, pp. 749-771

[10] Zheng Xianrong and Chen Qiang, "Study and Design of Notification Service Integration Based on CORBA ComponentModel," Computer Application Research, vol.8, 2005., pp. 47-48

[11] Qin Ke and Yang Gelan. "CORBA Technology Introduction," SHANXI Science & Tecnology, vol.1, 2006, pp. 22-23

[12] Din Yan, Guo Changguo and Wang Huaimin, "Design and Implementation of Real-time Notification Service," Computer Engineering, vol.5, 2005, pp. 94-95

[13] Chen Jian and Li Maoqing, "The Application of CORBA Notification Service in Network Management System," Journal of SHANXI University of Science & Technology, vol.4, 2005, pp. 85-88

[14] Lu Xin and Peng Laixia, "Research and Application in Network Management of CORBA," Modern Electronic Technology, vol.10, 2006, pp. 47-49

[15] Ding Yan, Dou Lei and Wang Huaimin, "Implementation of Event Filtering in OMG Notification Service," Computer Engineering & Science, vol.6, 2003, pp. 57-60

Enterprise Network Security Analysis and its Basic Solving Scheme

Xiaodong Zhao[1,2] Chune Zhang[3] Sufei Yang[4]

1 Hebei University of Technology, Tianjin 300130, China

2 Hebei University of Science and Technology, Shijiazhuang, Hebei 050054, China

Email: zhaoxiaodong10082@126.com

3 Computer Science&Engineering Department, North China Institute of Aerospace Engineering, Langfang, Hebei 065000, China

Email:zcxe76@sohu.com

4 Hebei University of Science and Technology Shijiazhuang, Hebei 050054, China

Email: yang_pingping_happy@126.com

Abstract

With the sustainable development of the network economy, the network made numerous organizations can sharing information and resources, such as enterprise, tissue, government departments, while enjoying the convenience brought by the network, also causes kinds of problems, the security problems particularly outstanding. So, understand kinds of threaten faced by network, prevent and eliminate this threaten, implement real information network safety have been the most important thing in the development of every organization's information. This paper analyzing the common risk, characteristic and security requirements of enterprise network security, given an integrated stereo network security solving scheme.

Keywords： Network; Network Security; Practical Technology; Solving Scheme

1 Introduction

Network information security means the hardware, software and its data of network should be protected, not be destroyed, changed and divulged because of accidental or wanton reasons, system can continuous, reliability and normal running, network service not interrupting. It involves a widely field, all fields involve the related technology and theory of network information's security, integrity, availability, authenticity and controllability are the research fields of network security. With the development of network and the emerging of the network attack, such as network virus, network hacker and so on, network security will be the important factors which directly influence the development of enterprise in the enterprise network construction.

2 Common Risk Analysis of Enterprise Network Information System

Because enterprise network information system is made up of internal network, external network and enterprise wan, so the network structure is very complex. The common threaten mainly from virus invasion, hacker intrusion, denial of service, password declassification, network eavesdropping, data tamper, waste mail, malicious scanning and so on. The factors which influence system security mainly have the following aspects:

1) Unintentionally failure of human

Many reasons will bring threaten to network security, such as the security vulnerability caused by the operator's not suitable configuration, the user's security

conscious not strong, user's password select escaped, user lent their own account to others or share with others and so on.

2) Malicious attacks of human

This kind attack mainly have two kinds: one is active attack, the other is passive attack. The former could selectively destruct the effectiveness and integrity of information by various ways; the latter is intercepting, stealing and deciphering the important secret information in the situation which not influencing the normal work of network system. These two kinds of attacks all will cause great harm to computer network and leak confidential data.

3) Vulnerability and backdoor of network software

The network software completely has no defects and vulnerabilities is impossible, and these defects and vulnerabilities are the first goal of hacker attack. In addition, the backdoor of software is setting by the software company's design programming personnel for their own convenient, it generally not know by other, but once the backdoor is open will cause unbearable result.

4) Security leak of net bridge

The net bridge is the interconnection equipment which independent to protocol, works in the second layer of OSI reference model, complete the forwarding of data-frame, main aim is provide transparent communication between connective networks. The usage of the net bridge is widely, but the interconnection of net bridge also has many problems. For example: because the net bridge not barrier the broadcast message in network, so there may cause broadcast storm when the scale of network is larger, made the whole network is filling by the broadcast information until complete paralysis; the net bridge will make the internal network and external network to one network when interconnection with external network, both parties open their network resources to counterpart completely; Because the net bridge transmission data information packet based on optimum effect, so there may cause data losing, bring more hidden trouble for network's security.

5) Security leak of router

The router works in the third layer of OSI reference model (network layer). The router has two kinds of selection strategy, that is static routing and dynamic routing, Corresponding to this, there are two kinds of routing table: static and dynamic. But the dynamic routing table has modifiability, so it may bring threaten to the network security.

3 Characteristic and Security Requirements Analysis of Enterprise Network Information System

3.1 Security risk characteristic of enterprise network information system

Any network system all exist kinds of security risk, the enterprise network system is also without exception. Exactly understand the enterprise's network system security is the basic of building reasonable security requirements. Through the analysis to enterprise information system's security threaten, know that network security risk has the following characteristic:

1) Different system environments have different risks

2) The risks will not product without causes, it product mainly by human

3) The threatens and results to network security caused by different risks are not the same

4) The possibility and severity of various risks are not the same

3.2 Structural characteristic and security requirement characteristic of enterprise network information system

Nowadays enterprise network basically adopt ethernet structure; calculation model based on server/client; on operating system, enterprise network's client basically is Windows platform, small and medium-sized enterprise's server commonly adopting Win NT/2000 system, part industry user or key business application of large enterprise's server adopting Unix operating system; TCP/IP protocol. The above four characteristics explain the bottom architecture,

application model, operating system and communication protocol of enterprise network all has surprisingly similar. So, enterprise information security requirement has the following common features:

1) Keeping the secret of various data

2) Keeping the integrity and accuracy of all information, data and various program in system

3) Ensure the access of legal visitors and accept the normal service

4) Ensure all works according to the standards, such as law, rules, license, contract and so on

3.3 Security problems need solving of enterprise network information system

The structure and information security requirement characteristic of enterprise network decide that enterprise network system need solve the following security problems:

1) The security problem in the LAN internal, include the division of network segment and the Implementation of VLAN

2) How to implement the security of network layer when connect to Internet

3) How to ensure the security of application system

4) How to prevent the intrusion to network, host and server by hackers

5) How to implement the security of information transmission in WAN

6) How to arrangement the encryption system, include establishing certificate management center, integrating application system and so on

7) How to implement the security of remote access

8) How to evaluate the integrate security of network system

4 Implement Technology of Enterprise Network Security

Based on the given of the above security problems, enterprise network information system commonly includes the following security mechanism: access control, security detection, attack monitoring, encryption communication, authentication, hiding network internal information (for example NAT) and so on. In order to ensure the security of enterprise network information system, there mainly have the following commonly using methods at present.

4.1 Security technology of LAN

1) Network segment

Network segment commonly be considered to one basic method to control network broadcast storm, its aim is isolating the illegal user and sensitive network resource, so as to prevent possible illegal interception. Network segment can be divided into physical segment and logical segment. Nowadays, enterprise LAN mostly adopt the network pattem which use switch as center, router as boundary, should mainly mining the center-switch access control function and three layer exchange function, comprehensive application physical segment and logical segment to implement the security control of LAN.

2) Use switching hub instead share-based hub

After network segment to the center switch of LAN, there also exists the risk of Ethernet interception. Because the access of network final users commonly through branch hub not center switch, and the widely used branch hub commonly the share-based hub. Thus, the data packet (namely Unicast Packet) between two computers will be intercepted by other users in the same hub when user and host carry on data communication. For example, user TELNET to one host, because the TELNET program lack encryption function, each character (includes user name, password etc. important information) the user key in all will be sending by plaintext, this will provide chances to hackers. So, should use switching hub instead share-based hub, make the umicast packet transmission only between two nodes, thus prevent illegal interception.

3) Division of VLAN

In order to overcome the broadcast problem of Ethernet, also can application VLAN technology, change the Ethernet communication to point to point

communicate, prevent most intrusion based on the network interception.

Under the environment of centralized network, we commonly integrate all center host system to one VLAN, in this VLAN not allow any user nodes, thus protect sensitive host resources well. Under the environment of distributed network, we can division the VLAN by the setting of mechanism or department. Each department's internal all servers and user nodes in their own VLAN，mutual nonaggression. The connection in the VLAN internal adopt exchange implement, and the connection between VLAN adopt router. Also can use external connected multi- Ethernet router instead switch implement the router function between VLAN, in this situation, the efficiency of router forwarding will decrease.

4.2 Security of network layer

The security protection of network layer is facing to IP packet, it mainly adopts firewall as the method of security protection, implements the primary security protection. It also can implement encryption protection based on some security protocol and corresponding intrusion detection.

1) Firewall

Utilization firewall can implement the isolation and access control between internal network (trust network) and external network (not trust network) or different security domain in internal network, ensure the availability of network system and network service.

At present, the mature firewall mainly have three kinds, one is packet filtration firewall, one is application surrogate firewall, also has the other kind is compound firewall, namely the binding of packet filtration and application surrogate firewall. Packet filtration firewall commonly filter data stream based on the source of IP data packet or target IP address, protocol type, protocol port number and so on, It has higher network performance and better application program transparency than other model firewall. surrogate firewall uses in application layer, commonly can surrogate to kinds of application protocol, and identification the user's identity, and provide more detailed log and audit information; Its disadvantage is all need provide corresponding surrogate program for each kind of application protocol, and the firewall based on surrogate often decrease the network performance obviously.

2) Intrusion detection system

When information security asset has the position that related to enterprise survival for some enterprises, deployment intrusion detection system will the very necessary protective measures of enterprise information security, it can make up the disadvantage of firewall. Asset detector with IDS in the firewall's internal and external can assistant judge the setting and operating of firewall whether suitable. IDS also can identification the network attack which firewall cannot perceive.

At present, most access mode of intrusion detection all adopt pass-by mode to interception the data stream in network, so this has limited the IDS itself block function, only can block part actions establishing on the foundation of TCP, such as Telnet, FTP, HTTP and so on, but inability to some actions establishing on the foundation of UDP. So, the linkage goal between IDS and firewall is more effectively block the attack event which had occurred, thus make the network hidden trouble decrease to lower limitation.

4.3 Security of application system

Application security is the security precaution established aim to specific application (such as WEB server, e-mail server and so on). Although some precaution maybe one replacement or overlap of network security business, For example, the encryption of WEB browser and WEB server to message in application layer all encryption through IP layer, but many applications all have its own specific requirement. Especially the enterprise involving business activities more intent to adopt various security measures in application layer but not in network layer. The security business in application layer can involve authentication,

access control, integrity, confidentiality, data integrity, non-repudiation, Web security and so on.

4.4 Security of host and server

The security of host and server mainly include: inspection and confirmation there has no known vulnerability (such as virus, Trojan horse and so on) in software installation; in order to make the system has minimum penetrating can adopt access control mechanism, allow access only through authentication; in management, for all access data must carry on audit, for system user carry on district security management; for intrusion carry on detection, audit, tracing and so on.

4.5 Data encryption and security mail

In the transmission and disposal of INTERNET, data encryption integrity technology is the important guarantee of information security. International encryption algorithm IDEA or data encryption standard DES suitable for the encryption of numerous secret information, Public key cyptosystem such as RSA commonly use in authentication and digital signature. E-mail can use S/MIME and PGP practical technology to guarantee security, they all following the IETF standard, S/MIME close to industrial standard, but PGP mostly using in personal security e-mail.

4.6 Integrity security

Network security should be a dynamic development process, should be a cycle process of detection – monitoring – security response. Dynamic development is the rule of system security. Network security detection is the important measures of risk assessment to network, through using network security analysis system can timely find the most weak link of network system, examination report the weakness, vulnerability and insecurity configuration exists in system, suggest remedy measure and security strategy, gain the aim of enhance network security.

5 Solving Scheme of Enterprise Network Security

According to the security requirement of enterprise network, the network security can be divided into three layers to implement solving scheme.

5.1 Basic protection system (packet filtration firewall + NAT)

1) User requirement: all or partly satisfy the following various:

2) Solving internal and external network boundary security, prevent external attack, protect internal network

3) Solving internal network security problem, isolation internal different network segment, establish VLAN

4) Filtration based on IP address, protocol type and port

5) Internal and external network adopt two set IP address, need network address exchange NAT function

6) Support security server network SSN

7) Through the correspondence of IP address and MAC address to prevent IP deception

8) Flow statistical and restriction based on IP address

9) Monochrome list based on IP address

10) Firewall running on the security operating system

11) Firewall is independent hardware

12) Firewall has no IP address

Solving scheme: adopt network firewall.

5.2 Standard protection system (packet filtration firewall + NAT + surrogate + VPN)

User requirement: all or partly satisfy the following various on the basis of basic protection system figuration:

1) Provide application surrogate service, isolate internal and external network

2) User authentication

3) Access control

4) Flow statistical and restriction based on users

5) Security management based on WEB

6) Support VPN and its management

7) Support transparent access

8) Has self protection ability, prevent the common attack to firewall

Solving scheme:（1）choose network firewall;（2）firewall basic configuration + network encryption equipment (IP protocol encryption equipment).

5.3 Strengthening protection system (packet filtration firewall + NAT + surrogate + VPN + network security inspection + monitoring)

User requirement: all or partly satisfy the following various on the basis of standard protection system figuration:

1) Network security inspection (include server, firewall, host and other TCP/IP related equipment

2) Operating system security inspection

3) Network monitoring and intrusion detection

Solving scheme: choose network firewall + network security analysis system + network monitor.

6 Conclusions

Network information security is a marginal nature comprehensive subject which involving computer technology, network technology, communication technology, password technology, information security technology, application math and so on many subject. Network security is subjected threaten have some reason, its main reason is the security vulnerability in network system and the rapid development of hacker technology; secondly also lies in human disadvantage security consciousness, not adopt effective security strategy and mechanism, and disadvantage advanced network security technology and so on. For one network's management, the important is the management of some key "point", so the network manager of enterprise need grasp the main problems which influence network security firstly. This paper analyzes the common risks, system characteristic and security requirement of enterprise network security, given an integrated stereo network security solving scheme, basic solving the security threaten of enterprise network information system caused by human factors and hardware network equipment.

References

[1] Gang Qian. Information system security management [M]. Jiangsu: Southeast University publishing house. 2004

[2] Eric Maiwald write, Qing-rong Li, Kai-zhi Huang etc. translation. Network security practical course (second edition) [M]. Beijing: Tsinghua University publishing house, 2003

[3] Xiao-feng Lin. Analysis of network security solving scheme. Science Mosaic, 2007.7: 99-101

[4] Jian Wu. Security risk and preventive measure of enterprise network. Shandong Communication Technology. 2005.6, 25（2）:19-22

Research on Text Clustering Based On Web Concept Semantic Tree

Yang Xiquan Dai Shu Zheng Dan

School of Computer Science, Northeast Normal University, ChangChun, Jilin, 130117

Email: yangxq375@nenu.edu.cn

Abstract

In this paper, the semantic tree consists of contents of How Net to eliminate the ambiguities of words and cluster semantic similar documents based on clustering of contents. we address a novel method to solve "key word obstacle" problem, namely concept semantic tree. The novel method based on semantic information of How Net makes the construction of the concept semantic tree more flexible, which can add and delete nodes easily. It also features combining relevancy to solve semantic ambiguity, which can well analyze semantics and improve the effect of text clustering.

Keywords：semantic tree; How net; semantic similarity; semantic relevancy; Web

1 Introduction

Clustering analysis is an important means of data mining, it is important in text mining. Text clustering is actually the clustering of text contents. (For example: the muti-file essay system of Biya university.) In the classical vector space model(VSM) based on text keywords, document vector Di＝{d1i，d2i，…, dmi} was composed of m keywords to state one document of the document set. But there are problems in this method. First, it takes the words as independent elements and there are no relationships between them when calculating the similarity of text vector spaces by inner product of vector. It can't clearly express the semantic meaning of the text. Second, the semantic VSM just matches the explicit words appeard in the texts, ignoring multiple meanings of a word and various expressions of text semantics.

The set of vocabulary entries can't exactly reflect the semantics of texts. But it can cluster the semantics of the texts by changing the method of text clustering. The semantic tree consists of contents of How Net to eliminate the ambiguities of words and cluster semantic similar documents based on clustering of contents. The second part of the paper explains the basic concepts of semantic analysis based on How Net. The third part analyzes the construction and query of the off-line semantic tree and the dynamic semantic tree. The fourth part is about the deducing calculation of semantic similarity and the analysis of experiments result .And the fifth part is the conclusion.

2 Semantic Analysis

There are three main parts of semantic analysis: meaning confirming, syntax analyzing of combining words, concepts similarity calculating. Meaning confirming means the analysis of concepts and the concepts (DEF) are described by sememes. As the definition of How Net, the characteristic used to definition in the DEF is at least one, but the number is not limited. But the first sememe of the DEF must be the main characteristic defined by How Net or it is taken as an error. And there is the Hypernym-Hyponym between the first sememes of DEF and there is not Hypernym-Hyponym between other sememes definitely. The sememes constitute a hierarchy like a tree, which is the basic of classification of words.

The How Net can combine the definitions and the

relative words quite well and provide a platform to construct the definition semantic tree.

3 Semantic Tree

The semantic tree[2] is composed of directed binary tree by marking a conjunction formula on each nodes.

3.1 Basic definition

Definition 1 if $\Gamma \subseteq S(F)$, then the semantic tree of Γ can be defined as follow:

i) the single node tree that the root tag is one element of Γ is semantic tree of Γ;

ii) if T is semantic tree of Γ, u is the leaf node of T and $\overline{u}$ is u or the ancestor of u, then deduce T' that is the semantic tree of Γ from T based on the following rules.

The semantic tree constructs the semantic space with tree-like models, which can overcome the disadvantage of the graph semantic space. There are two characteristics of semantic tree.

i) The elements which are used to construct the semantic tree are defined beforehand but the semantic tree is dynamically constructed when it is needed.

ii) The semantic tree dynamically built can be handled well, and it is easy to add or delete nodes.

Definition 2 The concepts are independent of languages and the words included in a sentence which are independent of languages are neither the syntax structure of the sentence nor the semantic structure, but they are the concept structure[3]. There are different relations between concepts, and the most important relation is affiliation between concepts. Above all, a kind of concepts is subject to a concept (Abstract concept).

3.2 Construction of semantic tree

There're two method to construct the semantic tree model, off-line construction of semantic tree and dynamic construction of semantic trees[4].

3.2.1 Off-line construction of semantic tree

Supposing TSim（p,q）is the similarity of concept p and concept q, and for an arbitrarily given concept p, describe the relations between it and all the other concepts by tree-like model, illustrating in Figure 1. the concept p is the root node and the word q is a leaf node, and the weight of the path between them is the similarity of the concept and the word. All the words from q1 to q r are ordered by the similarities between the words and concept p, and the word which is most similar with p is leftmost to satisfy the following formula:

Tsim（p,q_1）≥ Tsim（p,q_2）≥ ……≥ Tsim（p,q_i）≥ ……≥ Tsim（p,q_m）

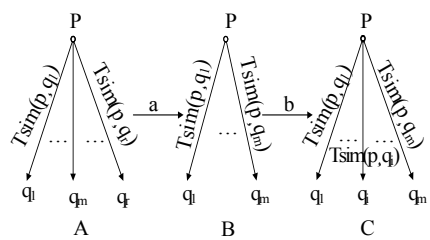


Figure 1 Off-line Semantic Tree

The method of calculating similarity is described by three trees in Figure 1 First, the leftmost m words are reserved, the others are deserted; Second, the interval of the similarity of qi and qm is（1～0.3）. Divide the m reserved words again, which are divided particularly into q_1 to q_i, q_i to q_m. And the interval of the similarity of q_1 and q_i is（1～0.7）and the interval of similarity of q_i and q_m is（0.7～0.3）. The B tree is divided here to compare the similarity of semantics further and to enhance the clustering effect of text and text set. The algorithm off-line semantic tree is as following:

In document set D, P is the concept vector space $\{p_1,p_2,\ldots\ldots,p_n\}$

buildPST(T)

Input: the query document q, and it is the vector space of word $\{q_1,q_2,\ldots q_n\}$

if $1 \le sim(p,q_m) \le 0.3$ *then;*

return q_m;

else

remove(T, q_m*);*

if $sim(p,q_i) \geq sim(p,q_m) \geq 0.3;$

$0.3 \leq sim(p,q_i) \cup sim(p,q_m) \leq 1;$

if $0.7 < sim(p,q_i) \leqslant 1$ *then;*

$0.3 \leqslant sim(p,q_m) < 0.7;$

then $PST(T');$

The off-line semantic tree is constructed accordingly to specific applications. P is the initial concept vector, which is the root node of the whole semantic tree.

3.2.2 Construction of dynamic semantic tree

It is known that the concepts are composed of sememes through HowNet. P=（$p_1,p_2,\ldots,p_i,\ldots,p_n$）is supposed to represent the initial sememe vector, and the value of n is random, and pi is the ith sememe. The CSTM (Concept Similarity Tree Model) is constructed by P, and v in CSTM (P, v, j) represents the depth of the hierarchies, and j represents every element of the CSTM (including i-best and m-best), which means that the CSTM is constructed by every root node at least connected with m leaf nodes, illustrating in Figure 2.

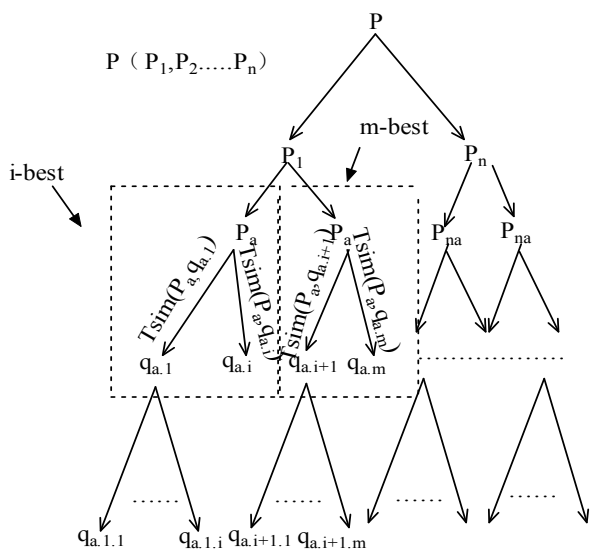


Figure 2 Dynamic Semantic Tree CSTM

The CSTM is constructed by several i-best trees and m-best trees on different hierarchies. The similarity of the root node and the leaf node can be easily obtained by CSTM.

3.2.3 The query algorithm based on semantic tree

The important step of the text research is that a set of effective subject concepts are constructed by the users' query. The key subject word takes an important role in subject concepts and the essential part is to expand the key subject words.

$$\begin{cases} sim(q,w) = \sum_{i=1}^{k} sim(q_i,w) \geq cv \\ overlay(CSTM(p,v,j),w) \geq percent \times k \end{cases} \quad (1)$$

It is supposed that the initial subject concept vector of users is P=（$p_1,p_2,\ldots,p_i,\ldots,p_n$）which includes n concepts and each concept is connected to a key subject word. The vector of key subject word is q= (q_1, q_2, …, q_j, …, q_k), which includes k words. And q_i represents the number ith word. If word w satisfies the following conditions, then w is the needed expanding query word.

Sim(q, w)is the similarity of the key subject word vector q and word w, and sim(q_j, w)is the similarity of q_j and w. The cv is the threshold value of the similarity. The value of Overlay(CSTM(p, v, j), w) is the number of the sub-trees which include word w. And "percent" is the threshold value of coverage

```
<?xml version="1.0" encoding=" gb2312"?>
 <root>
     school
    <node>
        College
      <name>
          Northeast Normal University
           <college> Computer College </college>
        </name>
     </node >
    </root>
```

4 The Relevancy of Semantics

The relevancy of semantics is an ambiguous concept. And there is no objective standard to measure. The relevancy of words is a concept which refers to the accidence, syntax, semantic diction and so on. And the most effective influence of the relevancy of words is the relevancy of semantic. The relevancy is defined as a

real number between 0 and 1.

Definition 1 The relevancy of semantic is the degree that two words of a phrase can compose a modificatory relation, subject-object relation or identical demonstrative relation.

Definition 2 In HowNet, w1 and w2 are two arbitrary words, and w_1 has n sememes, s_{11}, s_{12}…s_{1n}; and w_2 has n sememes, s_{21}, s_{22}…s_{2m}, if $s_{1i}=s_{2j}$, $1 \le i \le n$, $1 \le j \le m$, then the relevancy of w1 and w2 is 1.

4.1 The calculation of similarity

Accordingly to the relevancy of the semantic, each sememe is divided. Calculating thetransverse and lengthways relevancy can obtain more exact relation between concepts. Eq.2 is deduced from Eq.1 by semantic relevancy.

$$\begin{cases} R(p,q) = \max\left[\eta_1 \sum_{i=1}^{4} \beta_i \prod_{j=1}^{i} H_j(S_1,S_2) + \eta_2\left(1-\frac{d(t_i,t_j)}{D}\right)\right] \\ sim(p,q) = \sum_{i=1}^{k} sim(p,q_i) \end{cases}$$

$$(\eta_1 + \eta_2 = 1) \qquad (2)$$

In Eq. 2, t_i, t_j are the first basic sememe of sememe terms S_i, S_j. D is the transverse similarity, and H_j (S_1, S_2) represent the four part sememes. β_i ($1 \le i \le 4$) is the adjustable parameter,and $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$ $\beta_1 \ge \beta_2 \ge \beta_3 \ge \beta_4$ is the effect to total similarity from H_1 to H_4 and only β_1 has a bigger weight. The sim (p, q_i) is the similarity of concept p and word q_i. The proof is given as follow.

i) p and q are two sets, $\beta \in$ p, $\alpha \in$ q, if $\alpha \in \beta$, and $\alpha = \beta$, and β is the ancestor of α, then p is semi-part-of q.

ii) the attribute of p is defined as $WS_i=(I_1, I_2, \ldots I_n, O_1, O_2, \ldots O_m)$, the range of value of p is PD_i, the domain is PR_i, $PD_i \supseteq (I_1,I_2,\ldots\ldots I_n)$, $PR_i \subseteq (O_1,O_2,\ldots\ldots O_m)$; For arbitrary p_i, q_j, $\forall$ i, j, if (i, j) $\in p_i$, then (i, j) $\in q_j$, and p_i is semi-part-of q_j.

iii) Arbitrary elements p_i and q_j, $\forall$ i, j, if i $\in PD_j$, j $\in PR_i$, then i $\in PD_j$, j $\in PR_j$, and p_i is semi-part-of q_j.

According to the above proof of Eq. 2 (The above poof is deduced by definitions in reference [6]), and the similarity of the concept vector p and the word vector q can be calculated. The query document is taken as W whose vector is (w_1, w_2 … w_k). According to the calculation of the W and q, the similarity and the relevancy can be compared. When the similarity and the relevancy are both high, W is similar with q. Then calculate the similarity and the relevancy of W and p. The document query structure is given in Fig 3.

4.2 Experiment: Input an XML document

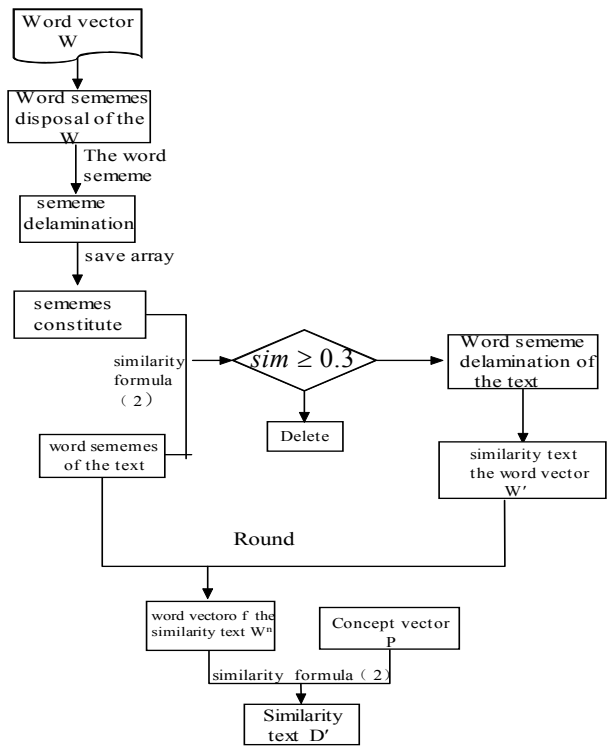


Figure 3 Document Clustering Structure

The similarity and the relevancy are compared transversely and lengthways. The concept semantic vector of the XML document tree in Fig 4, and college →Computer College is taken as an example.

college → Computer College:

Computer College $\xrightarrow{\text{similairy formula2}}$ Computer Department, similarity is 0.65

Computer College $\xrightarrow{\text{similairy formula2}}$ Computer, similarity is 0.32

Computer College $\xrightarrow{\text{similairy formula2}}$ Diannao, similarity is 0.15

While the semantic tree is constructed, the

similarity is divided further. If similarity＜0.3, then it is

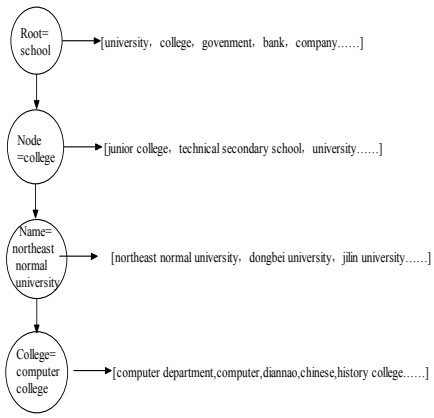


Figure 4 XML doucment

deleted. If similarity ≥0.3, calculate the similarity again ,if similarity ≥0.6 go to next step of calculation. According to the example above, the Computer Department will be returned. For the given parameters β and ∂, the values of similarities is as follow in table 1.

Table 1 Similarities by the given parameters

β_1	β_2	β_3	β_4	β_4	similarity
0.30	0.20	0.15	0.1	1.7	0.3
0.45	0.30	0.15	0.1	1.7	0.4
0.45	0.40	0.10	0.1	1.7	0.5
0.50	0.25	0.15	0.1	1.7	0.65
0.63	0.23	0.10	0.1	1.7	0.7

According to the data above, the bigger the first sememe, the larger the similarity is. The following Figure is the comparing result.

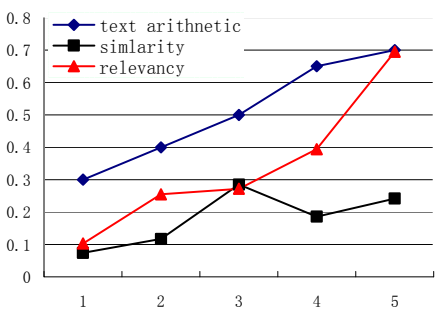


Figure 5 Data Comparability

The algorithm approached is based on the comparability of the similarities of concepts. It is obvious that the algorithm can exactly calculate the relations between concepts and words from Fig 4.

5 Conclusion

CSTM is approached here based on TSTM, which can calculate the similarity of word vector and concept vector effectively by the sememes in HowNet to cluster the text and the text set. This method can explain the words quite well and can solve the ambiguity problem by transverse and lengthways relations. The next work is to enhance the parameter adapting configuration to achieve the semantic clustering of documents.

References

[1] LU Peng, SUN Ming-yong, LU Ru-zhan HowNet-Based Word Semantic Automatic Classification System, Computer Simulation.2004

[2] Li Zhoujun, Wang Bingshan The Semantic Tableaux Method and Its Soundness and Completeness Journal of National University of Defense Technology.1994

[3] Yao Tian-shun, Zhu Jing-bo Comprehension of Nature Language-a Study on human language which can be understood by machines

[4] Zhao Jun, Jin Qian-Li, Xu Bo Semantic Computation for Text Retrieval Chinese Journal of Computers, 2005

[5] Xu Yun, Fan Xiao-zhong, Zhang Feng Semantic Relevancy Computing Based on HowNet Journal of Beiing Institute of Technology, 2005

[6] Cui Juntao ,Liu Jiamao ,Wu Yujin, Gu Ning,An Ontology Modeling Method in Semantic Composition of Web Services.Department of Computing and Information Technology

[7] Chen Ning, Chen An, Zhou Long xiang. An Incremental Grid Density-Based Clustering Algorithm.Journal of Soft-ware, 2002

[8] Frigui H, Krishnapuram R. Clustering by Competitive Agglomeration [J].Pattern Recognition, 1997, 30（7）:11091119

[9] A. K.Jain, R.C.Duties, Algorithms for Clustering Data.NJ: Prentice-Hall, 1988

[10] R. Kosala and H. Blockeel, Web Mining Research: A Survey. ACM SIGKDD Explorations Newsletter.vol.2, no.1, pp. 1-15, 2000

A Scheme of Integrated RSVP for QoS Support in Mobile IPv6 Network*

Gang Nie[1] Lei Li[2]

1 College of Computer Science, Wuhan University of Science & Engineer ,Wuhan, Hubei ,430073, China

Email: ng@wuse.edu.cn

2 School of Computer and Technology, Wuhan University of Technology ,Wuhan, Hubei ,430070, China

Email: lilei_lisa@sina.com

Abstract

Providing Quality of Service (QoS) guarantees and mobility support for Internet devices has become a hot research topic in the Next Generation Internet technologies, since mobile computing is getting more widespread. The Internet Engineering Task Force (IETF) has developed ReSerVation Protocol (RSVP) and Mobile IPv6 to provide IPv6 QoS and mobility support, respectively. However, an integrated and efficient interworking of these two mechanisms is still not present. In this paper, we propose a model based on flow transparent to solve this problem. We examine this desired model and illustrate how it overcomes the problem with the existing approach and achieves a more efficient RSVP and Mobile IPv6 integration. The analysis show that the proposed scheme works well in mobile environments.

Keywords: QoS; RSVP; Mobile IPv6; Scheme

1 Important information

With the rapid increasing number of portable devices, such as laptop computers, palmtops, and Personal Digital Assistants (PDAs), and the emergence of a variety of wireless access technologies (Bluetooth, Hiperlan2, WLAN, UMTS, etc.), mobile computing applications may become more practical. Real-time services such as Internet telephony, video conferencing, Video-on-Demand, in mobile environments also should be realized. Therefore it is important for the Internet to provide QoS guarantees and mobility support in the near future. The Internet Engineering Task Force (IETF) has standardized the ReSerVation Protocol (RSVP) [1] and Mobile IP [2] to support Internet QoS and mobility, respectively. In order to support QoS and mobility simultaneously, there is a need to integrate RSVP and Mobile IP.

To provide simultaneous QoS and mobility support for wireless real-time services in IPv6 environment, a number of studies on interworking of these two protocols were proposed [3, 4, 5]. For examples, M. Thomas et al. [6] proposed an RSVP and Mobile IPv6 [7] integration model. Under their model, resources are initially reserved between the Correspondent Node (CN) and Mobile Node (MN)'s original location. Whenever the MN changes its location, which incurs a path change, an RSVP signaling needs to be performed end-to-end between the CN and MN's new location to reserve resources for the new path. A problem with this model is long resource reservation delays and signaling overheads incurred during handoff [8]. Each time an RSVP renegotiation has to be performed end-to-end no matter how significant the handoff affects the path between CN and MN. Before this RSVP renegotiation completes, service degradation could occur due to lack of QoS guarantee in the newly added portion of the path between the CN and MN.

In this paper, we propose a method to automatically

* This research was supported by the Educational Ministry of Hubei Province, China under Grant D200717005.

limit the handoff RSVP renegotiation process within the newly added portion of the path between CN and MN. Thus, handoff resource reservation delays and signaling overheads can be minimized which in turn minimizes the handoff service degradation.

The rest of the paper is organized as follows. Section II identifies the problem in the existing IPv6 QoS with mobility support model. Section III introduce and analyze our proposed IPv6 QoS with mobility support model. Finally section IV concludes the paper and presents future work.

2 Problem in the existing IPv6 QoS with mobility support model

The existing IPv6 QoS with mobility support model is shown in Figure 1 [9]. A router is responsible for multiple cells. An IP level handoff only takes place when the MN crosses two cells which belong to different subnets. The Mobile IPv6 and RSVP interworking can be illustrated with a typical wireless mobile Internet telephony application. Both telephony parties are mobile and have wireless access to the Internet. A major challenge in this model is the handoff problem. As far as handoff is concerned, we may consider one party as CN and the other party as MN without losing any generality because during a telephony session, both parties function symmetrically, i.e., both as Sender and Receiver.

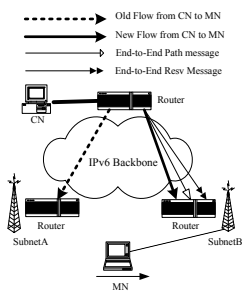


Figure 1 Existing IPv6 QoS with Mobility Support Model

The main idea of Mobile IPv6 and RSVP inter working is to use RSVP to reserve resources along the direct path between the CN and MN without going through their Home Agents since Mobile IPv6 has Route Optimization [10]. Whenever the MN performs a handoff which incurs a path change, a new RSVP signaling process must be invoked immediately to reserve resources along the new path, instead of waiting for the next periodic RSVP state update associated with RSVP Soft-state mechanism. In Figure 1, when the MN performs a handoff from subnet A to subnet B, it obtains a new care-of address and subsequently sends a Binding Update to the CN. The CN then triggers a Path message associated with the new flow from CN to MN. Upon receiving this Path message, the MN replies with a Resv message immediately to reserve resources for the new Flow.

In this model, the RSVP Session and the flow destination are identified by the MN care-of addresses. Therefore, both Session and flow identity changes whenever the MN obtains a new care-of address during handoff although the application level session is the same. For each handoff, the MN as receiver has to wait for a new Path message from the CN and only after that it can issue a new Resv message to the CN. All these RSVP renegotiations have to be conducted end-to-end even though the path change may only affect a few routers within the whole path during a single handoff. The long handoff resource reservation delays and large signaling overheads caused by this end-to-end RSVP renegotiation process could lead to notable service degradations because during this period, there might not be enough resources in the newly added portion of the new path between CN and MN.

For interactive services such as Internet telephony, the MN acts as both Sender and Receiver. It is not difficult to extend the above model to accommodate Sender mobility as well. When MN is functioning as a Sender and performs a handoff from subnet A to subnet B, it obtains a new care-of address and consequently sends to the CN a Path message associated with the new flow from MN to CN. Upon receiving the Path message, the CN replies with a Resv message to reserve resources for the new flow.

Unfortunately, this process suffers exactly the same problem as when MN is a Receiver. Since change of MN care-of address during each handoff changes the flow source identity, multiple network layer flows are

required to support one application data flow during node mobility. RSVP renegotiation must be performed end-to-end during each handoff which causes long resource reservation delays and large signaling overheads.

3 Proposed scheme

In the previous section, we identified the problem in the existing IPv6 QoS with mobility support model, namely, long resource reservation delays and large signaling overheads are introduced during each node handoff which could cause notable service degradation. The main cause of this problem is that RSVP and Mobile IPv6 are originally developed independently and thus lack of intrinsic collaboration; a direct combination of the two can not lead to optimized interworking. Consequently, a desired model should be able to minimize handoff resource reservation delays and signaling overheads through a more efficient integration of RSVP and Mobile IPv6.

3.1 Flow Transparency Concept

Node mobility incurs change of node address and in turn changes of flow identity with MN as source or destination. This results in the same application data flow being perceived as different flows at the network layer. Since router processing needs to be based on flow, each time a flow change at the network layer makes it necessary to rebuild information in all intermediate routers along the flow path. A natural idea of solving this end-to-end renegotiation problem is the routers which reside in the common portion of the new and old flow path should be exempted from performing handoff update; and only those routers that are in the newly added portion of the flow path need to be involved in the update process. This requires the underlying mobility support protocol to keep the node mobility completely transparent to the network layer flow handling mechanism. Similar to the Transport Layer Protocol Transparency concept which keeps node mobility invisible to transport layer protocols, we define this transparency at the network layer as Network Layer Flow Transparency which is illustrated in Figure 2. The figure shows only the flow from CN to MN, Flow Transparency for the opposite flow from MN to CN can be similarly achieved. It is important to note that with node mobility, the number and identity of routers involved in the same flow are dynamic and usually unpredictable. Only the routers that are common to both new and old path of the flow constitute the scope of Flow Transparency. This implies that an automatic flow handling adaptation mechanism for those routers in the newly added path is also required in order to exploit the Flow Transparency concept. The major advantage of Flow Transparency is that it allows the network layer flow handling mechanism to function normally regardless of node mobility which in turn brings about performance improvements for mobile QoS mechanism. The essence of providing Flow Transparency is to maintain a unique flow identity irrespective of the change of MN address.

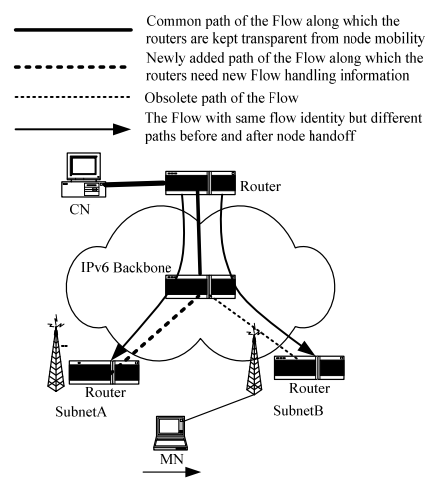


Figure 2 Flow Transparency Concept

3.2 Desired IPv6 QoS with Mobility Support Model

With the Flow Transparency concept defined, we propose a desired IPv6 QoS with mobility support model based on RSVP and a flow transparent Mobile IPv6. In

this section we describe the basic mechanism of this model. A flow transparent Mobile IPv6 always keeps a constant flow source and flow destination for a specific application data Flow so that the RSVP Session and the network layer flow identity are constant for the router flow handling mechanism (e.g., the packet classifier) regardless of node mobility. Figure 3 shows a similar architecture as in the existing model and the MN performs a handoff from Subnet A to Subnet B. Two scenarios are discussed. In the first scenario, the MN acts as a Sender, and in the second, the MN acts as a Receiver.

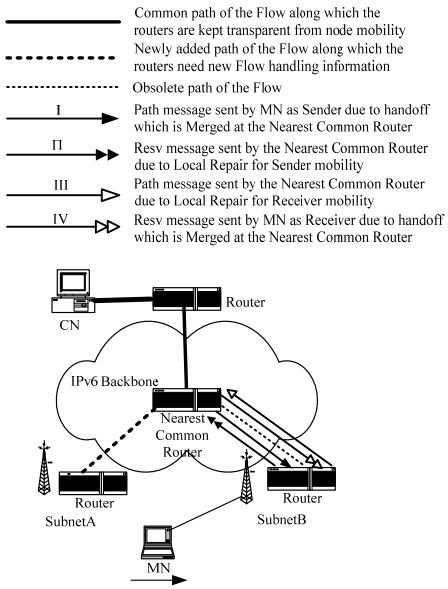


Figure 3 Desired IPv6 QoS with Mobility Support Model

Scenario 1: Mobile Node As Sender: If the MN is acting as a Sender, it immediately triggers a Path message (Message I) to the CN – with the same source flow identity as the one before handoff. According to the Merge functionality of RSVP, this Path message will be merged at a router where there is already a Path state for that flow which is created during previous RSVP message exchanges. In this case, the router where merge occurs is also the nearest router common to both the old and new path (Nearest Common Router) for the flow from MN to CN. It sees the same Path message arriving with a previous hop address that differs from the one stored in the original Path state. This is exactly the condition RSVP needs to trigger a Local Repair for Sender route change, which is actually due to Sender mobility. So it will immediately send a Resv message (Message II) associated with that flow to reserve resources along the newly added path to the MN. It is important to note that all handoff Path and Resv message exchanges only involve routers within the newly added path and only these routers need to perform handoff update.

Scenario 2: Mobile Node As Receiver: The situation becomes more complicated when the MN is acting as a Receiver because RSVP functions asymmetrically for Sender and Receiver. Firstly, RSVP does not allow a Receiver to send Resv message before the associated Path message is received while it is not desirable to wait for an end-to-end Path message from the CN during each handoff. Secondly, although RSVP can detect Sender route change and trigger Local Repair for Sender automatically, it relies on extra mechanism to detect Receiver route change to trigger Local Repair for Receiver. The first problem is solved by letting the Nearest Common Router issue a Path message to the mobile Receiver. With the flow transparent mobility support, the information required for this Path message already exists in the Nearest Common Router during the previous RSVP message exchanges. Solving the second problem needs some minor extensions to RSVP. The Receiver should be able to inform the Nearest Common Router of its handoff information, which contains the flow destination and the MN's current address. The flow destination identifies the RSVP Session and is used to determine which Path message to send and the MN current address is used to determine where to send the Path message.

This handoff information may either be carried in a separate message if MN acts solely as a Receiver, or it can be piggybacked in the Path message sent by MN itself for Sender mobility if the MN acts as both Sender and Receiver. In both cases, the message containing this information only needs to traverse as far as the Nearest Common Router where there is existing RSVP state information for the flow from CN to MN. Figure 6 shows the second case, the MN mobility information is piggybacked in the Path message (Message I) sent by MN due to Sender mobility. Upon receiving this Path message,

the Nearest Common Router will trigger a Local Repair mechanism for Sender mobility as described above, i.e., sending Message II. Then it also triggers Local Re-pair for Receiver route change which is actually due to Receiver mobility. Specifically, it immediately sends to the MN a Path message (Message III) for the flow from CN to MN. Upon receiving this Path message, the MN replies with a Resv message (Message IV) to reserve resources along the newly added path. Again because of RSVP Merge functionality, this Resv message will not be forwarded farther than the Nearest Common Router since there is already a reservation for this flow made from that router onwards during the previous RSVP message exchanges. Hence, all the Path and Resv message exchanges due to Receiver mobility only involve routers in the newly added path and only these routers need to perform handoff update.

In conclusion, the Flow Transparency concept enables the above model to automatically limit handoff update only to the routers in the newly added path by exploiting two existing RSVP functionalities, Merge and Local Repair. The handoff RSVP signaling messages are nicely merged at the appropriate place to minimize signaling overheads; and the Local Repair operation enables a fast RSVP response to handoff and results in minimized handoff resource reservation delays. As a consequence, this model obtains a natural RSVP mobility adaptation through a more efficient integration of RSVP and Mobile IPv6 than the existing one which requires end-to-end RSVP renegotiation during each handoff.

4 Conclusions and future work

In this paper we have examined a fundamental requirement for accommodating real-time services in mobile IPv6 network, specifically, the need for integrated QoS and mobility support in a mobile environment. The problem with RSVP and Mobile IPv6 interworking in the existing IPv6 QoS with mobility support model is identified, i.e., although normally a handoff only affects a few routers, an end-to-end RSVP renegotiation must be performed for each node handoff. This problem causes unnecessarily long handoff resource reservation delays and large signaling overheads which could lead to notable service degradations. In solving this problem, we have introduced a Flow Transparency concept which requires the mobility support scheme to provide a constant flow identity for an application data flow at the network layer regardless of node mobility. We have illustrated that a flow transparent mobility scheme is essential for a desired IPv6 QoS with mobility support model. This model exploits existing RSVP features to obtain a natural RSVP mobility adaptation and achieve minimized handoff resource reservation delays and signaling overheads, thus results in a more efficient integration of RSVP and Mobile IPv6. The future work includes performance studies through simulation and comparisons with the existing model.

References

[1] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, “Resource ReSerVation Protocol (RSVP) Version 1 Functional Specification”, RFC 2205, September 1997

[2] C. Perkins, “IP mobility support,” RFC 2002, IETF, October 1996

[3] M. Thomas, “Analysis of Mobile IP and RSVP Interactions”, INTERNET-DRAFT, February 2001.

[4] H. Chaskar, R. Koodli, “A Framework for QoS Support in Mobile IPv6”, Internet Draft, work in progress, Mar 2001

[5] Stefan Schmid, “RSVP Extensions for IPv6 Flow Label Support”, INTERNET-DRAFT, August 1998

[6] Q. Shen, A. Lo, W. Seah, “Performance Evaluation of Flow Transparent Mobile IPv6 and RSVP Integration”, Florida USA: Proc. SCI/ISAS 2001, July 2001

[7] D. Johnson and C. Perkins, “Mobility support in IPv6," IETF Internet Draft, April 2000, work in progress

[8] Terzis, A., Srivastava, M., Zhang, L., “A Simple QoS Signaling Protocol for Mobile Hosts in the Integrated Services Internet,” Proceedings of IEEE INFOCOM, Vol. 3, p. 1011-1018, March 1999

[9] Lopez, A., Vlahos, H., Manner, J., “Reservation Based QoS Provision for Mobile Environments,” 1st IEEE Workshop on Services and Applications on the Wireless Public Interface, Volume 7, July 2001

[10] Chaskar, H., Koodli, R., “QoS support in Mobile IP version 6,” IEEE Broadband Wireless Summit, May 2001

Personalized Recommendation of Campus Network Educational Resources Based on Collaborative Fiterring*

Junwei Li Qing Yang Yuying Huang

Department of Soft Engineering and Information System, Center China of Normal University, Wuhan, Hubei, 430079, China

Email:lijunwei1228@qq.com

Abstract

As campus-network has popularized in universities, students can get lots of educational resources from the network. It is getting harder and harder to find the interesting and suitable resources among these massive resources because of so many faculties and amount of knowledge. How to recommend better educational resources to students becomes a critical issue. This paper adopt the collaborative filtering technology to recommend the suitable resources to students. Thereby increase the interest of students to visit the network, wide their knowledge and advance the education quality of the university.

Keywords: collaborative filtering; personalize; recommendation; campus network

1 Introduction

Getting educational resources form network is wide spread in universities. Many universities have established their own campus-network. Students in campus can look for their interested resources from website. These universities are all pay attention to put more and more resources to the website but neglect how to recommend the resources to students, especially for the freshman, they know little about the computer and network, so, it is hard for them to find the suitable knowledge among the large number of resources if there is no recommendation to them. They will lose interest on getting knowledge from network gradually, then many resources are waste. Putting useful resources on website and using good recommendation system are both important in building up website. It can increase the students' interest and the visit rate of the website, also better for the university's development in future.

Collaborative filtering is a technology that successfully and widely used in online shop and e-commerce. Such as Amazon, Netflex and so on, they all make grate achievement of using the technology of collaborative. Some domestic e-commerce companies also adopt this technology. Such as China-Pub and Dang-Dang web. In the Ref.[4], Collaborative Filtering technology was used in e-commerce's online auction system to recommend auction goods to users who is browsing the website. But this recommended method is recommend goods to single user, it is one-to-one service.. We can classify students into different groups in campus network because of their specific structure in university. Students in same group have similarity interest and needs the same educational resources, so, the recommendation we introduce in this paper should recommend for a user group not only a user. The traditional recommendation technology is also for one user. Therefore this study proposes a method to recommend resources for a group by adopt collaborative filtering combine with the characteristic of campus network.

* This research was supported by National Society Science foundation of China (No.07BYY033)

2 Collaborative filtering technology

2.1 Basic conception and theory

Collaborative filtering suppose the user is human of the society, his interest will be affected by other people. The system will follow the tracks of all users and classify the users into different categories by there interest. Similar users will share the evaluation of same information [3]. Its basic idea has two sides, one is the users who have similarity interest may interest in same resources, the other is the user may interest in resources that relate to the things that he had concerned before [5]. A problem in collaborative filtering we are going to solve is building a recommendation system in campus network to recommend educational resources. Once the students visit website, we can show him the educational resources that he might be used actively and he can choose the best. There are two key points of the recommendation system in this paper, one is students' interest, the other is others' interest that is similar to this student.

2.2 Main points for Collaborative Filtering

Collaborative filtering technology has two solutions, one is user-based and the other is content-based.

User-based collaborative filtering: Classify the users who have similarity interest into groups. The resources recommend to a user is based-on other students' interest. High interest content will be recommended.

Content-based collaborative filtering: Resources similar to the ones that user preferred in the past will be recommended. It needs to get the relativity of the resources first.

As we know, different students in different major preferred different resources in universities. Students major the same subjects substantially have the same interest. So we can use user-based collaborative filtering. But there are some students who are interested in other majors. So we can use content-based collaborative filtering. Recommend for student by analyze the history of his visit records and the rate he give to the resources. Therefore this two kinds technology will be combined to design the recommendation system.

3 User group classification and res ources recommendation

3.1 Similarity evaluation

Similarity evaluation is to evaluate the similarity between a student who is visiting the web and other students. We can get the neighbors of current visiting user. Then we can classify the users into different part who have similar interest with each other. There are some methods of similarity evaluation. Cosine similarity is one of these methods which be used widely. It solves the problem that similarity can not rely on the number of 0.

Definition: suppose x an y are two documents vector, then

$$\cos(x, y) = x \bullet y \Big/ \|x\|\|y\| \qquad (1)$$

On the above, " $\bullet$ " means vector dot product, $x \bullet y = \sum_{k-1}^{n} x_k y_k$, $\|x\|$ is the length of x, $\|x\| = \sqrt{\sum_{k-1}^{n} x_k^2}$.

3.2 Classify students' group

It is easy to classify the students in university. We can get a tree-structure graph which have six levels. It was shown in Figure 1.

After classified by Figure 1, Every student can be given an only ID number which is got by formulate in Figure 2

For example, we can give a ID number 00214050945 to the student in level 6's Class, then we can calculate the similarity between different users according to their ID number. Seen from the Figure .1, students in level 4 and have the same major and grade have high similarity. It means these students will in a same group and suppose this group is $G = (S_1, S_2 ..... S_n)$, S_i represents a student in the group. There is another problem that how to classify the students who have interest on another major, it can be solved by three steps below.

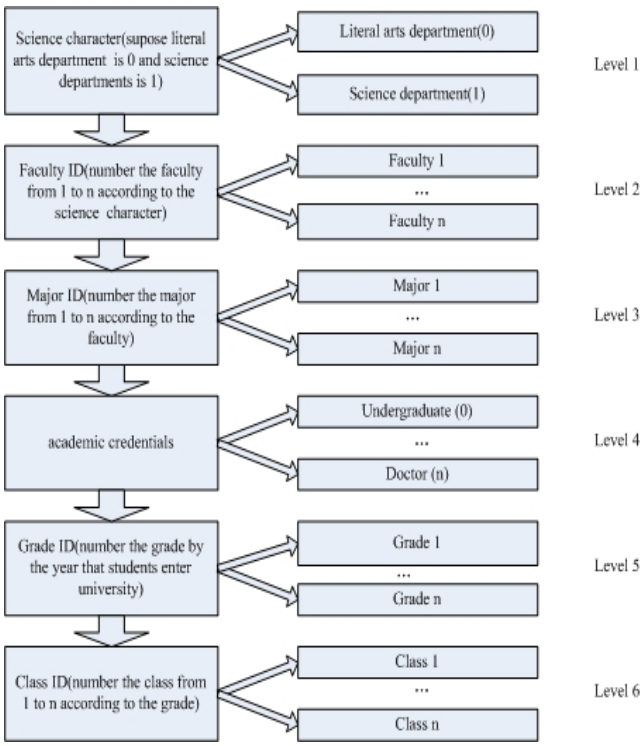


Figure 1 Classify of the students in the university

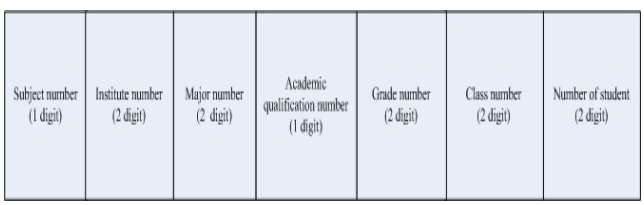


Figure 2 Formulate method of Student ID number

① Get the score matrix. Students have to score resources they have concerned before. Time is an important factor, users' interest may changed after a long time, so, set a time threshold value T, during the time T, suppose there are m strips education resources. We can get the evaluation matrix of the recently n times .the matrix is $\begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$, $x_n = (a_{n1,} a_{n2} \dots .. a_{nm})$ .Then evaluate the group of average score $\bar{x} = (\sum_{i=1}^{n} a_{i1} \Big/ n$, $\sum_{i=1}^{n} a_{i2} \Big/ n \dots .. \sum_{i=1}^{n} a_{im} \Big/ n)$ of the m trips resources.

② Calculate the group of average score $(\bar{x_1}, \bar{x_2} \dots .. \bar{x_n})$ of m pieces of information.

③ Get similarity. Use the cosine similarity of （1） to calculate the $\cos(\bar{x}, \bar{x_n})$. First, set a standard range of similarity, if cosine value is among the range, it can be in the same group. G can be extended to $G = (S_1, S_2, \dots .. S_n, T_1, T_2 \dots .. T_m)$ after we get $G' = (T_1, T_2 \dots . T_m)$, T_i is the student not in the same grade with the current user.

Example: The Current user pay attention to 5 pieces of information 4 times in recent 10 days. Then we can get the evaluate matrix $\begin{bmatrix} 4,5,2,0 \\ 5,1,3,0 \\ 2,2,3,0 \\ 1,0,0,0 \end{bmatrix}$ of the user. Calculate the group of average grade $\bar{x} = (3,2,2,0)$ of the 5 pieces of information. In the same way, calculate other users such as B, C. The group of average grade of B,C are $\bar{x_B} = (1,4,0,3), \bar{x_C} = (0,3,5,2)$ .so, $\cos(\bar{x}, \bar{x_b}) = 0.524, \cos(\bar{x}, \bar{x_C}) = 0.629$,therefore, it is clear that B has the similar interest with current user, they can be put in a same group.

3.3 Information recommend

It can recommend resources to student after ascertain the neighbors of the user.. The traditional recommendation is to one user. This paper proposes a recommendation based on a group. The resources may recommend to a group who are similar to each other. Of course, all these work should do based on one user recommend. Suppose the similar interest group of current students is

$G = G_1 \bigcup G_2, G_1 = (S_1, S_2, \dots .. S_n), G_2 = (T_1, T_2 \dots .. T_m)$ ' S_i, T_j are present to be a student. Recommendation can be divided into two parts. G_1 is a group that `have highly same interest with current user. G_2 is the group that have same interest with current user but not sure whether have same interest with the student in G_1 .so, we have to consider two parts when we recommend.

（1）Consider G_1. This group is the students who

have the same first 9 ID number with current student. History evaluation of the resources are precondition of how to recommend, Different from traditional TOP-N recommendation, we put forward a new TOP-N recommendation method. Calculate the average grade group of the resources which is evaluated by the group, then select the TOP-N high grade information to recommend to this group.

（2）Consider G_2. This group is the students who have different first 9 ID number with current student. We only consider the current user because group G_2 only has the same interest with him. Get the average and use the TOP-N recommendation to recommend information to the students in G_2.

According to the two parts on the above. A student can get two parts resources recommend when he login the website, one is resources same to current student's major get by method (1).The other is different to his major by method (2).So, when we design the website of the education resources, we can set up two special columns. Once the student login the website, he will get the information he want quickly and accurately. The efficiency is increased and make the website personalized.

4 Conclusion

This paper use the collaborative technology combine with the special characteristic of the university's student structure, and make some improve on the traditional similarity evaluation, Propose a method to recommend educational resources to the students .It can increase the interest of students to visit the campus-network to get the resources and wide there knowledge. The target user what we recommend to are students in university. We hope that the teacher can also get the information from the campus-network. There is another problem that what we recommend to students is the educational resources. In fact the students' life is niche and colorful nowadays. They also want to know some life and entertain information. If we want to recommend these resources to students quickly and accurately, we need to improve on the recommend method in this paper. All the above problems will be discussed in the later work.

References

[1] Thorsten Joachims, Dayne Freitag, Tom Mitchell. WebWatch: A personalized Recommender System Based on Explanation Facilities Using Collaborative Filtering,The Fourth International Conference on Electronic Business, 2004. 382-387

[2] Yen-Liang Chen, Li-Chen Cheng, Ching-Nan Chuang. A group recommendation system with consideration of interactions among group members. Expert Systems with Applications .2008:2082-2090

[3] Xu Xiaolin, Que Xirong, Cheng Shiduan. Information Filtering and User Modeling.Computer Engineering and Application.2003,9:182-184

[4] LI Xuefeng,LIU Lu,ZHANG Zhao; Recommendation of Online Auction Items Based on Collaborative Filtering; Computer Egineering,2006,23:18-20

[5] Zhang Fuguo. A survey of E-Commerce Recommendation System Based on Collaborative Filtering.science square, 2006, 8:7-9

[6] CHEN Xianhong,SHEN Jie,GU Tianzhu,WU Yan,ZHANG Shu,LI Hui。Collaborative Filtering Based on Users' Underlying Preference Model,2007,4: 42-44

[7] LI Junhuai,SUN Jian,ZHANG Jing,LIU Lijuan。Method of Personalized Portal Construction Based on Web-page Fragmentation.Computer Engineering,2001,1:14-18

[8] YUE Xun, MIAO Liang,GONG Jun-hua,YUE Rong。Personalized Recommender Systems Based on "Matrix Clustering" for E-commerce. MINI-MICRO SYSTEMS, 2003,11:1922-1926

[9] http://book.csdn.net/bookfiles/327/10032713191.shtml

[10] http://articles.e-works.net.cn/oa/Article42868.htm

[11] Mei-Hua Hsu. A personalized English learning recommender System for ELS students. Expert Systems with Applications .2008:683-688

Design and Implementation of WebGIS in Coal Mine Excavating and Joining

Kaixing Wu Li Liu

College of Information and Electrical Engineering, Hebei University of Engineering, Handan, Hebei, 056038, China

Email: llziyou123@163.com

Abstract

With the analysis of user requirements about the coal mine excavating and joining, the overall framework based on ASP.NET platform and the technology of WebGIS was explored in the paper. Some key techniques such as components, SSO (Single Sign On) and so on are used in the developing course. Multi-layer distributed application model and OO (Object Oriented) thoughts are adopted to design each model. The functions of the system such as Web browsing of map data, thematic map making, analyzing the excavating and joining, managing the maps, the exchange visits of graphics and database, etc. are realized. It's an example used for reference to the application of coal mine excavating and joining.

Keywords： WebGIS; Coal Mine; Excavating and Joining; ASP.NET; Architecture

1 Introduction

With the development of Internet and Web technology, WebGIS, which processes geographical spatial data on Internet, is developing rapidly[1], it's convenient to issue and share the information of maps based on WebGIS [2].

However, WebGIS faces some problems such as the difficultly of data transaction, the complexity of analyzing the map data. Most of WebGIS products only provide functions of graphical display and query, such as zoom in, zoom out, roam, properties query and so on. It's hard to clearly know the rate of progress about excavating and to display the maps in coal mine excavating and joining. It requires high configured system and provides low efficiency to use these products.

As the technologies of ASP.NET and WebGIS are becoming more and more mature, a new technique to realize WebGIS information system comes on stage. This technique makes controllable and visual WebGIS become possible. This paper sets forth how to implement the applicable and popular coal mine excavating and joining information system [3] based on WebGIS in the platform of ASP.NET with SuperMap.

2 Overall Design

2.1 Requirement analysis

With the analysis of user requirements about the functions, the workers who work at the control center of the coal mine can query the information such as the maps and the data of the working scene during excavating and joining, the operations of the interrelated equipment, the schedules of excavating and joining, examining the best path between the wells and the laneways, providing the decision-making and services for the excavating and joining mine, managing the maps, etc.

2.2 Overall framework design

The framework of the Microsoft.Net provides constructing the Multi-layer application for the solid platform. The applications sever [4] is divided into 3 layers: the user interface layer, business logic layer, data access layer. It simplifies the complex issues about the enterprise application systems development, deployment

and management, providing a great convenience for the enterprise server.

SuperMap IS. NET is a development tool based on the technology of Microsoft. NET and SuperMap Objects component, allowing developers to create a wide range of WebGIS maps, data and applications, providing these results for the users. Ajax Map using Ajax technology encapsulates the map controls in order to create a richer, more dynamic Web user interface. In practical application, as the compatibleness with COM technology, the programming languages such as Visual Basic, Visual C + +, C #, etc all can use SuperMap IS.NET to develop again.

The overall technical programme based on the above analysis of basic user requirement with .NET architecture and the platform of SuperMap IS.NET is determined as the follows.

(1) Physical logic structure

Fiber network is used as the main network road, realizing of a three wide-area network: the mine control centre, the main work site, the subsidiary work site. The application server and database server which are set up in the coal LAN, using slap-up PC servers which not only can complete the operation of all the electronic map task in the coal mine, but also can unified manage the digital map data.

(2) Software architecture

According to the physical logic structure, the software architecture about the system which uses .NET three-tiered distributed framework and the platform [5] based on SuperMap IS .NET has been designed as the following says.

Generally speaking, the three-tiered distributed software architecture is divided into the user interface layer, the business logic layer and the data layer. In this system, we intend to build the multi-layer distributed framework based on the three-tiered framework in details. The detailed descriptions about the software architecture are designed as follows.

Map information services are oriented to the three types of the workers who work at control centre, the main work site and the subsidiary work site. The realistic system is running on the WAN, using the designed multi-layer architecture about the customer/user interface layer/business logic layer/data layer.

The multi-layer distributed coal mine excavating and joining system based on WebGIS and ASP.NET platform are designed as show the Figure1

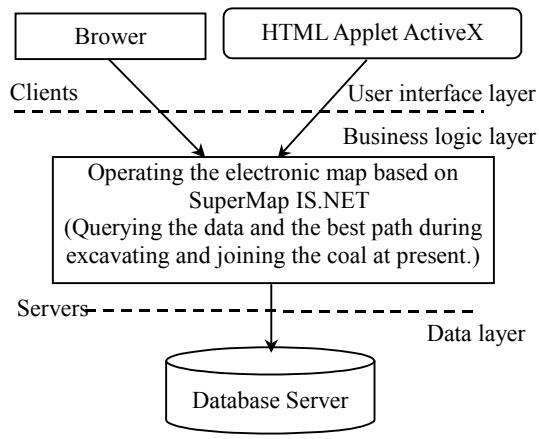


Figure1 the Software Architecture

The clients are the customers based on the IE Web browser. The users can browser the information about the system freely with the corresponding authorities given by the administrator.

The user interface layer is mainly ASPX pages, directly providing the visual pages such as the coal mine maps, information about the new and so on. It is convenient for the users to use the ASP.NET submitting the HTML pages to the client.

Business logic layer which is based on SuperMap IS.NET and made up of some class documents and business logic programmes mainly provides electronic maps for publishing and gives the related services, such as laneway, well and other enquiries, the optimal paths, etc.

Data layer is made up of the classes of the database management and modules, running on the database server of the system, using the ADO.NET technology to fleetly and safely complete the communications with the background database and to achieve the query of the spatial electronic map data and the attribute data deposited in the database.

Application server and database server use slap-up PC server with a dual hot backup solution to ensure the robustness and the stability of the system.

(3) Technical programme characteristics [6]

.NET architecture is divided into customer, Web layer, the business layer and the database layer. Only browser is installed at the customer, other features are achieved in the server to ease the workload of the development and the latter maintenance.

The user interface is consistent and friendly by the way of using the connection between the browser and Web business application server without changing the software and the procedures.

It is much safer to put the programmes about the business components in the LAN at the Web/ business applications server. The maintenance men will only redeploy the component programmes which are changed and expanded to the Web/business applications server without affecting the browse.

The system adopted the multi-layer architecture will also can use a multi-server to balance the loads with the expansion of the system and increase in users. Taking putting the Web layer and the service layer in the two servers for example will improve the response speed.

It is flexible to integrate .NET and SuperMap in the overall design. Establishing the multi-layer distributed architecture based on the WebGIS fully embodies the trends that the modern information management system is maintainable, secure and scaleable.

3. System Implementation

3.1 Application function implementation

Through the detailed analysis, the system function is divided into 8 modules as shown in Fig. 2. The information management system about the coal mine excavating and joining has been designed in Chinese. The main functions of the system are introduced as the follows:

(1) Basic operation

The maps of the system can be freely zoomed in and zoomed out in the way of clicking the position or holding down the mouse-drag a rectangular frame. The maps can be dragged freely, displaying the information in the direction of roaming. All the mine electronic information maps can be displayed with a variety of colors indicating the various wells, laneways and so on. The attribute information can be accessed while the mouse is clicking the vector graphics such as wells, laneways, facilities, measuring points and so on. The system can automatically calculate the distance and area while the mouse is clicking a designated straight line or region on maps. All the electronic maps can be saved and printed.

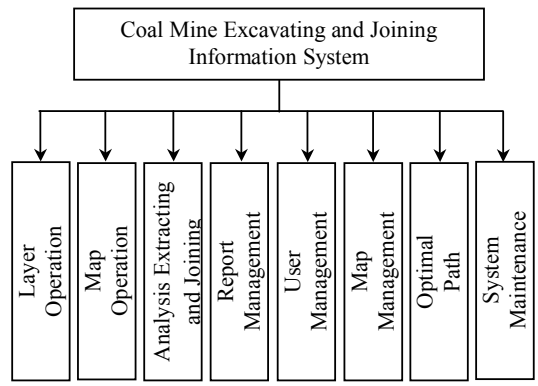


Figure2 System Function Division

(2) Eagle-eye view

There is an eagle-eye view in the upper right corner in the Fig.3. The red box in the region shows the current visual range of the electronic maps. The showing region can quickly switch to zoom in the electronic map in the eagle-eye view box while the mouse is clicking the eagle-eye view.

(3) Query and position

Information querying is the key, according to their respective targets in the different layers, the information such as finding and locating match with the key words input by the users can be find.

(4) Map editing

The operation provides the users for the functions of drawing the points and lines accurately online on the Internet, so that the users can edit the maps twice freely with the conveniences of real-time recording the information about the work progresses and updating the maps.

(5) Excavating and joining analysis

The progress is designed as follows: Inputting the planed time about planned to excavate and join the

laneway, confirming the laneway information, accessing the name and the type of laneway, supporting manner, cross sectional area, Excavating-Joining teams, inputting the two parameters of the total footage and single footage, getting the days and the time completed the laneway, judging of the interval whether or not less than three months, drawing the network map, outputting the laneway map and related information at the end. This progress is repeated until the requirements are over. The flow chart of the algorithm processes is illustrated in Figure.3.

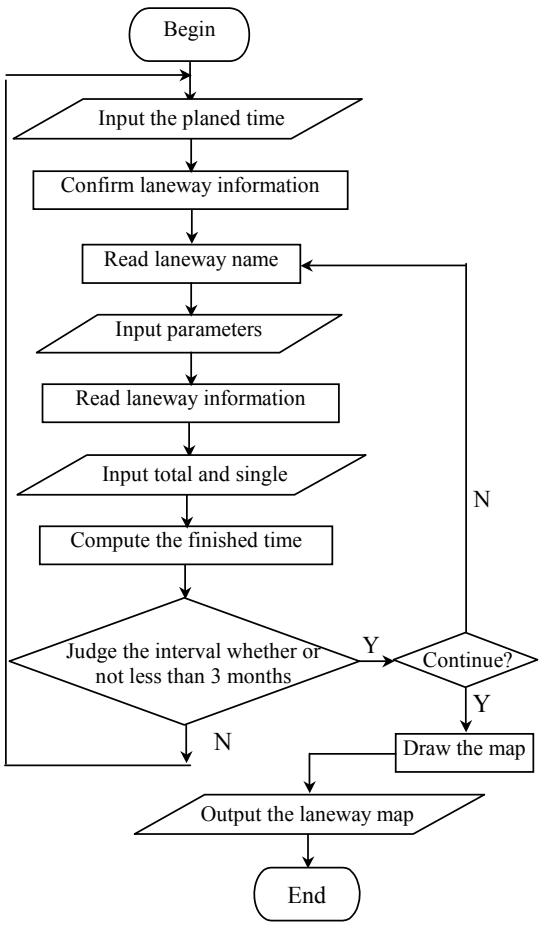


Figure 3 the Flow Chart of the Algorithm Processes

If the intervals between the planned time and the finished time computed by the computer are less than three months, we continue to do the work, otherwise we need return to the beginning. The system will automatic draw the map and output the laneway map after we determine to do the excavating and joining according to the process.

The analyzing of the excavating and joining the laneways in the coal mines can timely and safely analysis the practical situations about the excavating and joining the laneways according to the different excavation area in term of month with the forms of maps and words for the user to provide scientific, comprehensive and timely decisions on support. Specially, it is an useful tool for the decision-makers.

(6) Map and user management

Map management provides users with a map to add, delete, validate and other functions. It is convenient for the users to manage the maps. Editing the maps provides user with a point and a line to precisely draw via Web browser. It is convenient for the users to edit and update the maps. User management functions can be on the use of the system's user management and authorization, therefore the system can be convenient for the administrator to restrict operation to ensure the system secure.

(7) Optimal path enquiry: Electronic maps can display the optimal path with significantly color to link the two places while users choose the search path of the beginning and the end position through clicking the enquiry area.

3.2 Key technologies

Some key technologies such as components[7], SSO, web service and so on are used in the coal mine excavating and joining system.

The components of ASP.NET and SuperMap IS.NET are used to encapsulate each module in the process of implementing the system. Map services and Web services are developed separately, making full use of the characteristics such as rapid development of component, easy to maintain, easy expansion, high cohesion, low coupling and so on in order to achieve the encapsulation of each module and the elimination of the dependence among modules. The model based on components can easily replace and expand a service module to improve the productivity and reduce risks in the development.SSO [8] composed of RBAC (Role Based Access Control) and AD (Active Directory) is

used in the security module of the system as the identity of the users' authentication. The levels structure in business/organization can be simulated by the information tree in AD. All users of AD in all categories are grouped according to the sort of the application system and then grouped secondarily in term of the functions. It can define of reading, writing, implementation and the full control through binding the privileges on the first and the second levels. The users can use the function that they have the right to use once they successfully input the user name and password when they login the system. SSO separates the security module from the system so that it can manage conformably and integrate the existing information system.

3.3 Implementation configurations

The multi-layer distributed coal mine excavating and joining system based on WebGIS and ASP.NET platform is constructed and has been implemented. The main implementation configurations include operating configurations for the servers and the clients in coal mines.

(1) Configurations for the servers

Configurations for the servers are needed to use the slap-up computer servers, with Microsoft Windows 2003 Server operation system, installing the IIS (Internet Information Server), SuperMap IS.NET and the SQL Server, establishing the correct connection with the Internet and configuring TCP/IP.

(2) Configurations for the clients

Configurations for the clients are simple, we only needed to use the common computers, with Microsoft Windows XP or Windows 2000 operation system, establishing the connections with the Internet and configuring the TCP/IP.

4 Conclusion

The three-tiered distributed coal mine excavating and joining system based on WebGIS and ASP.NET platform is constructed in the paper. Combining the component technology of .NET, the modules of the system have some particular characteristics such as rapid development, easy maintain, easy expansion, high cohesion, low coupling and so on. The system has been used in the relevant unit with the good results, guaranteeing the safety, accuracy in the coal mine excavating and joining.

References

[1] Yuan Junru, Huang Weiwei , Liu Donglin, "The Visualization of Statistic Data in the Second Exploitation of WebGIS", Geomatics & Spatial Information Technology, Vol.28, No.2, 2006, pp. 42-43

[2] Zheli Liu, Shuming Wang, Yongjian Yang, "A Distributed Model of WebGIS Based on Java Servlet", Journal of Communication and Computer, 2006, pp.49-51

[3] Xiaoping Rui, Chongjun Yang, Dong Peng etc., "Coal mine WebGIS developing with Java". Geoscience and Remote Sensing Symposium, 2003, pp. 2659-2660

[4] Yingwei Luo, Xiaolin Wang, and Zhuoqun Xu, "Design and Implementation of Map Visualization Objects in Component-based WebGIS".,the Third International Conference on Web Information Systems Engineering, 2002

[5] Gottschalk K, Graham S, "Kreger H. Introduction to web Services architecture". IBM Systems Journal, 2002, 170-177

[6] Ma Rong, Wan Biyu, Hu Qingyu, "A multi-agent system using in spatial information sharing on Web-based GIS", Language Processing and Knowledge Engineering IEEE, 2005

[7] Hong Qi, Su fang Yu, Wenyi Fan, "Development of component geographic information systems applying in forest resources management", Journal of Forestry Research, Vol.16, No.1,2005

[8] Santa Fe, New Mexico, "Single Sign-On in In-VIGO, Role-Based Access via Delegation Mechanisms Using Short-Lived User Identities", l8th International Parallel and Distributed Processing Symposium (IPDPS'04), 2004

A Mobile Multi-Agent and Location-Aware Based Framework for Advertisement

Yang Liu Chunting Yang

School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, China

E-mail: hzliuyang@gmail.com; yangct@zust.edu.cn

Abstract

This paper focuses on providing a method that will make mobile advertising contextually sensitive. We proposed a mobile multi-agent and location-aware based framework for advertisement. This framework includes three modules which are location-aware module, mobile multi-agent module, and information transform module. Through this process, we can send advertisement which is user specific and location sensitive to the customers.

Keywords： mobile multi-agent; location-aware, advertisement

1 Introduction

Mobile advertisement is also a very important class of m-commerce applications [1]. But at the present time, mobile advertisement has some problems:

Low pertinence: Most existing mobile advertisement is push advertisement. This advertisement system provides a uniform message to all users who pass by the store that is broadcasting the advertisement. It provides advertisements that may not be expected by the customer.

Lack of timeliness: According to the advertising theory, the most appropriate time for sending the advertisement to the customer is when the customer is selecting the merchandise. If advertisement sent to a customer is not location-sensitive that will lead time mismatch.

Advertisements sent to a user can also be location-sensitive and meet to user's interest. In this paper, we focus on designing an advertisement sending framework to sending advertisement which is user specific and location sensitive to the customers. Using information on the current location of mobile users and his/her tastes, targeted advertising can be done.

2 Mobile Multi-Agent

Mobile agents move the data processing elements to the location of the data, whose transmissions in the raw which wouldn't incur most of the energy expenditures of the nodes. In addition, mobile agent systems introduce a higher degree of re-tasking flexibility, compared to other approaches, and facilitate collaborative information processing. In article [2], the author proposed seven good reasons for mobile agents:

(1) They reduce the network load.

(2) They overcome network latency

(3) They encapsulate protocols.

(4) They execute asynchronously and autonomously.

(5) They adapt dynamically.

(6) They are naturally heterogeneous.

(7) They are robust and fault-tolerant.

Mobile agents can work either as single processing units or as a distributed collection of components that can cooperate to achieve a given task. Agent cooperation is an important consideration, because this communications mechanism played an important role in reducing energy consumption in the data processing. In essence, information sharing enables agents to learn what other agents have already learned, enabling a faster task completion time, potential bandwidth savings, and energy conservation. Currently, Multi-Agent Systems are viewed as an appropriate technology for the delivery

of services to mobile and wearable devices.

3 Service Based on Location-aware

The main problem with context adaptation is that the context cannot be easily identified or measured. The location of the user is an element of the context that currently can be measured more or less accurately depending on the positioning system in use. Location-aware services are defined as context-aware services that utilize the location of the user to adapt the service accordingly. A location-aware or location-based service is a service which is mostly driven by location information [3]. Location-aware mobile advertising which is based on GPS or Bluetooth positioning and WAP push is a special case of location-based services. A Location-aware system uses location information to provide relevant information or services to the user.

4 Framework of Moblie Advertisement

This paper focuses on providing a method that will make mobile advertising contextually sensitive. First our method aims to deliver relevant advertisements which are location sensitive. Second our method aims to deliver relevant advertisements by considering user's interest.Fig.1 shows the architecture of the mobile advertisement system. Mobile devices, such as PDAs and cellular phones, are used to locate customer.

To make it personalized, the customer character has to be considered and checked before delivering a particular advertisement. Customer information is used to determining the customer character and is stored in the information server.This information include name, sex, age, profession, interests, purchasing habits and etc.It can use user IP address (which is unique) for matching database and then find out the corresponding mobile account.

The mobile agents compute out advertisements which are relating to the user location. Then the mobile agents analyze those advertisements by customer character and search out advertisements which may be interested by the customer. If those advertisements haven't be sent to the user, sent them.

After the customer received such advertisements, they can request for more information regarding a particular advertisement that catches his or her attention.

Through this process, we can send advertisement which is user specific and location sensitive to the customers. This framework includes three modules which are location-aware module, mobile multi-agent module, and information transform module.

4.1 Location-Aware Module

We could locate customer with different positioning systems. If the mobile device includes a GPS (Global Positioning System) module, the customer's location can be defined very accurately within 2–20 meters. The location is calculated in the customer's mobile device and it has to be sent to the service provider in order to get location-aware services. Then customer's location data is transferred to the mobile agent platform through mobile agent API.

GPS cannot be used indoors. If the customer enters into the indoor place, the customer should be identified at a service point, utilizing e.g. WLAN (Wireless Local Area Network), Bluetooth TM or infrared technologies. These kinds of proximity positioning systems require a dense network of access points. The customer's mobile device needs special equipment such as WLAN and Bluetooth.

4.2 Mobile Multi-Agent Module

The mobile multi-agent module is the most important part.Fig.2 shows the architecture of mobile multi-agent. Multi-agent cooperation potentially can enable to perform tasks faster and more efficiently by sharing information between agents.

After locating the customer, mobile agent starts to work. There are five main mobile agents that are resource agent, inquire agent, decision-making agent, cooperative agent and management agent. These mobile agents work cooperatively and give the compute result based on data service.

Resource agent: Resource agent plays an important role to dynamically interface with database resources. The resource agent would take the place of the web sites to bridge the customer to the databases.

Inquire agent: Inquire agent responds the requests of the customer.

Decision-making agent: Decision-making agent can be treated as a data analysis engine. It is capable of analyzing data which collected from resource agent.

Cooperative agent: It provides base function to control inter-agent communication. It comprises cooperative means for the agents, mechanisms to migrate data to other agent, and service security features.

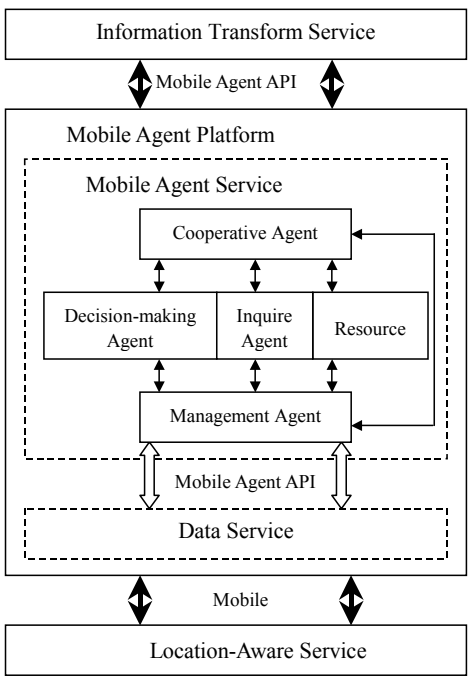


Figure2 The architecture of mobile multi-agent

Management agent: Management agent monitors the conditions of mobile agent compute in real time.

The mobile multi-agent performs task based on data service. A mass of information such as customer information, geography information, expert knowledge and advertisement information have to be stored. Database and knowledge base are two main bases of this framework. This data resource is the target which mobile agents access. Data server must have ability to manage database and knowledge base and provide interactive interface to the manager. Data server also has ability to receive mobile agents and provide operating environment for them.Figure.3 shows the architecture of data service.

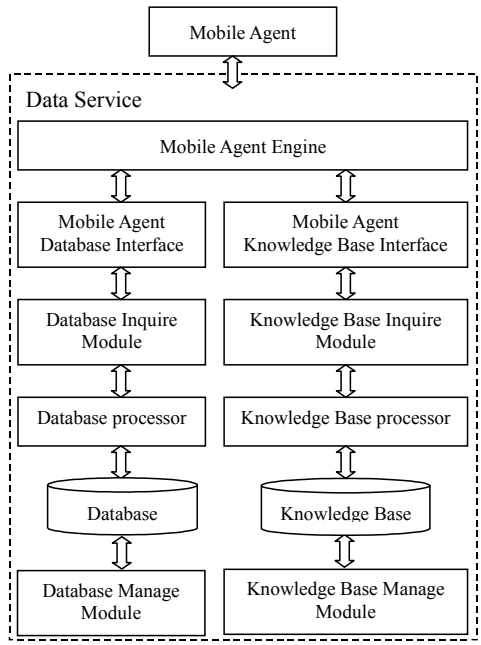


Figure3 The architecture of data service

4.3 Information Transform Module

During the period of advertisement service, there are basically two types of advertising for mobile devices, the push and pull advertising. Initially, the system will send a push advertisement to the user. From the list of advertisements, the user can request/pull for more information regarding a particular advertisement that catches his/her attention.

In our system, a Push Initiator in WAP Push terminal which in the information server transmits push content and delivery instructions to a Push Proxy Gateway using the Push Access Protocol, which uses XML over HTTP. The wireless gateway encodes the pushed message based on different mobile device to a binary over-the-air format and uses a Short Message Service Center (SMSC) with SMS as the bearer to deliver it. The Pull method operates as the same way.Fig.4 shows the architecture of the information transform.

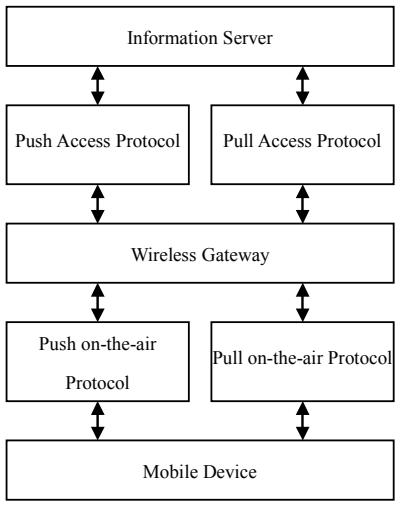


Figure4 The architecture of information transform

5 Conclusions

This paper focuses on providing a unique and personalized advertisement suitable for each customer by considering his/her interest. A mobile multi-agent and location-aware framework for advertisement as we proposed could send the most appropriate advertisement to the customer. This appropriate advertisement means sending appropriate content to the customer in appropriate time .By this method the advertising messages can be personalized based on information provided by consulting the customer's location and his/her purchasing interest.

References

[1] Varshney U, Vetter R.Mobile Commerce: Framework, Applications and Networking Support. Mobile Networks and Applications, Volume 7, Number 3, June 2002, 185-198(14)

[2] D.Lange, O.Mitsuru.Seven good reasons for mobile agents. Communications of the ACM.1999, 42(3):88-89

[3] Lauri Aalto, Nicklas Göthlin, Jani Korhonen, Timo Ojala, Bluetooth and WAP push based location-aware mobile advertising system, International Conference on Mobile Systems, Applications and Services, 2004, Pages: 49 - 58

[4] Chuan Jun Su.Mobile multi-agent based, distributed information platform (MADIP) for wide-area e-health monitoring. Computers in Industry. Volume 59, Issue 1 Pages: 55-68

[5] P. Tarasewich, Mobile Commerce Opportunities and Challenges: Designing Mobile Commerce Applications. Communications of ACM, V. 46, Issue 12. December 2003, 57-60

[6] Weiming Shen; Hamada Ghenniwa; Yinsheng Li; Agent-Based Service-Oriented Computing and Applications; 2006 1st International Symposium on Pervasive Computing and Applications, 3-5 Aug. 2006 Page(s):8 - 9

[7] T. Berners-Lee, J. Hendler, and 0. Lassila. The semantic web. Scientific American, pages 34 - 43, 2001

[8] T. R. Gruber. A translation approach to portable ontology specification. Knowledge Acquisition, (5):199-220, 1993

[9] D Zhang, M Chen, L Zhou ,Dynamic and Personalized Web Services Composition in E-Business, Information Systems Management, 2005

[10] X. Yan. Mobile Data Communications in China. Communications of the ACM, V. 46 Issue 12. December 2003, 81-85

System Design of Lean Supply Chain Based on Green Manufacturing

Yuyan Jiang Jie Li

School of Management Science and Engineering Anhui University of Technology, 243002, China

Email: yuyan_j@yahoo.com.cn

Abstract

Confronted with global resource exhaustion, increasing environmental deterioration and the trend of economic integration, to keep sustainable development, the enterprises have to minimize their environmental contamination and wasting of resources while managing to control the production cost. In this paper, green manufacturing and lean production are simply introduced firstly. And then, how to integrate these two kinds of advanced manufacturing systems into supply chain management is discussed and a structure model of green lean supply chain is illustrated. After that, the operational mechanism of green lean supply chain is discussed as well as its construction principles. Finally, the problems that could occur in implementing green lean supply chain management are analyzed along with their feasible means of settlement.

Keywords： Sustainable development; green manufacturing; lean production; green lean supply chain

1 Introduction

At present, the whole world shortage of resources and deterioration of the ecological environment has been seriously hampering the business and the socio-economic development, and a direct threat to the survival of humanity. From a sustainable development point of view, the 21st century modes of production of enterprises must be from purely profit-oriented and cost-oriented changes to the environment and resources, that is, in the production function of the same products at the same time, as far as possible to reduce the consumption of resources and reduce environmental pollution.

Along with the global industrial structure adjustment and economic integration development, supply chain management (SCM) can effectively control its costs have been increasing attention academia and the business community[1], and to continue to improve and eliminate all the waste the core Lean Production (LP)for the enterprises is a prominent advantage[2]. In recent years, there has been a optimize the use of resources and environmental protection management integration of the new manufacturing paradigm - Green Manufacturing (GM)[3], its purpose is to achieve the ecological and economic harmony, and promote the sustainable development of society. The current view of the above two kinds of advanced manufacturing and supply chain management and the integration of academic research are also being implemented gradually, but it will be integrated as one of the three related research is still rare. Accordingly, based on green paper on the green Lean manufacturing supply chain design, building and operation of the principle issues analyzed and discussed with a view to keep the business sustainable development provides a new way of thinking and reference.

* [The project of the fund] social sciences research project of education department of Anhui Province (Serial number: 2006skj164)

Brief introduction of author: JIANG Yuyan (1966-)，Female, XuanCheng Anhui , associate professor, the main research direction is the management information system , CSCW theory and application

2 Green manufacturing and lean production

2.1 The connotation and ideological essence of Green Manufacturing

The green manufacturing's concept and the connotation are still in the exploration and development phases, therefore does not have a unified standard definition. According to the Society of Manufacturing Engineers 1996 Blue Book on green manufacturing, "Green Manufacturing, also called as clean production, its goal is to make products from the design, production, transportation, and spent the entire process of dealing with the environment to minimize the negative effects" [4] and that its meaning is the product life cycle the entire process will have to be "green", that is compatible with the environment.

The green Manufacturing includes the green design, the green craft plan, the green production (cleaner production), the green packaging, and other basic elements. Its ideological essence manifested mainly in the following three aspects [5]: First, manufacture problems. Green manufacturing requirements in the entire product life cycle consider of the various phases of resources and environmental factors, stressed in product design, manufacturing, packaging, distribution, consumption and scrapped, and other aspects of the green, and its objectives include increasing the conversion of resources efficiency and reduce the pollutants generated by the type and quantity of materials, such as the use of effective recovery. Second is the issue of environmental protection. Green manufacture emphasis manufacturing process the "green", which demands not only the negative impact on the environment minimum, but also to protect the environment. Third, it is optimal resource utilization. Extensive production causes the exhaustion of the resources is the biggest problems which the human sustainable development are facing, green manufacturing will be able to ensure that product features, quality, cost under the premise that the maximum utilization of resources, and the lowest energy consumption.

Clearly, green manufacturing is a comprehensive consideration of environmental impacts and resource efficiency of advanced manufacturing model, in the pursuit of its products from design, manufacture, use to scrap the entire life cycle of pollution minimization, and energy and resource conservation, sustainable development concepts embodied in the manufacturing sector, and thus the furtherdevelopment of the enterprise become an inevitable trend and the only way.

2.2 Lean production and the concept of the basic principle

Lean Production from Japan Toyota Production System (TPS), is a kind of order to minimize production resources and occupied by the enterprise management and reduce operating costs as the main objectives of the mode of production, but it also is a concept, a culture[6]. By American Production and Inventory Control Society (APICS), lean production is made for the activities of the enterprises all the necessary resources to meet the minimum [7].

Lean production of the basic principle is: continuous improvement to eliminate waste, and work together, flexible production. Lean production as a means to streamline the removal of all non-value-added production and supply of materials in the total demand-driven pull-type production, to JIT. Lean Production stressed that the role of, and give full play of human potential, and work tasks and responsibilities transferred to the maximum value-added products directly to the workers, and the mandate from the group of workers collaborate commitment to require workers proficient in a variety of work, and increased workers on the production of autonomy. Lean production manufacturing equipment is not the blind pursuit of a high degree of automation and modernization, but stressed that the transformation of the existing equipment and in accordance with the actual needs of the use of advanced technology and the basis of this principle to increase the efficiency and flexibility

equipment. Lean Production of the "perfect" as the tireless pursuit of the objective of continuously improving production and eliminate waste, reduce inventory, cut costs and diversify product variety, and ultimately realizing maximize customer value.

Lean production system through the structure, organization, operation mode and the market supply and demand, and other aspects of the change in the production system so that users can adapt to the rapidly changing needs, and will enable the production process all useless and superfluous things been streamlined, and eventually achieve market supply and marketing, including the production of all sectors of the best results. Clearly, it is inevitable choice that lean manufacturing enterprises reduce costs, conserve resources, and enhance their competitiveness.

3 Based on green manufacturing lean supply chain design

3.1 Lean supply chain content

Since the 1990s, supply chain management research and practice had become enhance their global competitiveness in an effective way. Supply chain around the core business, through information flow, logistics, and capital flow control, from the procurement of raw materials, intermediate products and final manufactured products, the product by the sales network to the hands of consumers will be suppliers, manufacturers , distributors, retailers and end-users to connect up to an overall structure of a functional chain network model [8]. Supply chain management is on the whole supply chain of the participating organizations and departments of logistics, information flow and capital flow planning, coordination and control, the performance of the enterprises in the strategic and tactical operations of enterprises throughout the optimization process, the purpose of through process optimization to enhance the speed and the associated uncertainties, all related to maximize the net value-added process and improve the efficiency and effectiveness of the Organization.

Lean concepts to be used in the original manufacturing process in a period, really brought a lot to the enterprise income. However, as the increasingly competitive, more and more enterprises found that to truly Lean, at the lowest cost, according to the correct way to put the right product at the right quantity at the right time to correct location, must be the whole supply chain, with other enterprises. This requires making Lean thinking expanded to the whole supply chain to focus on from the purchase of raw materials to manufacturing, to the production of the final product and its delivery to clients throughout the entire process of Lean, Lean Construction of a supply chain and then by raising the whole supply chain to improve the competitiveness of their own competitiveness[9].

Therefore, if the core manufacturers in product design aspect of the joint design tools, internal manufacturing sector using lean production system, its logistics systems (including procurement, distribution system in the delivery of the shipment, etc.) to meet the needs of the lean production system requirements, and to extend this lean thinking to the whole supply chain, and the entire supply chain using lean production system. Lean supply chain can make the least value of the waste stream generated between customers and suppliers, products with the greatest efficiency flows in order to adapt to the new competitive environment of the enterprise market operation of the production of high-quality, highly flexible and low-cost requirements.

Lean supply chain in the development and implementation of, first and foremost, a key task is to coordinate the flow of strategic materials [10]. Can be seen from Figure 1, JIT procurement Lean supply chain management is the key interface, suppliers and downstream manufacturers, distributors and materials must ensure that the "continuous flow." In addition, the lean supply chain should have at least the following conditions and characteristics: simple structure can reduce the uncertainty on the negative impact of the supply chain, production and business processes more transparent; driven by the orders of procurement, supply, sell integrated supply chain model to achieve "a supply

flow, a flow production, a flow distribution"; openness of the enterprise information system between enterprises have better information transparency with the organizational form of dynamic alliance formed to a contract-based, loosely, collaborative nonlinear coupling system, thus realizing the "1 +1> 2" overall effect; production model of intelligent neurons, each business only to focus on their best at work, and can quickly respond to market changes and timely adjust production plans and production technology.

3.2 Lean green structure of the supply chain

Faced on increasingly fierce competition in the market, with enterprises in their efforts to the entire supply chain, lean, can effectively reduce costs and improve their competitiveness. However, as mentioned earlier, the 21st century enterprises to maintain sustainable development, we must strive to reduce the cost of taking into account the resource efficiency and environmental impact, that is, in the product life cycle to follow the entire process of environmental compatibility with the principle of realizing the product for the design, manufacture, packaging, distribution, consumption and recycling, and other aspects of the green. At present, the study and implementation of the lean supply chain management of the supply chain as a whole "green" have not yet given sufficient attention. Although some of the supply chain in the core business has developed and implemented stringent internal standards, but the upstream suppliers and downstream distributors did not comply with the same standards, which undermined the green supply chain design holistic and systemic. From the point of the lean supply chain if not scrapped for good solution to the recovery and re-use, will inevitably bring pollution to the environment and a waste of resources at the same time, it is also contrary to the purpose of Lean.

Lean on green manufacturing supply chain, in addition to the Lean all the characteristics of the supply chain, at least should have the following characteristics:

(1) "green" logistics. Lean green supply chain logistics involves not only the raw materials, intermediate products and final products, but also including product design, manufacturing, distribution to use in every aspect of the waste, flotsam, damage, and pieces of scrap materials; Green Supply Chain Logistics no termination points, such as: we can continue to use the processed products can be re-sold, re-use after the demolition of parts can be returned to the factory to re-melted down the faulty part can be used as the use of raw materials, and so on.

(2) "green" information flow. As same as the general supply chain, supply chain Lean Green has large flows of information sharing, which also includes various enterprise products and green design standards to achieve the state of "green" information. These large flows of "green" information for all enterprises in the green can create integrated upstream and downstream of their own characteristics and analysis, thus achieving better supply chain as a whole of the green.

(3) recycling enterprise integration. The globalization of the economy and abandoned the increasing number of products, allowing for the recycle and reuse of the problem has become increasingly serious. If their recycling enterprises by themselves, it is bound to distract corporation energy, and detriment of the core competitiveness of enterprises improved, if have a dedicated collector’s intervention can better reduce costs and conserve resources and protect the environment.

Lean green supply chain structure of the complexity of the product supply chain process to go through many areas, and its basic processes is that Lean primary supplier in accordance with the principle of providing raw materials for its downstream Green manufacturers; then the downstream manufacturers to follow supply chain members agreed the green design and green manufacturing standards, using lean production methods, its successive processing into a "green" parts, components and finished products; finished products after green packaging and transport, by distributors based on the lean principle to the end-user; users of scrap used by the recycling business products to be timely and properly deal with the separation, thus completing a closed-loop ecological

cycle. Thus, the lean supply chain truly has realized the green, has achieved to the resources full use and to the environment effective protected object.

Lean supply chain in the green, the recycling business can be as a supply chain from the supplier to the final consumer nodes customers, and may become upstream distributors, manufacturers and suppliers of suppliers, logistics has become multi-loop The "green" logistics. In addition, among the members of the supply chain there is a great amount of mutual flow of information, which also includes the flow of information in a large number of "lean" information and "green" information.

3.3 Lean supply chain Green Construction and Operation

Construction of a supply chain is the focus of choice in the supply chain strategic cooperative partnership, based on green manufacturing and lean supply chain in the selection of partners more strict than ordinary. Because these partners are not only information and the sharing of interests and more awareness is a standard synchronous relationship to the agreement.

First, suppliers located in the upper reaches of the entire chain, will transfer its operations to a chain of all nodes. Usually, select of supply chain providers major consider in product quality, price, delivery date, flexible quantities and varieties of diversity. But based on green manufacturing lean in the supply chain, environmental factor is the key factor. This is because the green suppliers to the supply chain is a very significant advantage: Green providers cost savings can be transmitted to the supply chain downstream of the various links, thereby improving the overall efficiency of the supplier's products meet the standards directly affect the green to the downstream enterprise whether a product complies with green standards, thus affecting the entire supply chain of green design. Therefore, the vendor selection process is the importance of environmental management in the selection and breeding enterprises have such a positive environmental management awareness of the enterprise, then its green form strategic partnership.

Secondly, in selected upstream suppliers, also deal with some of the lower reaches of the cooperative enterprise to choose. First of all these enterprises should also be a "green" consciousness, and it is necessary to follow the chain of a series of green standards, particularly in the choice of distributors, can the green marketing should be an important selection criteria.

Finally, it need to all enterprises in the supply chain together to discuss and define the principles of cooperation and follow a set of criteria after selected upstream and downstream partners. For example: the suppliers to provide raw materials to follow what kind of green standards, product recovery after scrapped or returned it to distributors continue to sell parts or demolition as a reproduction, and so on. That need to enterprises in the chain agreed to a uniform standard and consciously abide by. Only in this series of standards, protocols and the related information system to complete after a green Lean supply chain to be truly established.

Lean green supply chain built up, how to conduct effective management to ensure that its "green", "lean" and dynamic operation, it has become the most important issue. The based of green Lean supply chain management operations is: combining existing materials enterprise management systems, information management systems and production systems incorporated into all green conditions Lean production process and be green, and then through the management system, the production process, technical specifications, such as authentication and verification, as well as spare parts and finished product testing and the operation of the supply chain to ensure that all of the production process can be qualified with the user's requirements, and to achieve the most efficient use of resources, waste of resources at least, environmental pollution the smallest management principles and objectives [11].

Green Lean supply chain management to the successful operation, in addition to the aforementioned lean supply chain to meet the conditions of application, it must also accomplish the following:

(1) enterprises by correctly identifying the

environmental impact, this is green supply chain and the operation of the foundation. Business enterprises must have a significant impact on the performance of environmental considerations into green manufacturing certification, the certification system, and external customer requirements and restrictions into the internal norms, and the establishment of enterprises within the operating system, materials management systems, information systems and timely production control system.

(2) members of the supply chain to strengthen the communication and coordination and forming a sense of synchronization, integration of a unified standard mode of operation, this is the success of the green supply chain key. The entire product life cycle of the green operation and supervision of members of the chain needs of participants. In a series of cooperative principles, as well as green standards, enterprises in the chain must be consciously abided by. We should, through the formation of strategic alliances formed a kind of information and sharing of benefits, risks, and the interests of the loss-sharing mechanism, thereby forming a sense of synchronization, integration of a unified standard mode of operation. Specific course of the operation, the core business with suppliers, research and development cooperation, is the fundamental guarantee green product design, speed up the supply chain innovation capacity and speed; core enterprises should take the initiative for those who are small-scale, green poor awareness of the enterprises to provide training and support ; chain enterprises should study together and recovery products to the recovery process, and to strictly follow the reunification of the recovery principles and standards, so as to enhance the entire supply chain competitiveness in the market.

(3) build a support network integration platform, which is green supply chain operation of the fundamental guarantee for success. The platform should have at least the following functions: with a more complete picture of the products in international environmental standards database; to provide the industry with raw materials, the environment of the various parameters of the test data; can provide intermediate products, finished product in the manufacture , use and maintenance of the resources in the process, the energy consumption of the relevant data; can provide different parts of the recovery; technology can provide the impact on the environment and resources, energy efficiency reference data so that enterprises rational use of green manufacturing process; can provide enterprise and industry products overview of the product market information; enterprises to provide support mass customization tools and systems, including business and customer collaborative design system, the standard management system ; built waste products recycling network database, and so on.

(4) Attention and give full play to the role of the government. Green Lean supply chain management will be involved in many aspects of the standard, which requires governmental organizations, formulate uniform. If not the Government's environmental protection policies and regulations binding, enterprises can hardly independently with environmental awareness. In addition, in the course of running the necessary logistics and transport systems and information systems building the government also needs to provide some support and coordination.

4 Concluding Remarks

In the face of the global tension resources, the deteriorating environment and fierce competition in global markets, enterprises in the effective control of production costs ,resources and environmental pollution has become enterprises survive and development’s problems. This paper will try to "green manufacturing" and "lean production" of these two ideas together into the current widely used to effectively control costs in the supply chain management, design a green Lean supply chain structure model and operation mechanism and implementation process, so as to maintain sustainable development for enterprises to provide a new supply chain management model and ideas. Of course, this is only for companies in the implementation of supply chain management on how the integrated use of various

kinds of advanced management methods and carried out by the idea of a framework of the Green Lean supply chain management and implementation of the concrete building, academia and business to be sector further in-depth study.

References

[1] Yangwei, Analysis on Gobalized E-Business and Supply Chain Management[J], Journal of Guangxi University of Finance and Economics. Vol.20 No.6 2007:70-73

[2] Luozhengbi,luojie. Probe into successful application lean production and management to Chinese industry[J], World Manufacturing Engineering & Market, No.6 2007:96-103

[3] Lvshaoyi Liufuyan, Technoeconics & Management Research [J] .NO.5 2001:104-105

[4] MELNGK S A, SMITH R T, Green Manufacturing[R], Dearborn, USA: Society of Manufacturing Engineers, 1996

[5] Lijing, Wangcha, Weidapeng. Promoting lean green manufacturing[J]. Economic Forum,2005,13:120-121,140

[6] Yangbin, A New Interpretation of the Connotation of Toyota's Production Fashion[J], Contemporary Economy of Japan, No.3.2006:43:48

[7] PAULS. With agility and adequate partnership strategies towards effective logistics networks [J]. Computer in Industry, 2000,42:33-42

[8] Mashihu, Linyong, Chenzhixiang. Supply chain Management [M].Beijing: The Publishing company Machinery Industry,2000:41,113-118

[9] MA Tongbing,JIN Yue, WANG Yingchun. Study on Design of Integration Lean Supply Chain Management[J]. Manufacturing Technology & Machine Tool, No.9. 2005:101-104

[10] Jeffrey P Wincel. Lean Supply Chain Management: A Handbook for Strategic Improvement[M]. Productivity Press,2004

[11] AO San-mei. Green manufacturing — the sustainable development model of modern manufacturing industries[j]. Journal of Nanjing University of Technology, VOL.27 No.4. 2005: 106-1

Research on Web Information Extraction System Based-on Multi-Agent Cooperation

Hua Fang[1] Jianliang Wang[2]

1 Department of information Shandong Science Technology Vocational College, Weifang, Shandong, 261053, China

Email：sdfanghua@sina.com

2 Department of management Shandong Jiao tong Vocational College, Weifang, Shandong, 261206, China

Email：sdfywjl@163.com

Abstract

The aim of Web information extraction is to explore usable information, to mine the implied data schema, and to restore the knowledge in the database. The paper firstly analyzes the current situation of the research of information extraction, and then proposes the architecture of information extraction system based on multi-agent cooperation. Moreover, this paper introduces the characteristics of each part in the model and shows the process of information extraction. In this model based on multi-agent cooperation, the complexity between task transitions is reduced effectively.

Keywords：multi-agent; information; extraction; rules generation; database; cooperation; data schema

1 Introduction

Up to the present, the global Web pages have reached the quantity of 11.5 billion or more [1], and they are still increasing exponentially. Facing with such huge information resources, people have an urgent need to develop automated tools to quickly find the information they are interested in so that they can inquire, analyze and report the information. Information extraction precisely begins its development under this situation. Information extraction technologies are extremely useful to extract the specific facts from massive documents. On the Web there is such a document library which provides the rich data pool for users to extract the interested information.

From a general perspective, Web information extraction refers to finding the hidden data schema from a large number of Web documents and then structuring useful data into a table-like form. The input of information extraction systems is the original documents, and the output is some information points containing fixed format. Information points are extracted from the Web documents and then them are integrated in a uniform form. At last, the information points are stored in a database. Information extraction has developed in recent several decades. Two factors are important for the development. One is the geometric increasing of online and offline documents. Another is the concerns of "Message Understanding Conference"(MUC) in this field. After the MUC, the main engine promoting the further development of information extraction is the automation content extraction evaluation conference (ACE) [5] organized by the National Institute of Standards and Technology of America.

At present, people have developed many kinds of information integration systems based on information extraction technologies. According to their principles and methods, these prototype systems can be classified as natural language understanding based information extraction systems, manual construction rules based information extraction systems, machine learning based information extraction systems and visual and interactive information extraction systems. The manual way to construct extraction rules is relatively simple, but

programmers have to develop different wrapper for different sources which leads to the lack of adaptability for the website. If the information resources have structural changes, the wrapper must constantly be updated. The machine learning way adopts automatic extraction method which obtains a highly intellectualized degree. The machine learning way has facilitated user's use in a certain degree, but the extraction rules only have limited expression ability, because before the use of the system we must carry on the massive samples training to it. Visual and interactive information extraction systems define their respective information extraction language. Extraction rules are generated through the visual man-machine interactive interface, but currently it does not have the uniform extraction language standard. Moreover, the ability of these extraction languages to describe semi-structured data was poor.

2 Structure and the model

According to the data characteristics of Web sites, a Web information extraction model based on multi-agent cooperation is presented. In this model, a number of agents carry on effective collaboration to complete the information extraction tasks. This model can extract many data items from Web pages at a time, which can largely improve the information extraction efficiency. Model structure is shown in Fig. 1.

2.1 Documents Pretreatment Agent (DPA)

The information of Web pages often appears in the form of HTML. HTML can flexibly represent all kinds of information in the browser and has brought a big convenience to page design. But it is just the flexibility of HTML that brings along with a number of problems, such as lacking of strict standards and restrictions, lacking of definite structure and schema of the represented data[7,8]. The data in Web pages can be easily understood by users, but it is hard for computers to analyze the semantics. Moreover, markers in the HTML source may not match or there exists empty markers. Therefore, in the process of information extraction, we need to carry on a pretreatment to the sample documents; correct the wrong markers of the documents; remove the empty marker and transforms the HTML documents to the well structured equal documents. Documents parsing agent is just designed to complete the above duty. For instance, we can use page clean tool Tidy [6] to correct the common mistakes in the HTML documents and transform the documents to the well formed equal documents.

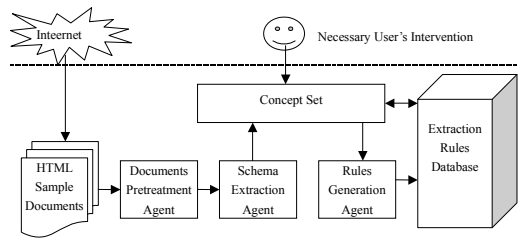


Figure1 Web Information Extraction Model Based on Multi-Agent Cooperation

2.2 Schema Extraction Agent (SEA)

Learning the page data schema from sample documents is an important task for the information extraction systems. SEA is just designed to extract the page data schema. Specifically speaking, firstly SEA will obtain alternative patterns from the page HTML source code. And then SEA will extract the effective data patterns based on some certain filtering rules. At present there are various methods to extract page data schema such as Road-Runner method [2], PAT-Tree method [3], DOM-Tree method [4] and so on. The process of extracting page data schema is shown in Figure2.

The HTML tags and elements such as product information compose the Web documents; HTML source code of each document is regard as the input of schema extraction. In fact, the general Web page structure is quite complex. It is impossible to discover the effective page data schema directly. In the process of schema extraction, we can firstly obtain the alternative data schema set, and then select the effective data schema from the alternative set.

In order to obtain the repeated substrings that are meaningful, in other words, effective data schema, we

firstly present some schema extraction rules. And then, we extract effective schema according to these rules. For instance, the PAT tree method is selecting effective schema according to the regularity and compactness. The regularity focuses on the even distribution while the compactness emphasizes no redundant information existing between schema substrings.

2.3 Concept Set

SEA can finally obtain page data schema according to sample documents [9]. The data schema may correspond to specific instances of data which is generally represented by a two-tuple :< name, property>.

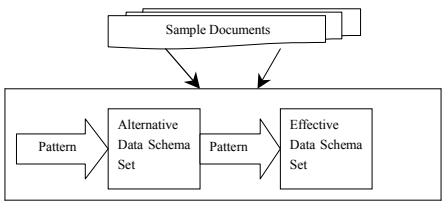


Figure2 The process of extracting page data schema

For example <name1, property>, <name2, property>,… <nameN, property >. The property is used to describe the data type, the position that data schema appears in the document and so on.

Furthermore, we can induce the concept set from the data instance corresponding to the page data schema. The concept is used to represent one class of data instances. It is the highly abstract and generation of data instances. Each concept is described by one group of concept properties. The property is used to describe some characteristic of concepts such as concept type, context information and so on. A concept is described by the following form:

Concept Concept_name
Type: Concept's type
Value: Concept's value
Prior Tag: Starting position tags in HTML code
Succeed Tag: Ending position tags in HTML code
Path: The beginning path in HTML code
Cardinality: Restraint conditions
... ...
End Concept_name

2.4 Rules Generation Agent(RGA)

RGA is designed to generate extraction rules from the concept set [10]. Extraction rules are the important guide to extract information from Web pages. In the future action, the system will take the extraction rules as a guide to extract information and knowledge from Web pages. The extraction rules are generally described by a five-tuple as following shows:

$$rule =< name, type, path, succeed\_tag, others >$$

The meaning of each item is:

①name: Rule name

②type: Type of content extracted by the rule, for example, String, Numeral, Date and so on.

③path: Beginning path of the information extracted in Web pages

④succeed_tag: End tags of the information extracted in Web Pages

⑤others: Some explanations

An information extraction rule corresponds to an information point. Extraction rules point out the positions where the information points appear in Web pages in detail. The path of a rule indicates the beginning of an information point. Once we find the beginning path, it means that we have found the beginning position of an information point in Web pages. Along the path until the succeed tag of a rule, we will find the information we need. The information type is marked by rule's type.

2.5 Rules Database

The key component of an information extraction system is a series of rules or schema. Its function is to determine which information will be extracted. Rules database is used to store the extraction rules generated by RGA. The rules can conduct the specific action of information extraction.

According to the information point, we remove the invalid rules, save the effective rules together and then get a rules database. Extraction rules are generated from the concept set; meanwhile, concept set is used as a guide to standardize extraction rules and exclude invalid

rules. Extraction rules and the concept set learn from each other and influence each other.

Therefore, in this model, each agent mutually cooperates to complete all the tasks in the process of information extraction together. Firstly DPA corrects the mistakes in the primitive Web sample documents and generate the well structured Web documents which will be passed to the SEA. And then SEA extract effective data schema from the sample documents. At the same time, concept set is produced by the data schema. Finally, RGA induces and learns extraction rules according to the concept set. Extraction rules form a rules database which is the basis to extract information. In the process of extraction rules' generation, concept set and extraction rules carry out effective interaction which further improves the generation of extraction and enhances the accuracy of information extraction.

3 Conclusion

Information extraction is a young research area. After several decades of its development, many successful information extraction systems have emerged, but there are still some problems. For example, the accuracy and robustness of information extraction system need to be raised. Secondly, the portability of them is poor because current information extraction systems only target at a specific area. If they are applied to other areas, we have to redesign the extraction patterns. Finally, the degree of automation to build an information extraction system in a new area is low, because in this situation we need the joint efforts of many domain experts and computer linguists in the field, which is time-consuming and laborious.

This paper proposes a Web information extraction model based on multi-agent cooperation. The agents in this model can cooperate to finish all the tasks. The way by cooperation between agents enhances efficiency and precision of information extraction. How to make this model extensible flexibly to adapt the requirement of dynamic network is our future work.

References

[1] A Gulli,A Signorini. The Index able Web is More than 11.5 billion pages[A],International World Wide Web Conference, 2005: 902-903

[2] Valter Crescenzi,Giansalvatore. RoadRunner: towards automatic data extraction from large Web sites[A], Proceedings of 27th International Conference on Very Large Database[C],2001:109-118

[3] Chia-Hui Chang,Shao-Chen Lui,and Yen-Chin Wu: Applying Pattern Mining to Web Information Extraction[J]. Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining,Pages: 4 – 16,2001

[4] Cui Jixin,Zhang Peng. DOM based Web information extraction [J],Journal of Agricultural University of Hebei,2005,28(3):90-93

[5] Zhang Zhixiong. Information Extraction and Its Functions in the Digital Library [J],New Technology of Library and Information Service,2004,No.6:1-5,23

[6] http://www.w3.org/People/Raggett/tidy/,W3C,2003

[7] N R Jennings. On agent-based software engineering. Artificial Intelligence, 2000, 117(2): 277～296

[8] F Zambonelli, N R Jennings, M Wooldridge. Organizational abstractions for the analysis and design of multi-agent systems. In: Proc of 1st Int'l Workshop on Agent-Oriented Software Engineering, Lecture Notes in AI, 1957. Limerick, Ireland: Springer-Verlag, 2000. 127～141

[9] J Ferber, O Gutknecht. A mata-model for the analysis and design of organization in multi-agent systems. In: Proc of ICMAS-98. Paris, France: IEEE CS Press, 1998. 128～135

[10] Xu Jinhui, Zhang Wei, Shi Chunyi. A structure-oriented method for agent organization formation. In: Zhongzhi Shi, B Faltings, M Musen eds. Proc of Conf on Intelligence Information Processing, Beijing: Publishing House of Electronics Industry, 2000. 251～258

Research on E-mail worms' propagation and control

Aiping Wang Shengwen Zhang Jian Song

Institute of Economics and Management, Dalian maritime university (DLMU), Dalian, Liaoning, China

Email: wap1130@163.com

Abstract

With the further application of internet, the threat of internet worm on the security of computer system is increasing day by day. Thus it obtained more attention of many people. Aiming at working principle and security issues of E-mail and transmission mode of network worms, a worm control system of multi-level firewall network which is composed of firewall system, distributed worm detection system, DNS worm control system, and network antivirus system is proposed.

Keywords：E-mail worm; multi-level firewall system; distributed worm detection system; DNS.

1 Introduction

With the Rapid development of Internet, E–mail has been an indispensable part of in the day-to-day work life, needs of using e-mail communication is increasing at an alarming rate. The issue of network security has become more serious, network intrusion and security incidents occur frequently.

1.1 Working principle of E-mail

Transmission of a mail system contains three parts of user agent, transfer agent and receive agent. User agent is a procedure of a client sending and receiving letter, in charge of packing according to a certain standard to mail server which sends or takes back the letter.

Transfer agent takes charge in letters' exchanging and transmitting, sends the letters to a proper mail host computer, then the letters to different mail box by receiver agent. Transfer agent must be able to receive the letter by user mail procedures, arrives destination accurately according to SMTP. Now generic transfer agent has adopted the procedure of "send mail" to fulfill task, then gets to mail host computer by POP for clients to read their own host computer.

That is to say, E-mail's transmission between client PC and ISP internet is done by POP3, however, its transmission on internet if achieved by SMTP. Following explain E-mail's transmitting process and work principle by the example of E-mail post office of sina and 163.(Figure1.1)

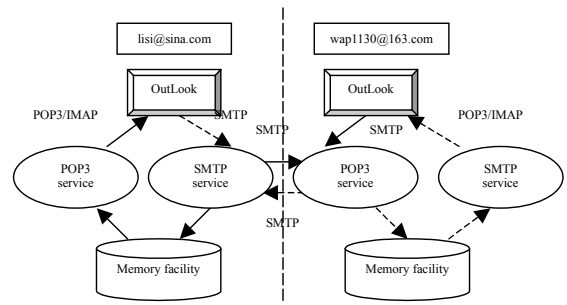


Figure1 e-mail transceivers process between sina and 163 post offices

1) Establish network connection between lisi@sina.com's mail client programs, and sina's SMTP server, and login, the send mail to SMTP server of 163 by SMTP.

2) After receive the mail which lisi@sina.com submits SMTP server of sina Firstly, determine whether the recipient's e-mail address belonging to the SMTP server's jurisdiction. If that is the case, storages E-mail to the recipient's mail box directly, otherwise, sina SMTP server enquiries MX record of domain name expressed by recipient's mail address suffix(163.com) from DNS server, and thus gains 163 SMTP server information, and then connect with 163 SMTP server,

sends mail to 163 e-mail using SMTP protocol.

3) After received the e-mail from SMTP Server, the SMTP Server of 163 judges the range of SMTP Server according to the recipient's address. If that is the case, storages E-mail to the recipient's mail box directly, otherwise (generally not be in such a situation), this e-mail may continue to be forwarded or discarded.

4) Client who has account of wap1130@163.com establishes network link with 163 POP3/IMAP server through mail client programs (assumed as outlook express), and uses wap1130 user name and password to log on, examines whether there are new messages by POP3 or IMAP, and if so, reads mail by POP3 or IMAP.

1.2 Security problems of E-mail

However, working mechanisms of E-mail show they have some loopholes exploited by malicious users easily. When e-mail messages related to commercial secrets, personal privacy, and other content, such information once are intercepted by malicious attacker, it will bring irrecoverable losses. On the existence of E-mail address feature, introduce E-mail's Security issues[3]:

1) Users set password which is too simple or has very obvious feature.

2) E-mail viruses. E-mail is the main way to transmit viruses and Trojans. As the more popular Panda burning incense virus: transmits mainly by infectious page document and e-mail, the virus code numerously hiding in the code page document, the majority of such viruses page documents are most types of advertising, cause uncontrolled proliferation in the Internet, resulting in the paralysis of the computer system, seriously hamper people's normal work and study.

3) Junk mail and e-mail bombs. Junk mail refers to little popular and avoidless e-mail or list which be sent to news groups or others without the users' permits. The complexity coming from such junk mail makes intelligent heuristic scanning engine and other anti-spam technologies difficult to detect the spam. Deceiving money, transmitting pornography and releasing reactive remarks in junk mail have brought great harm to the society. Mail bombs specifically refer E-mail sent to the same receiver continuously in a very short time through special E-mail software, Inbox certainly cannot afford overburden facing these tens of millions of large-capacity, and ultimately "explosion killed" occurred.

4) E-mail easily intercepted. As a network application service, E-mail adopts SMTP (Simple Mail Transport Protocol). SMTP has been a mail transferring standard on the Internet , seeing to E-mail transferring on the network, providing how E-mail to transfer in mail server. But from the perspective of safety, SMTP is almost undefended protocol, SMTP information transmission adopted express form and be fixed at 25-port, therefore, easily to be monitored and attacked. So that transmitted data has any encryption. Since e-mail sending is forwarding through different routers, arrives at final reception host all long, attackers can head E-mail data packets off which we cannot find.

5) Security loopholes of E-mail receiving client software

Now many people send and receive E-mail by E-mail receiving client software which is very convenient, but its design flaws could cause E-mail loopholes. Such as Outlook Express is powerful and can be integrated with the operating system, having considerable users, but it has a loophole of address book, some people can make u send a letter which would send to your friend send his Inbox, having possibility of being cheated.

6) Mail server loopholes. The most common e-mail server program Sendmail, Qmai, have security flaws in varying degrees. Take Sendmail for example, in the old version, telnet to 25-port, input "wiz", then "shell", "rootshell" can be get, and debug order can also obtain root privileges.

2 Worm propagation model

Transmission channels of Network worms are as follows:

2.1 transmit using system loopholes[2][4]

Such worms transmit mostly through some loophole of Microsoft Windows OS. Typical example is the "shock wave" worm (W32.Blaster.Worm) which has caused tremendous damage.

2.2 transmit using application procedures loopholes

Such worms transmit through network application procedures loopholes. Some transmit using a certain FTP service procedures' loopholes, and this kind of worm is relatively rare and the harm is not very great, Typical example is "Linus.Ramen.Worm".

2.3 Transmit using browser

It can add a small JavaScript code to HTML or ASP document through changing the content of web server, some version of IE can automatically carry out this code, thus local host infected. Nimd worm is the first virus using this mechanism which makes worm code penetrates through the firewall.

2.4 transmit using E-mail

It can get mail address list in infectious computer's address list through MAPI, and then send worm code as mail annex to other host by Windows mail client. Not patched IE will automatically carry out e-mail annex and activate worm, even playing the patch, as long as users open an attachment, the worm code will carry out. Accordingly more hosts could be infected.

2.5 Depending on network sharing

Depending on network sharing is one of important ways of worm transmission, and network worms transmit using network resources sharing, as well as Nidam worm does.

General transmission process of Internet worm for scan, attack, the scene disposal, copy, as shown in Figure 2.1.

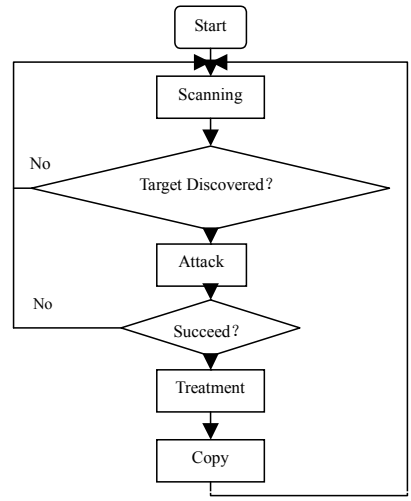


Figure2.1 General dissemination process of worm Procedures

Worm Procedures

1. Scan: Scanning function module take charge of detecting the mainframe with loopholes. After the procedure sends information of detecting loopholes to some mainframe and receives successful feedback, a transmissible target will be obtained.

2. Attack: attack module automatically steps to attack the target in step1 and gets this mainframe's jurisdiction (normally administrator privileges), obtains a shell.

3. Scene disposal: make the computer hold a backdoor after infected to start a distributed denial of service attack.

4. Copy: Copy module will copy worm programs to the new mainframe through interactive program between source mainframe and new mainframe and activate it.

Network worms' workflow process by e-mail transmission is different with the above, and the workflow described as: obtain e-mail address from e-mail address book--send e-mail with worm procedures in group--e-mail is activated, worm procedures start, as shown in Figure 2.2.

3 Worm control

As the main network security technology, firewall technology is widely used in network worm control, but

it is difficult for single firewall to control rapid and large areas transmission of network worms. Therefore, we have proposed an E-mail worm integrated control system based on a multi-level firewall, and this system is composed of multi-level firewall system, network worm detection, control system and anti-virus system. As shown in figure.3.1.

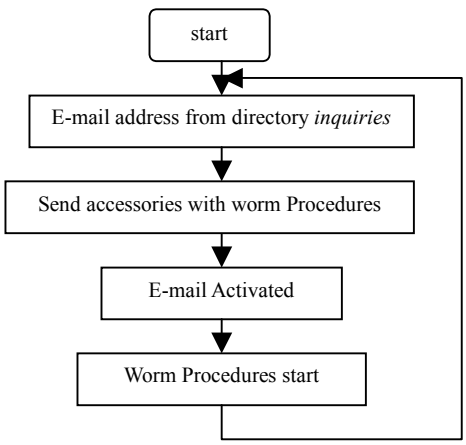


Figure2.2 Workflow of E-mail network worm

3.1 Multi-level firewall system

1) Edge firewall

Worms begun to spread mainly from the extranet, at network exports it is important to set up defense which is put at Network exit connecting with Internet. Logically, firewall is a segregator, limiter and also an analyzer, monitor any activities between internal network and Internet to ensure security of internal network.

2) Internal firewall

In order to protect information among the internal of subnets, users need to isolate flow of intranet segments through setting up two firewalls and make reasonable control.

3) Personal Firewall

Personal firewall software installs on the user workstations, be able to make respond and record of the various scanning and malicious attacking, in a certain extent and reduce the possibility of a virus attack in a certain extent. E-mail telecommunicates through specific ports, and firewall software will be closed most of the commonly unused ports, therefore, you must set firewall security rules before using the firewall.

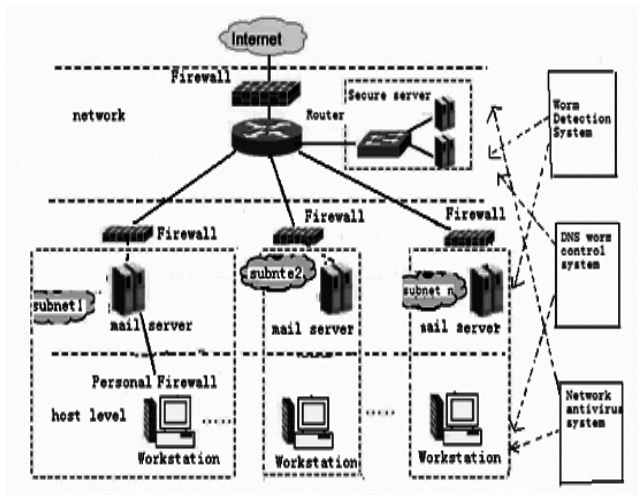


Figure3.1 Worm integrated control system based on a multi-level firewall

3.2 Distributed worm detection system

The development of high-speed Internet and enlargement of the scale of the network make data flow larger and larger, signal network worm detection is difficult to achieve the large flow monitoring. We designed a distributed network worm detection system which consists of many intrusion detection systems which run Snort, and adopted three-tier architecture.

The first layer is a sensor layer which can monitor the flow, take charge of "rip data packets", and transfer data packets to the second layer. The second layer is a server layer which collects alarm data from the first layer and translates into a readable form. In the second layer, you can match the collected data using Snort rule, and put Alarm data into the database. The third layer is console of analyst, and this layer is a partial system for security administrators to analyze and show data. System Components shown in Figure3.2

3.3 Network antivirus system

At present, computer network has become the largest source of the transmission of the virus, E-mail and network information transmission open the high-speed channel for transmission of the virus. The

high efficient of Network virus contagion highlighted new requirements, network anti-virus system should be divided into a three-tier structure:[6]

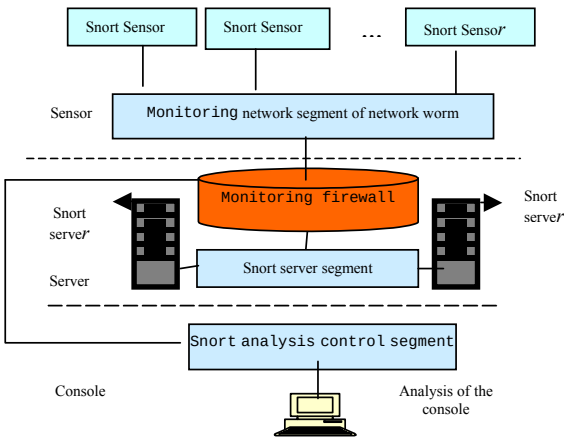


Figure3.2 Distributed worm detection system

1) Prevent virus at network entrance

Be able to head off worms at the network entrance, detect the new virus hidden in the annex of e-mail at any time.

2) Network anti-virus

Be able to monitor virus invasion of each node, to protect the integrity and accuracy of network operating system from the damage from virus damage; Do dynamic alarming, killing to virus from the node invasion, making defensive ability of virus of network system maintain on the same level; Configure an overall system's timetable used for virus detection in addition to inspection, for the realization of the virus detection and periodicity and plan of inspection; Do security audit to virus events, provide evidence to system administrator for tracking and tracing various possible virus incidents.

3) Single anti-virus

Single anti-virus mainly used for working with network anti-virus, as a supplement of network anti-virus to achieve dynamic defense and static antivirus combined. Network anti-virus system layouts include e-mail anti-virus gateway, Internet version of anti-virus software server and client.

3.4 DNS worm control system

DNS is one of the most commonly used services in the Internet. Generally, before users use other Internet services, mostly use the DNS which is the most frequent users of network services. In the enterprise internal networks generally set up a DNS server, users generally use the Enterprise Network internal DNS server. Therefore, DNS can be used to control the spread of network worms[5].

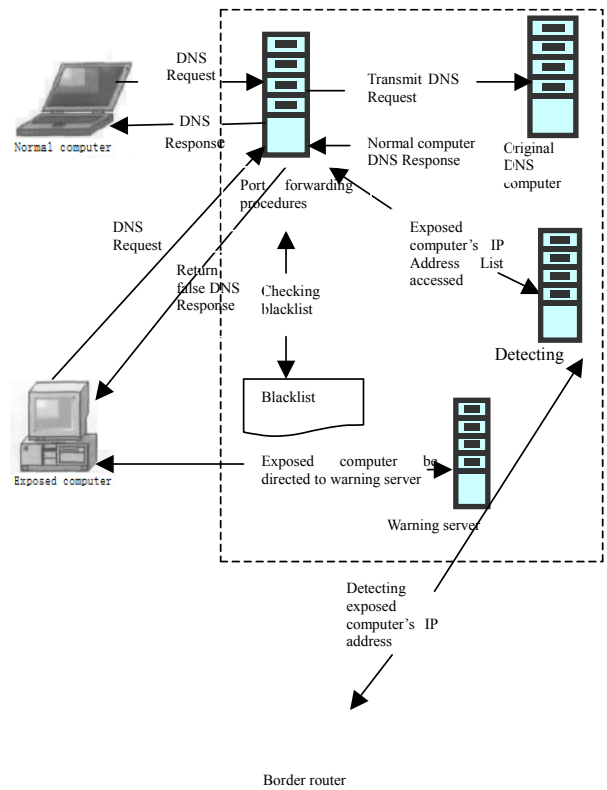


Figure3.3 Overall system architecture based on port ending

Main principle: In a pre-determined circumstance, DNS services in the specific procedures of the 53 ports to monitor DNS request message from users, if we replace the original DNS service procedures, the client procedures is not aware of this. In addition, the DNS requesting and responding message transmit adopted UDP protocol, is not sensitive to delay, so even if alternative procedures can adopt a more complex handling mechanism, it will not cause great impact to users communications. When the alternative procedures receive DNS request message, firstly judge whether the petitioner in the list according to the identified host list infected by worms, if that is the case, return false DNS response message, guide users to visit WIS; otherwise, the alternative procedures transmit request message to

normal DNS service program running in other ports (or host), then forward the responding messages returned by DNS service procedures to DNS requester.

4 Conclusion

On the basis of in-depth study of the current firewall technology, in order to effectively control network worm rapidly, large areas of transmission, proposes a multi-level firewall based on network worm integrated control system. The system in the network edge, the various sub-networks and on the mainframe users disposes multi-level firewall system, with network worm detection and control systems, anti-virus systems come into a comprehensive network orm ontrol system.

References

[1] Zhao Dongmei. Comprehensive risk assessment of the information system is based on entropy theory. Progress in Intelligence Computation and Applications. Changsha, China,2005

[2] Zhang Yunkai. Worm Propagation Modeling and Analysis Based on Quarantine. Proceedings of the THIRD International Conference on Information Security (Infoseeu'04). Shanghai, China, ACM press. 2004:69-75

[3] Qing Sihan. Cryptography and computer network security. Beijing: Tsinghua University Press,2001

[4] Zhang Yunkai. A Worm Transmission and Control Model based on Firewall. Xi'an University of Electronic Science and Technology Journal. 2006-1

[5] http://netsafety.163.net/

[6] Shigang Chen, Yong Tang. Slowing Down Internet Worms. Proceedings of the 24`" International Conference on Distributed Computing Systems, 2004

[7] Laurent Oudot. Fighting worms with honeypots: honeyd vs msblast.exe. [R] http://lists.insecure. org/lists/honeypots/2003/Jul-Sep/0071.htm1,August 2003

[8] Eugene H. Spafford, "The Internet worm program: an analysis", ACM Computer Communication Review, 1989, 19 (1):17~57

[9] David Moore, Colleen Shannon, Geoffrey Voelker, and Stefan Savage. Internet Quarantine: Requirements for Containing Self-Propagating Code. [R]. In proceedings of the 2003 IEEE Infocom Conference, April 2003

An Ontology System and Semantic Integration Architecture for Intelligent Transportation System of China*

Jun Zhai Miao Lv Yiduo Liang Jiatao Jiang Qinglian Wang

School of Economics and Management, Dalian Maritime University Dalian, Liaoning 116026, China
Email: zhaijun_dlmu@yahoo.com.cn

Abstract

Information is one of the key components in Intelligent Transportation Systems (ITS). Heterogeneity is inevitable because the concerned systems are often developed by autonomous participants. Ontology is a new method describing conception-level architecture and semantic model. In this paper, we study solving the semantic integration of ITS by ontology. Firstly, we bring forward an ontology system for ITS of China, which includes basic ontology, domain ontology and applied ontology. Then we propose the architecture for semantic integration of ITS, which provides with transparent service based on semantic for consumers so that it forms a virtual homogeneous environment for the applications. The research shows that ontology is a good tool to integrate traffic information at semantic level.

Keywords: Ontology; Semantic Integration; Intelligent Transportation Systems (ITS); Traffic Information

1 Introduction

Intelligent Transportation Systems (ITS) is composed by a series of subsystems, which are independent relatively and collaborative entirely [1]. Exertion of whole benefit of ITS depends mostly on the coordination and integration of each subsystem. The research result indicates there is great potential for integration to improve systematic performance of ITS. At present the research concentrates on the layer of architecture and data integration. For example, Smith and William have brought forward an integrated ITS architecture called IITS[2]. They discuss the design idea of combining concentrative database with data warehouse and data mining. Through establishing the common data model, René et al. proposed a framework for integrating existing and novel intelligent transportation systems [3]. Li et al. brought forward the platform of integrated transportation information, the key of which laid on integrating the distributed heterogeneous data source based on XML [4].

Integration mainly solves the heterogeneity of information system. The integration of information faces four-level heterogeneity[5]. Firstly system heterogeneity is among operation systems and hardware. Secondly, syntax means different languages and data manifest. Thirdly, structure heterogeneity includes different data models. Fourthly semantic heterogeneity means words and concept have different meanings in the different context and that is concerned with natural language, which is not formal.

The knowledge sharing and integration among systems should be built on understanding knowledge commonly. The first problem solved is semantic mismatch among systems, i.e. the integration of ITS in semantic level.

Ontology is a new method describing conception-level architecture and semantic model [6]. It is an efficient method to solve knowledge alternation, sharing and reusability and utilized in many domains such as knowledge management, intelligent information searches and so on. Ontology is also utilized widely in information system integration and interoperation based on semantic for solving information sharing of

* This work is supported by the Project of the Educational Department of Liaoning Province (Leading Laboratory Project) under Grant NO.20060083.

distributed heterogeneous system [7]. Ontology also is applied in the GIS and urban management, which domains are connected closely with intelligent transportation.

This paper achieves semantic integration of ITS by ontology. To fulfill that objective the paper is organized as follows. In section 2 briefly introduce the concept of ontology and in section 3 propose the three-level ontology system for ITS of China. In sections 4 we describe the semantic integration architecture. Finally, in section 5 we present some conclusions and indicate directions for future work.

2 Basic Definition of Ontology

Ontology is a philosophic concept originally which is used to describe the abstract essence of objective praxis. In 1991, Neches et al. introduced the ontology to artificial intelligence [8]. They thought ontology was the basic term and relation according to the glossaries of relevant domain and the rule to regulate the extension of these glossaries. In 1993, Gruber gave the most popular definition of ontology, which was explicit and normative illumination of conceptualization [9]. Now the most exact definition is given by Studer etc [10]. They think ontology is the explicit and formal illumination of sharing conceptualization, which including 4-level meanings, conceptualization, explicit, formal and share.

An ontology organizes domain knowledge in terms of concepts, properties and relations and can be formally defined as follows:

Definition (Ontology) - An ontology O is a triplet of the form O=(C, P, R), where:

(1) C is a set of concepts defined for the domain. A concept is often defined as a class in an ontology.

(2) P is a set of concept properties. A property p is defined as an instance of a ternary relation of the form p(c,v,f), where c is an ontology concept, v is a property value associated with c and f defines restriction facets on v.

(3) R is a set of binary semantic relations defined between concepts in O.

A set of basic relations is defined as $R_b = \{\approx, \uparrow, \nabla\}$ which have the following interpretations:

(1) For any two ontological concepts $c_i, c_j \in C$, $\approx$ denotes the equivalence relation. $c_i \approx c_j \Rightarrow c_i$ is equivalent to c_j. The synonym relation of natural language is modeled in an ontology using the equivalence relation.

(2) $\uparrow$ denotes the generalization relation. $c_i \uparrow c_j \Rightarrow c_i$ is a generalization of c_j. When an ontology specifies that c_i is a generalization of c_j, then c_j inherits all property descriptors associated with c_i, and these need not be repeated for c_j while specifying the ontology.

(3) $c_i \nabla c_j \Rightarrow c_i$ has part c_j. In an ontology, a concept which is defined as aggregation of other concepts is expressed using the relation ∇.

3 Ontology System for Its of China

This paper gives the definition of ontology system of intelligent transportation, which is sharing and conceptual form of knowledge system and illumination criterion showed in domain of applied transportation. The concept of intelligent transportation ontology forms hierarchical structure, which can be originated from conceptual sort system existing in transportation domain. The concept of ontology can be described by attribute set, which forms multidimensional feature vector space.

We divide the intelligent transportation ontology system into three-level architecture shown in Figure 1, basic ontology system, domain ontology system and applied ontology system. Basic ontology is composed by a group of perfect and inseparable basic concept that is atom ontology. The domain ontology system is a conceptual system in the level architecture. Every concept can be described by a group of attribute set. The attribute set of concept is the subset of basic ontology system. So the concept in the different ontology system can be described by basic ontology and that lays a steady foundation for integration and interoperation of ontology in the different domain. The applied ontology system is the projection in the application of domain ontology combining the need of consumer.

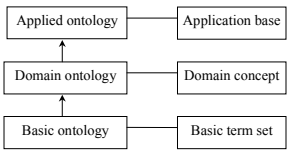


Figure 1 The structure of three-level ontology system

The basic ontology is a group of terms, irrelevant with given task and provides base for integration of different domains. The introduction of basic ontology provides the same standard for conceptual semantic compare of different system. For example, if two concept names are different and the attribute set and the value set of every attribute are the same, the semantic of two concepts are the same.

The domain ontology system can be established by ontology as follows according to National Intelligent Transport System Architecture of China.

(1) Service domain ontology

It includes 8 sub-domains. They are traffic management and planning, electronic payment service, traveler information system, vehicle safety and driving assistance, emergency and security, transportation operation management, intermodel transportation and automated highway system. Figure 2 shows the structure of service domain ontology.

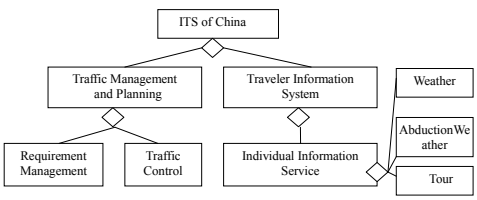


Figure 2 Structure of service domain ontology system (portion)

(2) Consumer ontology

It includes 6 sub-domains. They are road users, road constructors, traffic management departments, operation managers, public security and safety departments and related organizations.

(3) Service ontology

It includes 9 sub-domains. They are traffic management centers, passenger transportation, traffic information service provider, emergency management, infrastructure management departments, freight transport service providers, products/equipment manufacturers, products/services providers and law enforcement management.

(4) Terminal ontology

It includes road users, roadway, vehicle, freight, consignor, banks and system operators.

Figure 3 shows the partial applied ontology for transport information service. The main concepts include roadway, vehicle, driver, hotel etc, and the main relations include generalization, equivalence etc, e.g. the "organization" is generalization of "hotel" and the "road" is equivalent to "roadway".

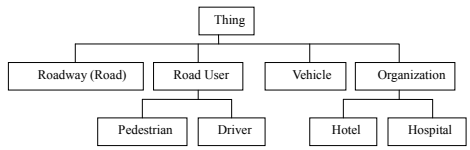


Figure 3 Applied ontology for transport information service (portion)

It should be point out that ontology is not confirmed uniquely and ontology cannot be established once. The ontology in all of the domains should be designed to be extensive to suit the concept of domain, which is developing unceasingly.

Furthermore, we use Web Ontology Language (OWL) to represent formally the ontology as following:

```
<owl: Class  rdf: ID= "Road User"/>
<owl: Class  rdf: ID= "Driver"/>
<owl: Class  rdf: ID= "Pedestrian"/>
<owl: Class  rdf: ID= "Roadway"/>
<owl: Class  rdf: ID= "Road"/>
<owl: Class  rdf: ID= "Driver">
    <rdfs: subClassOf  rdf: resource= "#Road User"/>
</owl: Class>
<owl: Class  rdf: ID= "Pedestrian">
    <rdfs: subClassOf rdf: resource= "#Road User"/>
</owl: Class>
<owl: Class  rdf: ID= "Roadway">
<owl: equivalentClass rdf: resource= "#Road">
</owl: Class>
```

4 Semantic Integration Architecture

We bring forward the integration architecture showed in Figure 4, so that ITS can achieve semantic integration in the different subsystem.

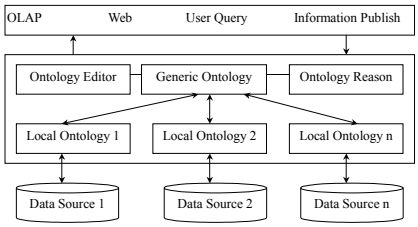


Figure 4 Semantic integration architecturebased on ontology for ITS

The core of the integration framework is generic ontology management system. It supervises the distributed heterogeneous subsystem and communication with consumers, provide with transparent service based on semantic for consumers so that it forms an abstract homogeneous environment for the application of consumer.

Through extracting the inner knowledge and pattern of each data sources, the independent local domain ontology is established. Further we can get the generic domain ontology by merging the local ontology, which provides the user access to the data with a uniform interface of semantic level.

In our simulative environment, the urban road can be divided into expressway, main road, submain road and lateral road. We denote ontology1= {expressway, main road, submain road, lateral road}. A remaining system uses old sort standard. It denotes ontology2= {the first-degree road, the second-degree road, the third-degree road, the fourth-degree road}. To achieve semantic integration of two systems in the integration framework, we establish a generic ontology for urban road, in which the following relations exist: expressway ≈ first-degree road, main road ≈ second-degree road, submain road ≈ third-degree road, lateral road ≈ fourth-degree road.

5 Conclusions

The ontology is introduced to integration of ITS in order to help to solve these problems:

(1) To supply universal identification term

Basic ontology supplies criterion essential and terms to describe the world. These terms are defined strictly and get recognized together.

(2) To show and describe recessive knowledge

People in the colony living environment use recessive knowledge existing in the conception system and expert knowledge unconsciously t and daily. Ontology plays a key role in the show and formal description of the knowledge.

Our further researches lay on the automatic integration among different domain ontology, especially among the transportation ontology and other domain ontology such as the GIS ontology and urban ontology.

References

[1] Lino Figueiredo, Isabel Jesus, J. A. Tenreiro Machado, Jose Rui Ferreira, J. L. Martins de Carvalho, “Towards the Development of Intelligent Transportation Systems”, In: Proceedings of the 2001 IEEE Intelligent Transportation Systems Conference, USA, pp.1206-1211, 2001

[2] B.L. Smith, T.S. William, “Development of integrated intelligent transportation system”, Transportation Research Record 1675, Washington D C: TRB, pp.85-90, 1999

[3] René Meier, Anthony Harrington, and Vinny Cahill, “A Framework for Integrating Existing and Novel Intelligent Transportation Systems”, In: Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria, September 13-16, pp.650-655, 2005

[4] Li Ruimin, Lu Huapu, Qian Zhen, Shi Qixin, “Research of in the Integrated Transportation Information Platform Based on XML”, In: Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems, Vienna, Austria, September 13-16, pp.214-219, 2005

[5] A.Sheth, “Changing focus on interoperability in information system: from system, syntax, structure to semantics”, Interoperation Geography Information Systems, kluwer, Academic Publishers, pp.5-30, 1998

[6] Stab S., Stude R, “Knowledge processes and ontologies”, IEEE Intelligent Systems, Jan/Feb, pp.26-34, 2001

[7] F.T. Fonseca, M.J. Egenhofer, “Ontologies and knowledge sharing in urban GIS”, Computers, Environment and Urban Systems, 24, pp.251-271, 2000

[8] R. Neches, et al, “Enabling technology for knowledge sharing”, AI Magazine, 12(3), pp.36-56, 1991

[9] T.R. Gruber, “A translation approach to portable ontology specification”, Knowledge acquisition, 5(2), pp.199-220, 1993

[10] R. Studer, V.R. Benjamins, D. Fensel, “Knowledge engineering, principles and methods”, Data and Knowledge Engineering, 25(1-2), pp.161-197, 1998

Analysis and Checking of Internet Banking Based on Safety Transition System

Wan Liang Huang Yiwang Li Xiang

Institute of Computer Software and Theory, Guizhou University, Guiyang, Guizhou 550025, China
Email: wanliangtr@163.com

Abstract

The temporal logic of actions is a logic for specifying and reasoning about concurrent systems. One kind of logic brought forward by Leslie Lamport[1]. And its syntax and complete formal semantics are summarized in about a page. TLA is extremely powerful, both in principle and in practice. In the process of researching Internet banking system we put forward safety transition condition, safety action and safety transition system based on TLA; then, specify Internet banking system using TLA+ which is based on safety transition system. The specification is checked by TLC, and the results show that the system based on safety transition is more secure.

Keywords: TLA; Safety transition condition; Safety action; Safety transition system; Internet banking

1 Introduction

With the networking vigorous development, the internet banking can develop rapidly. The internet banking is refers to the bank take own computer system as a main body, with the aid of the Internet technology, provides the bank service through Internet to the customer. The internet banking change traditional pattern, It provide for the bank customer any day, any time, low transaction cost, high grade convenient service. However in recent years, the false website, the fraudulent email, and the malicious wooden horse virus and so on, each new crime method emerged one after another incessantly, the bank security brought us the huge challenge.

At present many banks take measures to strengthen the security of internet banking such as the certificate & client side technology, the password card technology, the document certificate technology, the handset short note confirmation password technology, the U shield technology, and the digital certificate technology. These technologies strengthen the security greatly, but the present security problems mainly appear in the payment process.

We put forward safety transition condition, safety action, safety transition system a series of definitions and a theorem based on TLA(The Temporal Logic of Actions)[1][2][3]. In the safety transition system each state is safe which guaranteed the security. Internet banking system based on the safety transition has safety runs throughout the payment process, thus strengthens the payment process security. Then we use TLA+ [1] specify internet banking system based on the safety transition, and checking the specification with TLC [1]. The results show that the system based on safety transition is more secure.

2 Transition System Based On TLA And Safety Property

2.1 Transition system and TLA

The temporal logic of actions is a logic for specifying and reasoning about concurrent systems. One kind of new logic brought forward by Leslie Lamport[1]. Systems and their properties are represented in the same logic. Its syntax and complete formal semantics are summarized in about a page. Yet, TLA is not just a logician's toy; it is extremely powerful, both in principle and in practice.

Definition 1. A labeled transition system $T=(Q,I,A,L)$

a (finite or infinite) set of states Q,

a set $I \subseteq Q$ of initial states,

a set A of actions(action names), and

a transition relation $L \subseteq Q \times A \times Q$ [3].

Definition 2. An action is a Boolean-valued expression containing constant symbols, variables, and primed variables [1].

Definition 3. For any action A, we defines Enabled A to be the predicate that is true for a state iff it is possible to take an A step starting in that state [3].

Definition 3. A run of T is a (finite or infinite) sequence $\rho = q_0 \xrightarrow{A_0} q_1 \xrightarrow{A_1} q_2 \cdots$ where $q_0 \in I$ and $(q_i, A_i, q_{i+1}) \in L$ holds for all i [3].

2.2 Safety property

Definition 1. Let Q and A be sets of states and actions. A (Q,A) -property Φ is a set of ω -sequences $\sigma = s_0 \xrightarrow{A_0} s_1 \xrightarrow{A_1} \cdots$ where $s_i \in \Phi$, $A_i \in A$. We inter- changeably write $\sigma \in \Phi$, and $\sigma \models \Phi$ [3].

Definition 2. Given a sequence $\sigma = s_0 \xrightarrow{A_0} s_1 \xrightarrow{A_1} s_2 \cdots$, we write $\sigma[..n]$ to denote the prefix $s_0 \xrightarrow{A_0} s_1 \cdots \xrightarrow{A_{n-1}} s_n$ [3].

Definition 3. For a property Φ and a finite sequence $\rho = s_0 \xrightarrow{A_0} s_1 \cdots \xrightarrow{A_{n-1}} s_n$, we write $\rho \models \Phi$ iff $\rho \circ \sigma \in \Phi$ for some infinite sequence σ [3].

Definition 4. Φ is a safety property iff for any infinite sequence σ: $\sigma \models \Phi$ if $\sigma[..n] \models \Phi$ for all $n \in \mathbb{N}$ [3].

3 Safety Transition System

Firstly, we give the definition of initial safety states, safety transition condition, safety action and a theorem. Through these formalized terms of an agreement, we may specify the Internet banking based on safety transition.

Definition 1. We write $s.u$ to denote value mapping from state variable u in state s, and write $A.u$ to denote the quantity action A has changed variable u.

Definition 2. A state portrays by many state variable:

$s = \{ < i, u_i, s.u_i > \mid u_i \text{is state value}; s.u_i \text{is mapping from variable to value}; 0 \le i \le m, m \in \mathbb{N} \}$

Definition 1. $(s, A, s') \subseteq L$, the safety transition condition Γ is composed by the predicate $\varphi_i(s.u_i, Z_i(u_i))$ and the predicate $\eta_j(A.u_j, \pi_j(u_j))$:

$\Gamma = (\varphi_i, \eta_j)$.

$\varphi_i(s.u_i, Z_i(u_i))$: indicates the value $s.u_i$ of the state variable u_i and the value $Z_i(u_i)$ have some kind of relations, the value $Z_i(u_i)$ is decided by the system, and the state values satisfy this relation form the set Z_s.

$\eta_j(A.u_j, \pi_j(u_j))$: indicated the value $A.u_j$ of the state variable u_j and the value $\pi_j(u_j)$ have some kind of relations, the value $\pi_j(u_j)$ is decided by the system, and the state values satisfy this relation form the set π_s.

Definition 2. The action A to the state s is the safety action iff the predicate holds:

$$\Gamma / s.A = \forall u_i \left(\left(u_i \in Z_s \rightarrow \varphi_i \left(s.u_i, Z_i \left(u_i \right) \right) \right) \wedge \left(u_i \in \pi_s \rightarrow \eta_i \left(A.u_i, \pi_i . u_i \right) \right) \right) \equiv true$$

And we write $A_{\Gamma s}$ to denote the action A to the state s is the safety action, and we write Λ_s to denote all safety actions in a system T.

Definition 3. We write S_θ to denote the set of initial safety states decided by system. $(s, A, s') \subseteq L$, if s is safety state and the action A to s is safety action, then s' is safety state.

Definition 4. A safety transition system is a five tuple: $T_s = (Q, S_\theta, \Lambda_s, L_s, \Gamma)$

a (finite or infinite) set of states Q,

a set $S_\theta \subseteq Q$ of initial safety states,

a set Λ_s of safety actions,

a transition relation $L_s \subseteq Q \times \Lambda_s \times Q$ and

safety transition condition Γ.

Theorem 1. In a safety transition system T_s, all states are safety states.

Proof: assume a run $\rho = q_0 \xrightarrow{A_0} q_1 \xrightarrow{A_1} q_2 \cdots$

in T_s and assume state q_i is not a safety state.

(1) According to safety transition system definition, any action in T_s is safety action.

(2) According to definition 13 if q_i is not safety state and the action A to q_{i-1} be safety action, then q_{i-1} is not safety state.

(3) From this may promote q_0 is not safety state.

(4) According to the definition, $q_0 \in S_\theta$.

Q.E.D. (From (3),(4), by contradiction)

4 Analysis and Checking Security a nd Concurrency of Internet Banking Based on Safety Transition System

4.1 Analysis of Internet banking system based on safety transition

In the discussion, definitely the bank is the credible side; the bank is the entire fund management and the security process control side. The question leaves in Internet account instruction, if user's account number password is intercepted by the intruder, he can do anything he want, such as transferring accounts willfully to other accounts. Thus is cannot guarantee security of the payment process. As shown in Figure 1.

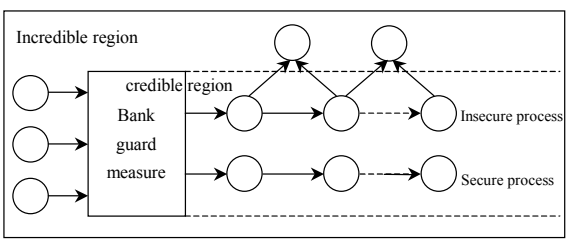


Figure 1 Internet banking model

Therefore strengthens the security facing the payment process is necessary. In order to achieve this goal, customer signs the agreement of safety transition conditions with the bank such as establishing the credible accounts and determining each account transfer quantity on the counter. It will form the initial safety state and safety transition conditions. When later on-line payment, actions satisfied the safety transition condition will be executed and actions dissatisfied will not be executed. Like this may form safety run in system, thus guaranteed the security facing the payment process. As shown in Figure 2.

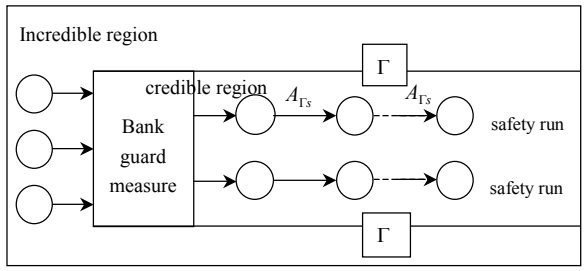


Figure 2 Internet banking model based on safety transition

4.2 Modeling and specifying with TLA+

Most TLA system specifications are of the form:

$Init \wedge \Box[Next]_v \wedge Liveness$ [1][2][3]

Init: state formula describing the initial state(s)

Next: action formula formalizing the transition relation usually a disjunction $A_1 \vee A_2 \vee \cdots A_n$ of possible actions (events) A_i

Liveness : temporal formula asserting liveness conditions usually a conjunction $WF_V(A_i) \wedge \cdots \wedge SF_V(A_j)$ of fairness conditions

4.2.1 Main constants and variables establishment

users = {1,2,3}: Expresses a main account has three users respectively is the user 1 to the user 3, or its under establishes three users.

AllAccounts = {0,1,2,3,4,5}: Expresses in the system altogether has six accounts, respectively is the account number 0 to the account number 5. Account number 0 expression current main account.

creAccounts = {0,1,2,3}: Expresses four credible accounts, the account number 1 to account number 3 is the credible accounts of the account number 0 which registers in the bank.

drawoutM: Expresses taking out or the account transferring amount.(In order to reduce the number of states, here takes 1 to 100 between numbers, other variable value also based on similar consideration)

conditionM: Expresses taking out quota.

saving: Expresses balance of current account.

ask: Array variable, expresses three user's requests, for example, ask[1]=”y” expresses user 1 has business to request.

pay: Array variable, with above correspondence, is to the request reply.

payAccount: Expresses already paying quantity.

4.2.2 Safety transition condition

the current account only carries on business with three credible accounts to process:

$$\varphi_1(s\_payAccount\_, Z_1(payAccount))$$
$$\Box\ payAccount \in creAccounts$$
$$\Box\ payAccount \in \{1,2,3\}$$

The account transfer business quota is 10:

$$\eta_2(A\_drawoutM\_, \pi_2(drawoutM))$$
$$\Box\ A\_drawoutM\_ \le 10$$

4.2.3 Intruder designed

the intruder may change information to remit account to incredible account, the amount also to be possible to change with willfully:

Intruder==/\ payAccount'= CHOOSE i \in AllAccounts: i >= 0

/\ drawoutM'=CHOOSE i \in (1..20): i>0

/\UNCHANGED<<saving,count,askcount, paycount, ask, pay, account, conditionM, in>>

5 Checking Content Design And Results

5.1 Checking content design

The model carries on the specifcation with TLA+, carries on the examination with TLC. Checking contents:

To all users, if business requested is responded, always carries on with the credible accounts:

\A u \in Users: [](pay[u]="y"=>payAccount \in cre Accounts)

To all users, if business requested is responded, payment quantity is always smaller than or is equal to the quota:

\A u \in Users: [](pay[u]="y" =>(drawoutM<=condi tionM))

The host account balance is always bigger than or is equal to the zero:

[](saving >=0)

In the bank the total payment is always bigger than or is equal to the payment quantity:

[](account>=paycount)

To all users, if the requested payment is smaller than or is equal to the remaining sum, can have the reply inevitably:

\A u \in Users: (saving>=drawoutM /\ ask[u]= "y")=> <>(pay[u]="y")

5.2 Main code

```
Renew_PVChan == /\ count=1
                /\ pvChan!Send("y")
                /\ count'=0
                /\ askcount'=askcount+1
                /\ drawoutM'=CHOOSE i \in (1..10): i>0
                /\ payAccount' =CHOOSE i \in
creAccounts: i>=0
                /\ ask'=[u \in Users |->"y"]
                /\ pay'=[u \in Users |->"n"]
                /\ UNCHANGED <<saving,paycount,
account, conditionM>>

Intruder == /\ payAccount'= CHOOSE i \in All
Accounts: i >= 0
/\ drawoutM'=CHOOSE i \in (1..20): i>0
                /\ UNCHANGED <<saving,count,
askcount, paycount,ask,pay,account,conditionM,in>>

RcvChan == /\ pvChan!Rcv
/\ UNCHANGED <<saving,drawoutM, count,
askcount,paycount,ask,pay,account,payAccount,
conditionM >>

Drawout(u) ==/\ askcount>0
                /\ ask[u]="y"
                /\ payAccount \in creAccounts \/
payAccount=0
```

```
        /\ drawoutM<=conditionM
        /\ account>=paycount+drawoutM
        /\ askcount'=askcount-1
        /\ pay'=[pay EXCEPT ![u] = "y"]
        /\ ask'=[ask EXCEPT ![u] = "n"]
        /\ in.val="y"
        /\ pvChan!Send("n")
        /\ saving' = IF (saving >= drawoutM) /\
(paycount + drawoutM <= account) THEN saving -
drawoutM ELSE saving
        /\ paycount' = IF (saving >= drawoutM)
/\ (paycount + drawoutM <= account) THEN paycount
+ drawoutM ELSE paycount
        /\ count' =1
        /\ UNCHANGED <<drawoutM,account,
payAccount,conditionM>>

Next == \/ Renew_PVChan
        \/ RcvChan
        \/ Intruder
        \/ \E u \in Users: Drawout(u)

PropertySafe1 == [](saving >=0)
PropertySafe2 == [](account>=paycount)
PropertySafe3 == \A u \in Users:
[](pay[u]="y"=> payAccount \in creAccounts)
PropertySafe4 == \A u \in Users:
[](pay[u]="y"=> (drawoutM<=conditionM))
PropertySafe5 == \A u \in Users:
(saving>=drawoutM /\ ask[u]="y")=><>(pay[u]="y")

vars == <<in, drawoutM, saving, count, askcount,
paycount, ask,pay,account,payAccount,conditionM>>
Spec == Init /\ [][Next]_vars /\ \A u \in Users:
SF_vars(Drawout(u)) /\ WF_vars(Next)
----------------------------------------------------------
THEOREM Spec => []TypeInvariant /\
PropertySafe1 /\ PropertySafe2 /\ PropertySafe3 /\
PropertySafe4 /\ PropertySafe5
```

5.3 Checking results

The results, show in figure 3, indicates Internet banking system based on safety transition is more secure. There are 216 states, 9 branches,43 depth and no error be found.

Under the safety transition condition (φ_1, η_2), action Drawout(u) causes a safety state to turn another safety state; The results show the properties are right, so the system satisfies the secure requests .

6 Acknowledgment

We thank everyone who helped us during all researching time, especially professor Li. He is also our teacher, a man of profound learning.

References

[1] Leslie Lamport. Specifying Systems. Addison -Wesley Longman Publishing Co., Inc. 2002

[2] Leslie Lamport. The Temporal Logic of Actions. ACM Transactions on Programming Languages and Systems.1994.5,16(3): 872-923

[3] Stephan Eerz. Modeling and Developing Systems Using TLA+. Escuela de Verano,2005

[4] Manna,Z. and Pnueli,A. The Temporal Logic of Reactive and Concurrent Systems. Springer-Verlag. 1991

[5] Alpern, B. and Schneider, F. B. 1985. Defining liveness. Information Pro-cessing Letters 21, 4 (Oct.), 181-185

[6] Apt, K. R. 1981. Ten years of Hoare's logic: A survey-part one. ACM Transactions on Programm -ing Languages and Systems 3, 4 (Oct.), 431-483

[7] Apt, K. R. and Olderog, E.-R. 1990. Veri_cation of Sequential and Con-current Programs. Texts and Monographs in Computer Science. New York,Berlin, Heidelberg, London, Paris, Tokyo, Hong Kong, Barcelona: Springer-Verlag

[8] Ashcroft, E. A. 1975. Proving assertions about parallel programs. Journal of Computer and System Sciences 10, 110-135

[9] Garland, S. J. and Guttag, J. V. 1989. An overview of LP, the Larch Prover.In N. Dershowitz (Ed.), Proceedings of the Third International Conferenceon Rewriting Techniques and Applications, Volume 355 of Lecture Notes onComputer Science, pp. 137{151. Springer-Verlag

[10] Hehner, E. C. R. 1984. Predicative programming. Communications of theACM 27, 2 (Feb.), 134-151

Research on the General Architecture of Ontology Learning System*

Kui Fu[1] Guihua Nie[2]

1 Department of Electronic Business, Wuhan University of Technology, Wuhan, Hubei 430070, China
Email: fukui@whut.edu.cn

2 Department of Electronic Business, Wuhan University of Technology, Wuhan, Hubei 430070, China
Email: niegh@whut.edu.cn

Abstract

This paper proposes a unified, open, extensible, domain independent, and scalable architecture of ontology learning systems. The unified functional structure and standardized access interfaces are defined for ontology learning systems. Our architecture consists of resource layer and five main function modules: resource management module, general resource read/write module, data preprocessing module, ontology extraction module and ontology evaluation and editing module. Functions and constitutions of resource layer and all modules are depicted in detail. Its main goal is to provide support for reuse of ontology learning methods and components, ontology industrialization producing, periodical ontology refinement.

Keywords: ontology; ontology learning; knowledge acquisition

1 Introduction

Ontology can support information exchange, knowledge sharing and reuse between human and machine, machine and machine, therefore, it gained increasing attention, research and application. However, the scarcity of domain ontology is one of the main bottlenecks which plagued ontology theoretical research and practical application, so ontology learning emerged as the times require. It is able to acquire ontology from various different data sources, using automatically or semi-automatically machine learning methods.

During the past few years some research works have been done in the field of ontology learning and some ontology learning tools and systems, such as TextToOnto[1], OntoLearn[2], OntoLift[3], Hasti[4] or OntoBuilder[5], are proposed. These systems may exploit statistical methods[6,7], linguistic methods[8-10] or hybrid of both to learn ontologies. Though some fruits are produced in these ontology learning systems, there are following problems which need to be solved. Firstly, most of above systems may usually learn ontological knowledge only from one type of data sources which include unstructured data source such as natural language texts, semi-structured data source such as XML and HTML documents or structured data source such as relational patterns. Secondly, lack of unified definition of functional structure and standardization of access interfaces results in non-reuse components of ontology learning systems. Thirdly, They assist on learning some parts of ontology but they are not able to learn how to do the learning process better. Finally, They are almost prototype systems without the ability of large-scale ontology acquisition.

As a solution of above problems, we present a unified, open, easily extensible, domain-independent, and scalable architecture of ontology learning systems, which consists of resource layer and five main function modules. Our main goal is to assist reuse of ontology learning methods and components, ontology industrialization producing, periodical ontology refinement. The rest of the paper is organized as follows:

* This research is supported by National Key Project of Scientific and Technical Supporting Programs under Grant NO.2006BAH02A08.

Section 2 presents an overview of our ontology learning system architecture. In Section 3, we introduce resource layer of ontology learning system. Section 4 details five main function modules of ontology learning system. Finally, conclusions are discussed in Section 5.

2 Overview of Proposed Architecture of Ontology Learning System

2.1 Design Principles of the Architecture General Appearance

1) The architecture must provide support for reuse of ontology learning methods or components.

2) The architecture must assist on ontology acquisition from various heterogeneous data sources.

3) The architecture has the ability of large-scale ontology learning.

4) The architecture must support the evaluation and refinement of ontology learning results.

5) The architecture has the self-study ability at a certain extent.

2.2 Proposed Architecture of ontology learning system

Our architecture of ontology learning system aims to reach the following main goals: to assist on reuse of ontology learning methods and components, to support ontology industrialization producing and ontology refinement, and to have the self-study ability, by means of the introduction of an iterative feedback into the system. As shown in Figure 1, the architecture is composed of resource layer and five main function modules: resource management module, general resource read/write module, data preprocessing module, ontology extraction module and ontology evaluation and editing module.

The architecture is hybrid in the sense that it uses different types of data sources, and provides an effective combination of different methods for extracting and analyzing information in data sources. The architecture has a modular design, composed by a set of modules covering all the steps of the ontology construction process. This design facilitates the extensibility and reusability of the components.

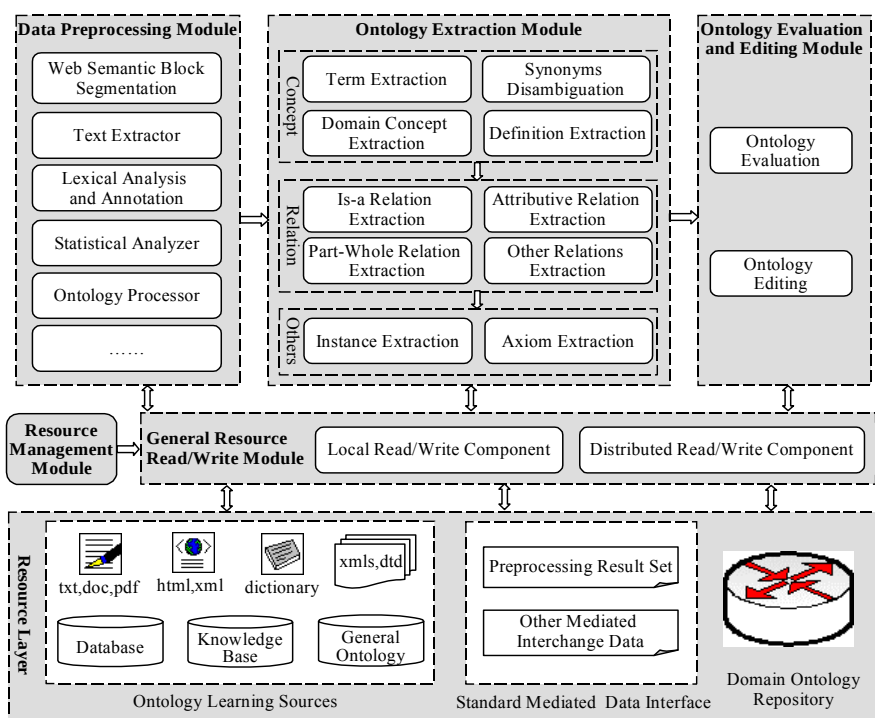


Figure 1 The Architecture of Ontology Learning System

3 Resource Layer

Resource layer is the aggregation of input data and output results of other modules in ontology learning systems. It is composed of there portions: ontology learning sources, standard mediated data interface and domain ontology repository. Data in the resource layer can be stored in either local systems or distributed computing platforms. Universal access interfaces to resource layer are provided by general resource read/write module, so that other modules can easily read or write data in the resource layer without considering complicated data structures of heterogeneous resources.

Ontology learning sources are the initial input of ontology learning systems. These sources can be divided into structured, semi-structured and unstructured sources with the difference of structured degree. There are familiar categories of ontology learning sources in Figure 2. Notably, html files can be viewed as unstructured sources or semi-structured sources according to different ontology learning methods.

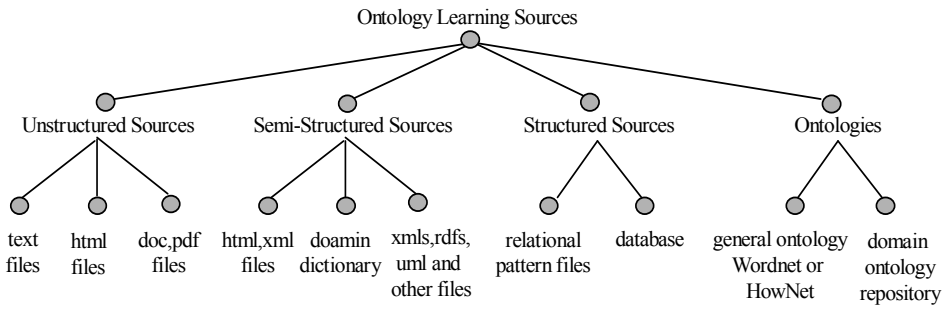


Figure 2 Categories of Ontology Learning Sources

Standard mediated data interfaces define the standard data formats according to the requirement of input and output of each module of ontology learning systems, which could support the mediated data interchanging among different modules. Standard mediated data interfaces are the basis of reuse of ontology learning components in a ontology learning system or among different ontology learning systems.

Domain ontology repository is the final arts of ontology learning systems. Because ontology learning is a continuously improving process, the ontology learning results of last period can be used to improve ontology learning of the next period

4 Five Main Function Modules

4.1 Resource Management Module

Resource management module is used to manage and maintain resources in resource layer. Its main functions include: resource crawling, resource browsing and maintenance, resource changes detection, version management and so on. It's the most important role to gather ontology learning resources.

The World Wide Web is a vast and growing source of information and web pages in the Internet are the main resources for many ontology learning systems, therefore, resource management module is designed to assist on the crawling of web pages. On the other hand, dynamic acquisition of learning resources is often required in the other phrases or modules of ontology learning, so resource manage module needs to meet the requirements of web pages' crawling and management from upper application modules.

4.2 General Resource Read/Write Module

General resource read/write module is the mediated component, which provides support for accessing various data sources in resource layer for other modules of ontology learning systems. It supports both the reading/writing data, and the definition of resource type.

This module shields the complicated methods of accessing to various heterogeneous resources. There are two benefits of separating resource read/write module from other modules: independence of ontology learning algorithms and possibility of large-scale ontology learning.

General resource read/write module is composed of local read/write component and distributed read/write component. The local read/write component supports the reading and writing of ontology learning resources under the concentrated environment. The distributed read/write component supports the reading and writing of ontology learning resources under the distributed environment.

In order to support large-scale ontology learning and ontology industrialization producing, we implement our distributed read/write component on the basis of the distributed computing platform, Hadoop[11], which is open-source distributed file storage system and distributed task execution system provided by the Apache Software Foundation. Our distributed read/write component has the ability of reading and writing resources under the distributed environment, and provides strong support for mass data processing of large-scale ontology learning systems.

4.3 Data Preprocessing Module

The main function of data preprocessing module is to adaptively choose appropriate methods and processes according to different data sources. Adaptive data preprocessing includes text extraction, syntactic analysis, statistical analysis, html structure analysis and so on. The analyzing results are stored in the standard mediated data interfaces through general resource read/write module. These results can be further used by ontology extraction module.

Data preprocessing module is the aggregation of familiar data preprocessing techniques often used in ontology learning systems. It consists of the following sub modules: web semantic block segmentation, text extractor, lexical analysis and annotation, statistical analyzer, ontology processor and so on. Web semantic block segmentation analyses the html structure of web page in order to detect and segment the main semantic blocks. The function of text extractor is to extract the natural language texts from txt files, web pages, doc files and pdf files. Lexical analysis and annotation's function is to split natural language texts into words and annotation the part of speech of each word. Statistical analyzer is responsible to calculate the frequency of words and other usable statistical information. Besides, data preprocessing module provides support for the preprocessing of dictionaries, data bases, knowledge bases and ontology repositories.

4.4 Ontology Extraction Module

Ontology extraction module is the most import module of whole learning system, which implement the core functions of ontology learning. It is able to acquire main elements of domain ontology using some ontology learning methods and policies. The input data of ontology extraction module is not the original data sources in resource layer, but standard mediated interchange data produced after preprocessing of the original data. Standard mediated data interfaces make it possible for the reuse of ontology learning methods and components. Besides, ontology extraction module may improve ontology learning with the help of general ontologies, such as WorNet and HowNet, domain dictionaries, knowledge bases or domain ontology learned in last learning period.

According to the difference of ontology learning tasks, ontology extraction module can be divided into three types of sub modules: concept extraction sub module, relation extraction sub module and other sub module. These sub modules are detailed in the following.

Concept extraction sub module is composed of term extraction component, synonyms disambiguation component, domain concept extraction component and definition extraction component. Term extraction component's function is to identify candidate terms. The difficulty of term extraction lies in automatic identification of compound terms. A concept may be

represented by different terms which compose a synonym set. Synonyms disambiguation component is used to disambiguate synonyms among candidates. Domain concept extraction component may calculate the domain relevance and domain consensus of extracted terms, and then filter irrelevant candidates according to the value of domain relevance and domain consensus. The results of domain concept extraction component are domain concept sets. Definition extraction component's function is to automatically acquire extracted concepts' definition.

Relation extraction sub module consists of is-a relation extraction component, attributive relation extraction component, part-whole relation extraction component and other relations extraction component, which are respectively able to extract is-a relation, attributive relation, part-whole relation and other relations among domain concepts. Other relations describe the domain-specific relations, such as the purchase relation between consumer concept and product concept.

Other sub modules mainly include instance extraction component and axiom extraction component. Instance extraction component can learn instances of domain concepts and relations between concepts. Axiom extraction component is used to extract axiom knowledge in the domain.

4.5 Ontology Evaluation and Editing Module

The ontology acquired through ontology extraction module should be refined and reorganized periodically. In this phase some new concepts will be created, new relations will be established between concepts and some old concepts will be merged, old and relations will be discarded. The main function of ontology evaluation and editing module is to automatically calculate the probability and confidence value of the learning results and to provide support for automatic ontology refinement or semi-automatic ontology refinement under ontology engineer's guidance. This module consists of ontology evaluation sub module and ontology editing sub module.

The functions of ontology evaluation sub module are to automatically evaluate the ontology learning results. Some ontology evaluation methods, such as probabilistic methods and gold standard, are provided in this sub module. Ontology learning systems may automatically execute the refinement task according to the evaluation results. Ontology engineers may also semi-automatically refine ontolgies according to the evaluation results.

Ontology editing sub module provides user interfaces to manually edit ontologies for ontology engineers. This sub module may implement its function through integrating existing ontology editing plug-ins, such as Protégé and OntoEdit.

5 Conclusion

This paper proposes a unified, open, extensible, domain independent, and scalable architecture of ontology learning systems. The architecture is composed of resource layer and five main function modules: resource management module, general resource read/write module, data preprocessing module, ontology extraction module and ontology evaluation and editing module. Functions and constitutions of resource layer and all modules are discussed in detail. The design and definition of unified functional structure and standardized access interfaces of learning systems have the ontology systems to posses the following virtues. Firstly, it provides support for ontology acquisition from various heterogeneous data sources. Secondly, it supports the reuse of ontology learning methods and components. Thirdly, it assists on ontology industrialization producing, ontology evaluation and periodical ontology refinement. Finally, it has the self-study ability through producing a feedback to the previous phases for future uses.

References

[1] Maedche A. Ontology Learning for the Semantic Web. Boston: Kluwer Academic Publishers, 2002

[2] Missikoff M, Navigli R, Velardi P. Integrated approach for web ontology learning and engineering. IEEE Computer,

2002,35(11), pp.60-63

[3] Volz R, Oberle D, Staab S, Studer R. OntoLiFT prototype. IST Project 2001-33052 WonderWeb Deliverable 11. 2003

[4] Shamsfard M., Barforoush A. Learning ontologies from natural language texts. International Journal of Human Computer Studies, 2004,60 (1), pp.17-63

[5] Modica G, Gal A, Jamil H.M. The use of machine-generated ontologies in dynamic information seeking. In: Proc. of the 9th Int'l Conf. on Cooperative Information Systems. Heidelberg: Springer-Verlag, 2001. 433-448

[6] Agirre E., Ansa O., Hovy E., and Martinez D. Enriching very large ontologies using the WWW. Proc 1st Workshop on Ontology Learning OL '2000. Berlin, Germany: CEUR Workshop, 2000

[7] K.T. Frantzi, S. Ananiadou.The C-Value/NC-Value domain independent method for multi-word term extraction. Journal of Natural Language Processing, 6(3):145-179,1999

[8] A. Smeaton, Quigley. Experiments on using semantics distances between words in image caption retrieval. In Proc. of 19th International Conrerence on Research and Development in Information Retrieval, Zurieh, Switzerland,1996

[9] Morin E. Automatic acquisition of semantic relations between terms from technical corpora. Proc 5th Int Congress on Terminology and Knowledge Eng. Vienna: TermNet, 1999

[10] Navigli R, Velardi P, Gangemi A. Ontology learning and its application to automated terminology translation. IEEE Intelligent Systems, 2003,18(1):22-31

[11] Hadoop Documentation. 2007, from http://lucene. apache. org/hadoop/docs/r0.15.2/index.html

Study on Minimum Exact Cover Problem of Group Key Distribution

Yaling Lu

Department of Electronic Information Engineering, Wuhan Polytechnic University, Wuhan, 430023, P.R. China
Email: colorfullu@hotmail.com

Abstract

To renew group key is necessary to ensure secrecy of multicast contents. However how to distribute updated key to all legitimate members efficiently is a hard problem in opening research. All existing works have complexity of O(*c*log*n*) in batch re-keying, where *c* is total additions/evictions. The Minimum Exact Cover Problem of Group Key Distribution is presented and studied, which can reduce the complexity down to O(1). Efficiency analysis and simulation test show that the achievement can improve efficiency of any tree-based group key management.

Keywords: minimum exact cover problem; key distribution; secure multicast

1 Introduction

Secure multicast is an efficient method for many emerging group oriented applications. All legitimate members in a multicast group share a session key (SK)[1]. The SK must change dynamically for user additions or member evictions to ensure forward secrecy or backward secrecy of multicast sessions[2]. It requires each updated SK be distributed only to the current legitimate members.

The group key distribution problem has been studied extensively. The key-tree based schemes are of practical interest for a variety applications because of its balance between communication complexity and storage complexity[3]. However, both group controller (GC) and each member should take O(*logn*) steps to finish group re-keying for each addition/eviction. So, all previous works have difficulties in batch re-keying for massive changes for the complexity of O(c*logn*), where c is total changes[4].

This paper aims at how to improve efficiency of group key distribution. The main content is organized as follows: Section 2 defines partial covering problem of group key distribution(GPCP) and summarizes existing works. Section 3 study on minimum exact GPCP and its solution. Section 4 evaluates its efficiency and compares it to typical schemes. Finally, Section 5 concludes the paper.

2 Partial Covering Problem of Group Key Distribution

2.1 Partial Covering Problem of Group Key Distribution

Figure 1 is an example of binary key tree. In the key tree, leaves present group members, the root is SK or data encrypt key (DEK), and all middle nodes are key encrypt key (KEK). Whenever the SK is updated, it should be distributed to all legitimate members secretly.

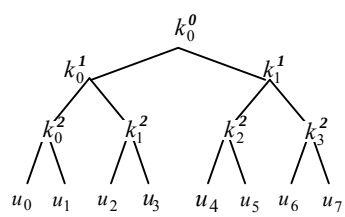


Figure 1 Logical key tree

Take the key tree as a graph $G(V,E)$, where $V=\{\{Root\},\{MiddleNodes\},\{Leaves\}\}$, $E=\{P_{i,j}\}$, $\forall i,j\in V$.

Let assign any node X to a tree **G(X),** which covers a set of leaves. Hereafter, a problem of which nodes should be selected to cover all legitimate members without any secure weakness is referred to Set Covering Problem [5].

Definition: Given a group key tree $G(V,E)$, a set of legitimate users **GM**, and an integer k

1. Partial covering set: a subset $\boldsymbol{GN} \subseteq \boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$ contains several nodes which cover all leaves in **GM**.

2. Partial covering problem[6]**:** whether there is a sub set $\boldsymbol{GN} \subseteq \boldsymbol{G}(\boldsymbol{V},\boldsymbol{E})$, which satisfies the equation

$$\begin{cases} \bigcup_{\mathbf{S}i \in GN} \mathbf{S_i} = \bigcup_{\mathbf{S}i \in \mathbf{G(V,E)}} \mathbf{S_i} = \mathbf{GM} \\ |\mathbf{GN}| \le k \end{cases}$$

It means that its solutions should assure that GN should cover all members in **GM** but illegitimate ones.

Existing works

There are two typical methods of group key distribution: Logical key hierarchy tree (LKH) [2,7] based schemes and One-way function tree (OFT) [8,9] based schemes, which named as top-down method and bottom-up method respectively in [9]. To explain the methods, here gives some notations:

U_a : Leaf node of user U_a

$S(X)$: The sibling node of node X.

P_a : Ancestor nodes of U_a (e.g. $P_1 = \{N_0^3, N_0^2, N_0^1\}$).

S_a : Sibling nodes of U_a , $S_a = S(P_a) \bigcup \{S(U_a)\}$.

Let C present the set of changing members, $c = |C|$. In top-down method, each updated upper node is distributed by its child nodes, till the root, so $GN = \bigcup P\{C_i\} \bigcup S\{C_i\}$, $\forall j : j = [1, c]$. Whereas in bottom-up method, each upper node is generated by two blinded keys of child nodes and each blinded key is distributed by its sibling node, so $GN = \bigcup S\{C_i\}$. From point view of distribution, top-down method has distribution complexity of $O(2c\log(n))$, each member should be covered from c to $c + \log n$ times, whereas bottom-up method has complexity of $O(c\log(n))$ and each member covered c times. It makes think out a way to lessen distribution complexity.

3 Minimum Exact Covering Problem of Group Key Distribution

3.1 Problem description

The best solution to group key distribution is to finish it in one step and to cover each legitimate member only once. That allows defining the minimum exact covering problem of group key distribution (GMECP).

Definition:

Sub-tree: $G(S)$ is a sub-tree of $G(T)$, if all members covered by $G(S)$ are covered by $G(T)$.

Conjoint node: node S and T are conjoint if $G(S)$ is contained in $G(T)$.

Disjoint tree: $G(S)$ and $G(T)$ are disjoint, if any of members covered by $G(S)$ is not covered by $G(T)$.

Independent set: Given $GN \subset G(V,E)$, GN is an independent set if any two nodes in it are not conjoint.

Minimum coverage: GN is a minimum covering set, if there isn't another exact part cover $GN' \subset G(V,E)$ which makes $|GN'| < |GN|$.

GMECP: Let L denote $\{leaves\}$. Given $T = G(V,E)$, $GM \in L$, whether there exists a set $GN \in G(V,E)$ with minimum cardinality which covers the GM exactly--each member in GM is covered only once and any member out GM mustn't be covered.

Let:

$L = \{L_1, L_2, ...L_n\}$

$V = \{V_1, V_2, ...V_{2n+1}\}$

$$x_j = \begin{cases} 1, & \text{If node } V_j \text{ is selected, } \forall V_j \in V \\ 0, & \text{Otherwise} \end{cases}$$

$$c_{i,j} = \begin{cases} 1, & \text{If } V_j \text{ covers leaf } L_i, \ \forall L_i \in L, \ \forall V_j \in V \\ 0, & \text{Otherwise} \end{cases}$$

The integer linear formulation of the problem can be stated as:

Target: $\min \sum_{j=1}^{|V|} x_j$

Subject to $\begin{cases} \sum_{j=1}^{|V|} c_{i,j} = 1, \ \forall L_i \in L \\ \sum_{j=1}^{|V|} c_{k,j} = 0, \ \forall L_k \in L / GM \end{cases}$

then $GN = \{V_j, \forall x_j = 1\}$ is the minimum coverage.

Any node in the tree $G(V,E)$ can be presented by set of members covered by it. Let $G_C(X)$ denote all members covered by tree $G(X)$. We can show that the GMECP is NP hard. Before giving the proof, we first introduce a well-studied NP-hard problem: Exact set covering problem[10].

Instance: Collection $F = \{S_1, S_2, ..., S_k\}$

Solution: A covering set for S, i.e., a subset $T \subseteq F$, which consist of several mutually exclusive sets.

MEASURE: $\bigcup_{\mathbf{S}i \in T} \mathbf{S_i} = \bigcup_{\mathbf{S}i \in F} \mathbf{S_i} = \{u_1, u_2, ..., u_n\}$

Proposition 1: The **GMECP** is NP-hard.

Proof: Restating the optimization problem **GMECP** as a decision problem, we want to determine if there is *k* sub-trees which can cover all legitimate members within lest cost. From the point view of coverage, we can take the tree $T = G(V,E)$ as a collection of V, which represented by the set of member $G_C(V_i), \forall V_j \in V$. So, the problem can be stated exact covering set problem as follows: whether can find a sub-set $GN \subseteq V$ consisted of several independent sets which can cover the given $GM \in L$.

Then the objective of finding the minimum number of sub-trees to cover the given member set **GM** is equivalent to minimizing the cardinality of set GN', where $GN' \subseteq GN$ such that every element in GM belongs to only one member of GN'. For L is a special case of given **GM**, the problem GMECP is NP-hard. This completes the proof.

3.2 Characters of GMECP

From the point of view of covering, the logical key tree is a special sorted set, some features of it can be illustrated.

Definition: 1 R_1 and R_2 are conjoint, if $G_C(R_1) \bigcap G_C(R_2) \neq \Phi$.

2 R_1 and R_2 are disjoint, if $G_C(R_1) \bigcap G_C(R_2) = \Phi$.

Proposition 2: Given a tree or sub-tree $G(R)$ and GM. $GN = R$, if $GM = G_C(R)$.

Proposition 3: Given a tree or sub-tree $G(R)$ and a user U_a, $G_C(R) = \{U_a\} \bigcup \{\bigcup G_C(S_a)\}$.

Proposition 4: Given a tree $G(V,E)$ and a node $N \in G(V,E)$, $G_C(V,E) = G_C(N) \bigcup \{\bigcup G_C(S(N))\}$

Proposition 5: Suppose that R_1 and R_2 are disjoint, $R_3 \in G(V,E)$. if $G_C(R_3) \subset G_C(R_2)$, R_1 and R_3 are disjoint.

Proposition 6: Given a tree $G(V,E)$ and a minimum covering set GN, $G_C(GN[i]) \bigcap G_C(GN[j]) = \Phi (i \neq j)$. In other words, all nodes in GN must be disjoint.

Proof: Obviously, there exist only two relationships between any two nodes in the binary tree: disjoint and conjoint. Any two nodes in a path node are conjoint, otherwise, they are disjoint.

If $G_C(R_1) \bigcap G_C(R_2) \neq \Phi$, R_1 and R_2 must be conjoint, so $G_C(R_1)$ *and* $G_C(R_2)$ must have a relationship of contain-contained. Suppose $G_C(R_1) \subset G_C(R_2)$,

then $\sum_{j=1}^{|V|} c_{i,j} > 1, \ \forall L_i \in G_C(R_1)$

which is in conflict with the definition of GMECP. So all node in GN must be disjoint, if the GN is the minimum exact cover set.

Features: 1 any two nodes in $\{\{U_a\}, P_a\}$ are conjoint, for $\{U_a\} \bigcap \{\bigcup G_C(P_a[i])\} = \{U_a\} \neq \Phi$, $\forall i \in [i, |P_a|]$.

2 any two nodes in $\{\{U_a\}, S_a\}$ are disjoint, for $G_C(S_a[i]) \bigcap G_C(S_a[j]) = \Phi$, $(i \neq j), \forall i, j \in [i, |S_a|]$.

3 $P_a \bigcap S_a = \Phi$.

4 $G_C(P_a[i]) \subset G_C(P_a[j])$ $(i < j)$ (we assume that the elements are sorted from the bottom to the root).

3.3 Solution to GMECP

As mentioned above, the GMECP is NP-hard, the time complexity of traversal searching is $O(n^{c\log(n)})$. However, the logical key tree is special sorted sets from point view of distribution coverage, so we can give the optimal solution by limiting to search best results in the set of all sibling nodes of members which should not be covered. The greedy algorithm shows as follow:

Obviously, the time complexity of the greedy algorithm is not polynomial, which costs:

$$O(\sum_{i=1}^{c\log(n)} C_{c\log(n)}^{i}) \approx O(c\log(n)^{c\log(n)})$$

(where $c = |\overline{GM}|$, $c << n$)

Greedy Algorithm:

$\overline{GM} \leftarrow L/GM$

$DS \leftarrow S(\overline{GM})$

$i \leftarrow 0$

while $i < |DS|$ do

$i++$; $GN \leftarrow \Phi$; $t \leftarrow \{Root\}$

for all $t \in DS$ do

$t \leftarrow \{t_j \in DS : \forall j = [1,i]\}$

if $\bigcup G_C(t) = GM$

$GN \leftarrow t$

Break

Endif

Endfor

Endwhile

In many multicast scenarios, group size is very big and group members change very frequently. Batch handling method is applied to improve efficiency of group key management, so c is always big, and the greedy algorithm will become time-consuming. To solve this issue, we can give heuristic solution to it based on the definition and some features of GMECP.

Heuristic Solution: according to the definition of GMECP, if $G_C(A_1) \subset G_C(A_2)$ and A_2 is selected as distributing node, A_1 should not be selected. So we can improve the greedy algorithm more. Some notations adopted in our solution are as follows:

$F(X)$: The parent node of node X.

$F(U_a, U_b)$: The lowest shared ancestor node of U_a and U_b, $F_{a,b}$ in brief. For example, $F_{2,7} = N_0^1$.

$S_a(F)$: Nodes in S_a under the node F.

$S_{a,b}(N)$: $S_{a,b}(N) = S_a(N) \bigcup S_b(N) - C(F)$.

We descript the solution by following theorems:

Theorem 7: when SK is distributed to all members except member U_a, $\boldsymbol{GN} = S_a$

It can be deduced from the theorem 3 and feature 2. For $\{\bigcup G_C(S_a)\} = G_C(R) - \{U_a\}$ and any two nodes in S_a are disjoint, any legitimate members are covered only once and the U_a is not covered. So, $\boldsymbol{GN} = S_a$.

Lemma 1: when SK is distributed to all members except members U_a and U_b,

$\boldsymbol{GN} = S_a \bigcup S_b - C(F_{a,b}) = S_{a,b}(Root)$

Proof: Let $\overline{G_C}(F_{a,b})$ denote $\{\bigcup G_C(S(F_{a,b}))\}$. From theorem 4, we can see $G_C(V,E) = G_C(F_{a,b}) + \overline{G_C}(F_{a,b})$. Let $LG(F_{a,b})$ and $RG(F_{a,b})$ denote the left sub-tree and the right one of $G(F_{a,b})$ respectively. Hence, $G_C(V,E) = LG_C(F_{a,b}) + RG_C(F_{a,b}) + \overline{G_C}(F_{a,b})$

$S_a(F)$, $S_b(F)$ and $S_a \bigcap S_b$ are the minimum exact coverage respectively for the right three parts of the equation. So,

$$\begin{aligned} GN &= S_a(F) \bigcup S_b(F) \bigcup (S_a \bigcap S_b) \\ &= S_a \bigcup S_b - C(F) \\ &= S_{a,b}(Root) \end{aligned}$$

Deduction 1: when SK is distributed to all members except $\{M_i, \forall i: i = [1,m]\}$, $GN = \bigcup S_{Mi} - \bigcup C(F_{Mi,Mi+1})$ (supposed that $\{M_i\}$ has been sorted on their position in the key tree).

Proof: we can obtain $\{F(M_i, M_{i+1}), \forall i : i = [1,m)\}$ and $F(M_1, M_m)$. We divide all legitimate members into m parts according the positions of $\{M_i, \forall i: i = [1,m]\}$, then

$G_C(V,E) = G_C(F(M_1,M_m)) + \overline{G}_C(F(M_1,M_m))$

since $G_C(F(M_1,M_m)) = \bigcup G_C(F(M_i,M_{i+1}))$, and $G_C(F(M_i,M_{i+1})) = LG_C(F(M_i,M_{i+1})) + RG_C(F(M_i,M_{i+1}))$, from the deduction 1, we can see

$$\begin{aligned} GN &= \bigcup S_{Mi,Mi+1}(F(M_i,M_{i+1})) + S_{F(M_1,M_m)} \\ &= \bigcup S_{Mi} - \bigcup C(F_{Mi,Mi+1}) \end{aligned}$$

$(i = 1,2,...m-1)$

According to these solutions, we can give an improved heuristic algorithm.

The Heuristic Algorithm searches the key tree level by level, so its time complexity is $O(n.c.\log(n))$ (where $c = |\overline{GM}|$).

Heuristic Algorithm:

$\overline{GM} \leftarrow L/GM$

$I \leftarrow P(\overline{GM})$

$S \leftarrow Root$

while $S \neq \Phi$ do

while $S_i \in S$ do

if $G_C(S_i) \bigcap I \neq \Phi$

insert child nodes of S_i into S

else

$GN \leftarrow S_i$

remove S_i from S

endif

endwhile

endwhile

4 Simulation and Efficiency Analysis

In this section, we compare the performance of my scheme GMECP and other schemes by both analysis and simulations.

The system is described as fellows:

1) Assume that the user-join numbers and the user-leave number are the same. This assumption may not be accurate for multicast service, but the advantages of GMECP over others do not seem to be sensitive to the choice of the model of user changing.

2) Both schemes apply the same user model to test efficiency, in which the leaf nodes of leaving users distribute in key tree randomly.

3) Both apply SHA1 to generate and renew keys, which length is 160 bits.

Key criteria of distribution efficiency focus on computation cost and scalability. The main factors include time and communication complexity.

Figure 2 shows member pattern applied in the simulation test. It can simulate many multicast applications, in which addition is more than eviction at first and contrarily in the end. Figure 3 and Figure 4 show that GMEC is superior to LKH+ and OFT+ in time cost and communication cost respectively, for which need distribute $O(c.\log(n))$ keys in $O(c.\log(n))$ steps, while the GMEC need distribute only root key in $O(1)$ step.

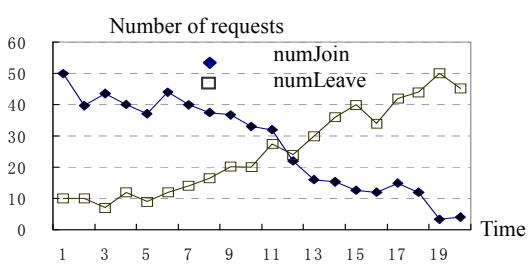


Figure 2 Member change pattern

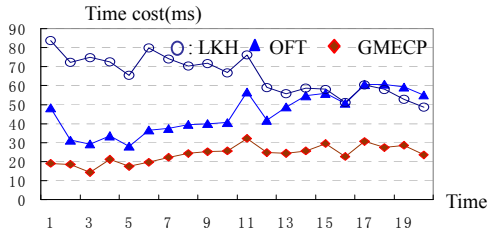


Figure 3 Comparison of time cost

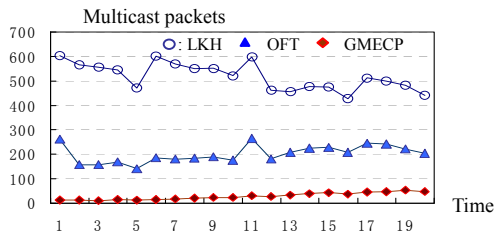


Figure 4 Comparison of distribution complexity

As for the client members, everyone in GMECP need only one step to get the new SK in each group rekeying circle, while it need $O(c.\log(n))$ steps in other schemes. The benefit of it is very important to improve QoS in many applications, such as multimedia multicasts, bandwidth or time cost restricted scenarios.

5 Conclusion

This paper gives the solution to GMECP, which can be applied to any tree-based group key management. Scheme based on GMECP can operate group rekeying more efficiently than other schemes no matter how dynamic a group is, which is very important to assure QoS for users in multimedia application scenario.

References

[1] H. Harney and C. Muckenhirn. Group Key Management Protocol (GKMP) Specification. RFC 2093 [Z]. 1997

[2] D. Wallner, E. Harder, and R. Agee. Key Management for Multicast: Issues and Architectures. RFC 2627 [Z] . 1999

[3] Sandro Rafaeli, David Hutchison. A survey of key management for secure group communication[J]. ACM Computing Surveys, 2003, vol.35(3)

[4] Duma C, Shahmehri N, Lambrix P. A hybrid key tree scheme for multicast to balance security and efficiency requirements [A]. 12th IEEE International Workshop on

Enabling Technologies - Infrastructure for Collaborative Enterprises[C]. Los Alamitos (USA), CA: IEEE Computer Society Press, 2003: 208 –213

[5] W C Huang, C Y Kao, J T Horng. A genetic algorithm approach for set covering problems[A]. Proceedings of the First IEEE Conference on Evolutionary Computation[C], Los Alamitos (USA), CA: IEEE Computer Society Press, 1994, vol.2:569 – 574

[6] Gandhi, Rajiv; Khuller, Samir; Srinivasan, Aravind. Approximation algorithms for partial covering problems[J]. Journal of Algorithms, 2004, vol.53(1): 55-84

[7] J. Pegueroles, F. Rico-Novella, J. Hernandez-Serrano, et al. Improved LKH for batch rekeying in multicast groups[A]. International Conference on Information Technology [C]. New York (USA): IEEE , 2003. 269 – 273

[8] D. Balenson, D. McGrew and A. Sherman. Key Management for Large Dynamic Groups: One-Way Function Trees and Amortized Initialization. IETF Internet draft [Z]. 2000

[9] A.T. Sherman, D.A. McGrew. Key establishment in large dynamic groups using one-way function trees[J]. IEEE Trans. on Software Engineering, 2003, vol. 29(5):444 – 458

[10] Chen, Jianer; Kanj, Iyad A.; Xia, Ge. Labeled search trees and amortized analysis: Improved upper bounds for NP-hard problems[J] Algorithmica , 2005, v 43(4, p 245-273)

Yaling Lu (1972-), female, from Hubei Yingshan, lecturer, master, major study fields are signal processing and image processing

Study and Implementation of University Information Portal Platform based on Web Service*

Deyu Kong Yuansheng Lou Lei Lu Hongtao Xu

College of Computer & Information Engineering, Hohai University Nan Jing, Jiang Su 210098, China
Email: kongdeyu@hotmail.com, wise.lou@163.com, qingchun723@yahoo.com.cn

Abstract

University Information Portal (UIP) is an information integration platform for all school. We can effectively integrate the campus information resources, and provide a convenient and efficient unified entrance via the mode of single sign-on (SSO) to assess the University information service. At the same time, it provides personalized information services for teachers and students by management style, page settings and other functions. This paper presents the framework and implementation of UIP, which includes portal technology, Web services, identity authentication and resources integration.

Keywords: Portal; Web Service; Identity Authentication; Resource Integration

1 Introduction

Currently, most colleges and universities have been established the digital campus information systems, which could offer independent information services to students and teachers. But most of the service, which are provided in a single business manner, can not integrate the dispersed applications and resources of the university, and share information. At the same time it is also lack of identity authentication and individualized service. Therefore, this paper presents a Web service-based university information portal platform, which offers services to all students. And we attain campus information services via the effective integration platform, which is accessed in the manner of SSO [1] that accommodates the convenient unified login interface and personalized information services.

2 UIP Integration Framework

2.1 UIP Integration Architecture

This part presents a framework of the Webservice-based UIP system. It adopts web-based multi-layer architecture, which includes user layer, web server layer, application server layer and resource layer. As shown in Figure 1.

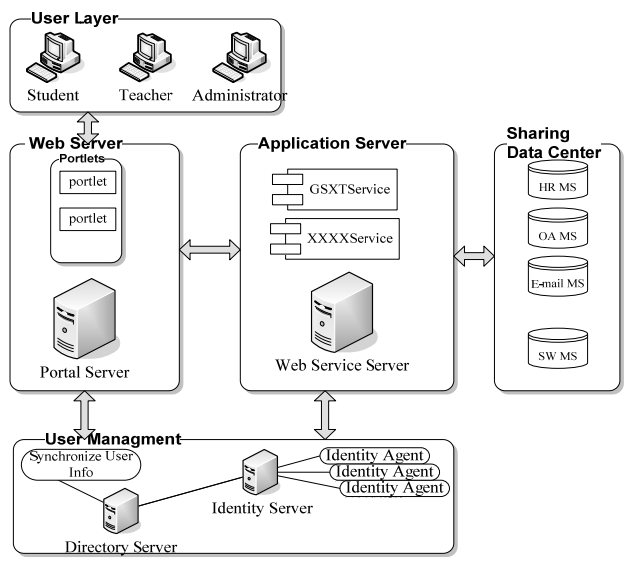


Figure 1 University portal system the basic hierarchical structure

Detailed information of each layer is as following:

1) User layer supports various terminals to access into UIP, such as computer, PDA, mobile phones and other devices' Web browser which can visit system resources and services by the unify login interface.

2) Web layer provides portal services, which can

* Supported by National Natural Science Fund (60573098) and the Key Project of Chinese Ministry of Education (107056).

process and display various users' data according to different portlet-application requests.

3) Application layer provides Web Service, which deals with business requests that received from Portal Server;

4) Shared database layer is the data standards of all the unified school system, which integrates shared data of all application information systems, provides consistent and accurate data sources and stores data in Application service layer.

We take an example to demonstrate the working procedures. A user submits a request to web layer through user layer. The web layer will select a particular portlet, which will send the request to the corresponding web service in application layer. And the web service will fetch the target data and send it back to the portlet in the web layer. And finally it will send the received data, which will be displayed in portal pages, to user layer.

2.2 Implementation Technique

The architecture adopts SUN Directory Server and SUN Identity Server platform, which provides a uniform identity authentication, for the entire portal platform. This certification interface as a middleware also offer an API for applications. Web console assesses Directory server, which is used for managing personnel data, via LDAP protocol. Web layer uses IBM Websphere Portal Server to provide portal services [2]. The application layer is provided by IBM Websphere, which also offers business supports to the entire framework.

3 Local web services resources intergration

This section starts from how Axis applies in Web Service developing, and then introduces the interaction between Portlet and Web service data resources [3]. And how Portlets and Web Services exchange data and how the platform integrates its data.

3.1 Appling Axis to develop Web Service

Axis framework comes from the Apache open-source organization; it is based on the latest SOAP standards (SOAP 1.2) of JAVA language and the open source codes implementation criterion by SOAP with Attachments [4]. There are many popular developing tools used to achieve AXIS as its function for Web services support, such as JBuilder and the famous Eclipse J2EE plug-in Lomboz. This article is based on the Eclipse J2EE plug-Lomboz and websphere application server.

This section shows how to use Axis in Web Service developing by the notice subsystem of the UIP. The main function of the subsystem is to facilitate users (students, teachers, administrative staff) setting the official information such as scholarships, grants or honorary title via the mode of SSO. Specific steps are as follows:

1) Developing client of Web Service [5]

Coding files of GsxtServiceImpl.java and GsxtService.java, of which the main function is to receive scholarships, and the honorary title of the XML information packets from the corresponding database;

2) Coding deploy.wsdd document

The main function of the file is to define the XML namespace, Axis provider; the parameter name defines class name of web service and the location of the package.

3) Deploying services

Preparing a batch file which named deploy.Bat, and Setting the path Axis lib, deploy.wsdd path, and so on;

4) Generating client stub documents

To visit server-side services from the browser, can be downloaded to GsxtService.wsdl file, an Axis tool and an org.apache.axis.wsdl.WSDL2Java class can be automatically generated from the WSDL document Web Service client code [6].

So far, as the above steps we can generate the information packets of XML Web Service of scholarship, honorary title in the local Websphere server.

3.2 Integration of Portlets and Web Service

Local applications are deployed into portlet Web Service which can conveniently and effectively interact

with shared database layer [7]. Specific local portlet architecture is shown in Figure 2.

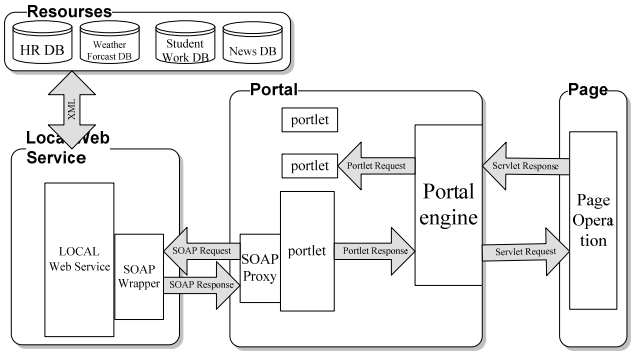


Figure 2 Architecture of portlet interact local Web Service

A user, like a student, loges into UIP, the local integrative components will response the servlet's request in the following procedures [8]:

1) If the request wraps with JSP or HTML Portlet, a response fragment will be sent back to portal, and then the portal will generate a HTML file.

2) If the request wraps with Web services, Portlet will first call the SOAP agent. The agent arranges the request parameter into SOAP requests, and then sends the request to the local Web services.

The local Web Service will unpack the received SOAP request, restore the requested parameters, call local Web services according to those parameters and complete the service requests. When results return, the SOAP wrapper will pack the results into SOAP requests which are programming language free and send them back to the SOAP agent, which will unpack the returned results. A portal with the relevant portlet fragment creates portal pages, sends the pages back to the user, and then generates the corresponding HTML.

We deploy the local portlets into the portal as the local integrate implementation. The procedures are as follows:

1) Creating source files, which include JSP files and the corresponding Java classes. The JSP files are the contents to be displayed in the Portlet; and the Java classes are the corresponding files relative with the contents.

2) Configuration Web.Xml. Configuration steps are the same as general Servlet configuration.

3) Configuration Portlet.Xml, which defines the rename of the Portlet, the relative servlet, and the portlet's attributes, such as display pattern and permission settings.

4) Packing for Web applications, which will pack all the files into a WAR file as a Web application in Websphere.

5) Setting portlet's parameters, which will set an address of the relative Web service.

6) Loading portlets and customizing relevant properties through portal management.

3.3 Data Integration

Data integration includes structured data and unstructured data. Structured data includes the data in the relational database; unstructured data includes text, pages, images and media [9]. The UIP is made of lots of shared data resources, such as personnel information management system, the student work information management system, and graduate student information management system, mail system and OA system. These resources include structured data and unstructured data, which are also called information isolated islands. In this part, UIP can be viewed as a VIEW layer, which manipulates Web services and XML-based data integration component for data integration. Both heterogeneous relational database and unstructured data are transformed into XML format data and sent to integration component through the unified interface. And XML format data is transferred to the UIP by the data integration component. The basic structure of data integration component shown in Figure3.

Data integration components are the center of data exchanging and data sharing, and each of them interacts through a standard Web service interface and transmits and exchanges data in XML formation. The data of shared database, in XML formation, is transferred to a Data integration component which will send the received data to UIP. Data integration component packs the data through Web service and takes the portal requests as certain service. The shared database exchanges and shares its data via SOAP's request and data integration component's response.

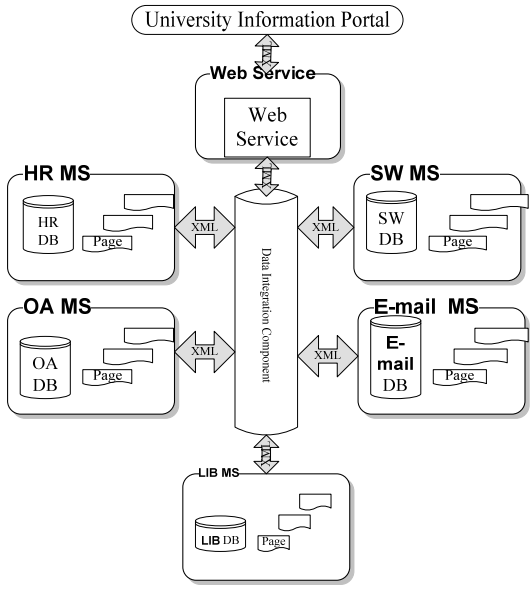


Figure 3 Basic structure of data integration component

4 Identity Authentication

This section starts with LDAP for the understanding of identity authentication directory service. And then introduces how SSO control user's login and visiting the limited resources request. Finally, introduces how UIP interact with the identity authentication server.

4.1 LDAP And Directory Server

The framework adopts the Sun ONE Identity Server as the public service. We take LDAP directory to write and read user's information, authorization management. Then the integrated portal system will be able to manage an identity authentication. LDAP is the Lightweight Directory Access Protocol, which is the cross-platform, standard protocols [10]. LDAP directory provides large-scale distributed environment for writing and reading user information data. And we make use of LDAP under two considerations: ① directory server in the data can be read on any platform, and very easy expansion in the user information ② Read operation takes most of the running time, which is more efficient than write operation, which is needed in registering users and modifying user's information. It is very suitable for the storage of user's information

4.2 SSO

The principle of SSO is that when users visit a number of authentication protected resource, IDS conversational server will generate an authentication credential (TokenID), which enables the user, who is in a conversation, to visit the resource without re-certification. For example, when a user visits a mail server, after certification, the mail server will call IDS's API, which will generate a SSO conversation token which records the identity of the user, while generating a random number linked to TokenID. The random number is sent to the user's browser as cookie preservation. When the user logs into the same domain of other servers, such as OA server, the previous cookie will be sent to the server, OA server can read the TokenID by SSO API from the current user's token. The user can log into the OA information system with verification.

The UIP make use of IDS authentication server certification to enter the right portal page, or right information resources. When a user logs into the UIP login interface, IDS will take the user's name and password to match uniform identity authentication service. If the user's name and password matches, the user is allowed to visit porlet's resource which is part of relevant role. Figure 4 shows the above procedures.

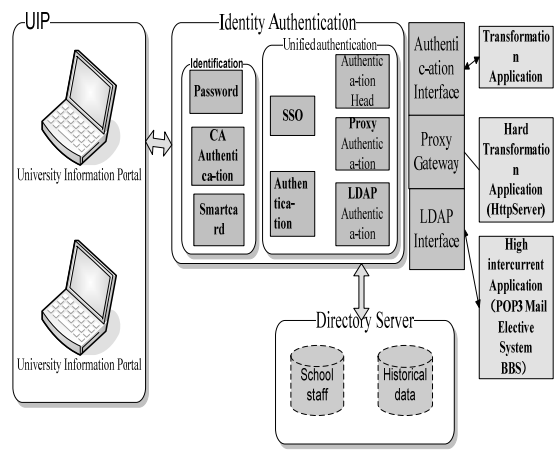


Figure 4 Access restricted resources portal framework

5 Conclusion

This paper presents a basic application framework, based on Portal, Web Service and identity authentication and describes the design and implementation of UIP. The webservice-based UIP can integrate resource and application with the technology of portlet, XML and identity authentication. It provides a unified SSO's login interface for students, teachers, administrators and other school users to visit the resources in public platform, which enhances the security, efficiency and extensibility of the UIP.

References

[1] Don Jones. Single Sign On Support in WebSphere Portal Server 1.2. http://www-900.ibm.com/developer

[2] WebSphere Portal. http://www.ibm.com/software/ websphere/portal (Accessed Jun. 2, 2004)

[3] R.Weinreich, T.Ziebermayr. Enhancing Presentation Level Integration of Remote Application and Services in Web Portals. *IEEE International Conference on Services Computing (SCC 2005)*, 2005. pp.224-236

[4] G. Alonso. Web services: Concepts, Architectures and Applications. Springer, 2004

[5] F. Bellas, "Standards for Second-Generation Portals," IEEE Internet Computing, vol. 8, 2004

[6] W3C Consortium.Web Services Description Language (WSDL). http://w3.org/2002/ws/desc/. 2004

[7] Java Community Process. JSR-168: Java Portlet Specification, version1.0 [EB/OL] http://www.jcp.Org/aboutJava/communityprocess/final/jsr 168/, 2003

[8] Wang Weiguo, LiFei, LuoZhe, Yan Baoping. Construction of Portal based Integrated Information Portal for Avian Influenza [J]. Application Research of Computer, 1001-3695(2007) 07-0279-04

[9] Shen Zhao-yang.Java, XML and Database applica- tion integration [M]. Beijing: Tinghua Press, 2002

[10] Yu Jian, ZhangHui, Zhao Hongmei.The Application of LDAP Directory Server in the Web Development [J]. Application of Computer, 2003, 23(10): 82-84

Research on Ontology Component and Description Logic Inference*

Wenjing Li [1] Yucheng Guo [2] Weizhi Liao [1] Rongwei Hang[1]

1 Department of Information Technology, Guangxi Teachers Education University, Nanning, 530001, China
Email: liwj@gxtc.edu.cn

2 Department of Computer Science, Wuhan University of Technology, Wuhan, 430063, China
Email: guoyucheng@whut.edu.cn

Abstract

The paper introduces the key technologies of the semantic web and the retrieval method of component description of the facet. According to the network architecture form of the facet component, a simple component knowledge library can be constructed. The logic axiom of component description deduced by basing on the OWL ontology component is consistent with the description logic axiom gained from the theory of component description logic and the actual inference.

Keywords: Semantic Web; Ontology Component; Description Logic; Description Logic Inference

1 Introduction

Along with the computer application domain's rapid expansion, the software scale and the complexity's unceasing enhancement, the software multiplexing already became an effective solution to avoid repetition work of the software development and to raise the software productivity. Through the Internet to gain software component is the efficient path to solve the software component shortage of resources. But, now the network engine's ambiguity and computer's semantic intelligibility do not cause on Web the component to be unable to use effectively. Semantic web put forwards that the definition and the linked data not only can be demonstrated but also can be automatic reeducated, integrated and entrusted with heavy responsibility by the computer. But gains each kind of the component through the next generation network Semantic web, its basic condition is the software component must have the good semantics. Therefore, we act according to the Semantic web the architecture, the domain ontology construction model, propose that based on XML and the RDF description logic's component ontology knowledge library, establishes the description logic reasoning rule and based on the OWL component ontology, complete through the description logic reasoning based on the semantic component retrieval.

2 Semantic Web's Key Technologies And Facet Component Description Method

Berners-Lee has given the Semantic web cascade strengthens the hierarchical structure for the first time in the XML2000 conference report [1]. These seven hierarchical structures are in turn:

First, UNICODE and URI, which is the lowest level of the entire semantic network, the Unicode processing resource's code, URI is responsible to mark the resources.

Second, XML+NS+xmlschema, uses in expressing the data content and the structure.

Third, RDF+rdf schema, uses in describing on the Web resources and the type.

Fourth, Ontology vocabulary provides the explicit

* Supported by Department of Education of Guangxi Zhuang Autonomous Region (0626120) and Guangxi Teachers Education University (0604A005).

formalized language, by accurate defined notion semantics and concept relations.

Fifth, Logic, realizes the inference with the logic description.

Sixth, Proof, gives the logical proof to the assertion of the existed information on the Web.

Seventh, Trust, guarantees the Web information credibly with the digital signature and the encryption technology.

Known from the seven hierarchical structures, the Semantic web's realization needs three key technologies XML, RDF, the Ontology's support, needs to describe logic to realize the inference, proves the credibility of Web information through the digital signature and the encryption technology.

2.1 XML technology

XML is the W3C recommendation standard, is an emerging technology on the Internet, and obtained widespread application. Compared with HTML language, it overcomes many limitations of the hypertext language. It has characters such as the mark extendibility, the separation of the content and the style, the formidable ultra link function, convenience for transmission of information among different systems. These formidable functions are inseparable with XML documents definition standard DTD and XML Schema, realization of the cross platform to visit XML the data documents object model DOM, utilization in demonstrating that the XML documents content XML correlation techniques and so on cascading style sheet CSS or XSL. Web pages have high flexibility based on XML technical, is a new technology to construct the establishment condition website and the production dynamic Web page; XML is also one semantic, structured, half structured language, it describes the documents structure and the semantics, users may define the special-purpose mark of themselves' domain very conveniently, the documents structure may also be random. When the XML documents are used in the different applications, it is only need to change the corresponding manifestation, not to revise the documents itself. Therefore, XML is a very ideal cross platform language used in the stored datum content and the structure [2].

2.2 RDF(S)

The resource description framework (RDF) is the W3C organization's recommendation of the language standard, which uses to descript the relation between resources, has characteristics such as simple, easy to expand, openness, easy to exchange and easy to synthesize. The RDF's goal is to provide one kind of general frame for Web resources description, RDF is a kind of metadata model, while XML is a kind of grammar form[3]. The RDF data model may use XML to indicate that may also use other grammar format descriptions. RDF Schema is a kind of realization to RDF based on XML, is to provide a basic system for the RDF model and define resources attributes. The definition the kind which describes, and combines to kind and the relational possibility carries on the restraint, simultaneously provides the restraint case of breaching the rules the examination mechanism [4].

2.3 Ontology

The Ontology is the formalized explanation of the shared conceptual model, describes the concept semantics through the relations among concepts. As a kind of effective model to display the concept's hierarchical structure and the semantic, Ontology's goal is to capture knowledge of related domain, provide common understanding to the knowledge of this domain, determines glossary which approves through this domain, and give clear definition of the relations among these glossaries[5].

2.4 Description methods of facet component

Description plan based on facet mainly composes three parts: Facet classification plan, component's facets description congregation as well as relations among each facet description terminology. For example: Prieto-Diaz raises most early the facet description plan

is two main facets: “function” and “environment”, and each main facet has 3 sub-facets separately (action, object, medium) and (application domain, system type, customer type), describes some component under this facet description plan to be possible to use the facet tree which as shown in Figure 1 to express [5] [6]. In Figure 1 "App" expresses application domain, "Sys "expresses system type, "Cus" expresses customer type.

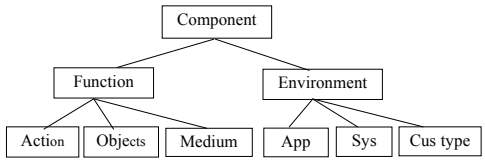


Figure 1 A facet tree of a component

3 Description Logic Component Knowledge Library And based on Owl Ontology Component

3.1 Description logic component knowledge library

In Figure 1 component facet tree, we take the chatting server component of the network communication domain as an examples, establishes its facet tree [3], As shown in Figure 2.

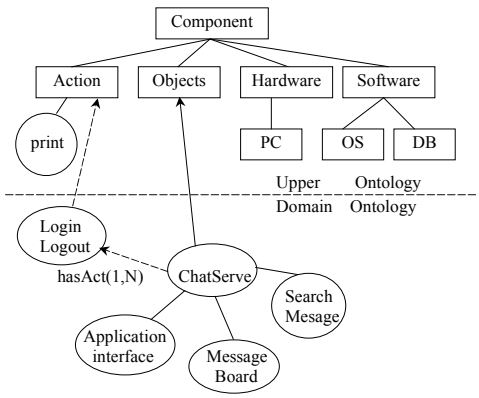


Figure 2 A ontology tree of ChatServer Component

Figure 2 the ellipse and pane node expresses concept, the character in the node is the concept name. Figure 2 the expresses knowledge library including Component, Action, Objects, Hardware, Software, Chatserver Objects, Login, Logout, Search Message, Message Board, OS, DB,PC concepts and so on, in the chart between the solid arrow expresses concept “Is-a” relation, for example between ChatServer and Component “Is-a” relation expresses “ChatServer is a Component Object”. The dashed line expresses dual relations between two concept examples, the starting point of the dashed line is the definition domain of the relation, and the termination node of the dashed line is the value domain of the relation. Inside the node of the ellipse expresses the example of the concept Action, writing nearby the ellipse expresses the relation name and the relation restraint, concept Action and concept ChatServer Object has the named “hasAct” relation, (1, N) is the relation has Action’s restraint on the value domain Action: 1 expresses the Action number's world of mortals, N expresses the Action number is the limited upper boundary. The facets also have the dependence relationship, this can be defined as “FacetDependenceProperty”, the concrete dependence name has for example: “Act On”. May used to express the dependence relationship between concept Action and concept Objects, the expression is “Action Act on Objects” [7].

Figure 2 contains the component domain knowledge library can be expressed by description logics component knowledge library shown in Figure 3.

$$Action \cup Objects \subseteq Component$$
$$Action \cup Objects \cup Hardware \cup Software \subseteq Component$$
$$Action \subseteq Component$$
$$Objects \subseteq Component$$
$$Hardware \subseteq Component$$
$$Software \subseteq Component$$
$$OS \subseteq Component$$
$$DB \subseteq Component$$
$$OS \cup DB \subseteq Software$$
$$ChatServer = Component \cap Objects$$
$$Login \subseteq Action$$
$$Logout \subseteq Action$$
$$ChatServer \subseteq \exists\, hasAct.Action$$
$$ChatServer \subseteq \forall\, hasAc.Action$$

Figure 3 Description logic expression component knowledge library

In the description logic component knowledge

library, Component, Action, Objects, System type, Client type, Chatserver Objects, Login, Logout, Message send, OS, DB and so on are regarded as concepts, "Is-a" is regarded as relationship. The knowledge library uses the description logic grammatical form $A \cup B$ to express concept mergence between A and B, to express the concept A implication with B by axiom $A \subseteq B$. Each axiom in Figure 2 may be transformed to the corresponding first-order predicate well-formed formula: For example, the axiom ChatServer hasAction.Action can be transformed into $(\forall x)$ ChatServer (x) has Action(x, y) Action(y). In the transformation, all concepts will be regarded as unitary predicate; relations will be regarded as dual predicate [5].

3.2 Component description based on OWL

Concept is composed of class and instance and property in OWL. Leafage facet was defined class in OWL [6], abstract language and embody language were defined class and instance in domain language. Class languages of the same leafage facet were defined that make up of a series of inherit relation. Leafage facet class corresponding is their same father class. Language of instance was defined that is instance of upper language class. For example, component is class, Action, Object, Hardware, Software is subclass of Component, ChatServer is subclass of Object. Print and Application interface are instances of Action and ChatServer. Besides inherit relation ontology can define relation else. Between Action class and Object class is "Act on" relation. This is a reliant relation of facet. Action class and ChatServer Object class have the named "hasAct" relation,(1, N) is the relation has Action's restriction on the value domain Action. In Figure 2 dashed upward part is upper ontology and is domain ontology under layer. There is parts materialize axiom and other parts materialize instance knowledge in Ontology component knowledge. Subclass relation of the same father is axiom too. Besides can define else axiom. Instance, Relation of class defined transitive relation. Figure 4 shows that OWL description snippet of under layer domain ontology [5] [8].

4 Description Logic Principle And Inference Rule

Defines 1 description logical grammar

Suppose Nc and N_R are countable and do not intersect atomic concept collection and atomic relation collection. The description logic's concept description regressive definition is:

(1) Random atomic concept $A \in N_c$ is description logical concept;

(2) Suppose A and B is description logical concept, R is atomic relation of description logical concept, then the $\neg$ A, $A \cup B$ (and), $A \cap B$(junction), $\exists$ R.A (existence restraint) and $\forall$ R.A(full title restraint) are description logical concepts.

Defines 2 description logic semantics

The description logic explanation is one dual to $I = (\Delta^I, \bullet^I)$, Δ^I is represents the universe of discourse the non-spatial set, $\bullet^I$ is explains the function, it each $A \in N_c$ mapping for Δ^I subset, each $R \in N_R$ mapping for $\Delta^I \times \Delta^I$ the subset, It is called the atomic concept and the atomic relation R explanation separately, records makes A^I and R^I.

One description logics knowledge library is mainly composed of TBox and the ABox two parts. TBox defines the component area of knowledge structure and contains a series of axioms, May through the existed concept to constitute the new concept. ABox states that the domain individual and the concept as well as the human body with the relation's subordination relations, is a group of limited assertion set, it contains concept examples in TBox. The description logic component knowledge library of Figure 3 defines the component facet description plan knowledge structure and a series of axioms, through the existed concept to constitute the new concept, constitutes new concept ChatServer Component through atomic concept Component and Objects. ABox contains concept examples in TBox, for example, concept Action's example Login, Logout and so on.

We will OWL DL translate into abstract syntax,

translate into description logic semantic any more, to construct description logic knowledge library. Suppose named space of ontology is ex in Figure 4,it's abstract syntax snippet for next[9]:

```
<owl:Class rdf:ID="Component"/>
<owl:Class rdf:ID="Hardware">
<rdfs:subClassOf rdf:resource="#Componet"/>
</owl:Class>
<owl:Class rdf:ID="Software">
<rdfs:subClassOf rdf:resource="#Componet"/>
</owl:Class>
<owl:Class rdf:ID="Action">
<rdfs:subClassOf rdf:resource="#Componet"/>
</owl:Class>
<owl:Class rdf:ID="Object">
<rdfs:subClassOf rdf:resource="#Componet"/>
</owl:Class>
<owl:Class rdf:ID="ChatServer Component">
<owl:unionOf rdf:parseType="Union">
    <owl:Class rdf:about="#Action">
    <owl:Class rdf:about="#Objct">
    <owl:Class rdf:about="Hardware">
    <owl:Class rdf:about="Software">
  </owl:UnionOf>
</Owl:Class>
<owl:Class rdf:ID="ChatServer">
<rdfs:subClassOf rdf:resource="#Componet"/>
</owl:Class>
<owl:Class rdf:about="#ChatServer">
<rdfs:subClassOf rdf:resource="#Objcet"/>
</owl:Class>
<owl:Class rdf:ID="Login">
<rdfs:subClassOf rdf:resource="#Action"/>
</owl:Class>
<owl:Class rdf:ID="ChatServer">
<owl:intersectionOf rdf:parseType="Collection ">
<owl:Class rdf:about="Componect">
<owl:Class rdf:about="Object">
</owl:intersectionOF>
</owl:Class>
<owl:ObjectProperty rdf:ID="hasAct"/>
<rdfs:domain rdf:resource="#Object"/>
<rdfs:range rdf:resource="#Action"/>
</owl: ObjectProperty>
<ChatServer rdf:ID="Application interface"/>
<Action rdf:ID="print"/>
<owl:Restriction>
<owl:onProperty rdf:resource="#hasAct"/>
<owl:allValuesFromrdf:resource="#Action"/>
<owl:hasValue rdf:resource="#Login"/>
</owl:Restriction>
```

Figure 4 OWL fragment of the Domain ontology

SubClassOf (ex:Action ex:Component)
SubClassOf (ex:Object ex:Component)
SubClassOf (ex:Hardware ex:Component)
SubClassOf (ex:Software ex:Component)
subClassOf (ex:ChatServer ex:Component)
subClassOf (ex:ChatServer ex:Object)
Class (ChatServer complete intersectionOf(ex: Component ex:Object))
ObjectProperty(ex:Object)
domain(ex:Action)
range(ex:Object)
restriction (ex:hasAct value (ex:Login))

We reason some axiom of simple description logic knowledge library from abstract syntax snippet so you say. Figure 5 show, these axiom are consistent from axiom of component description logics knowledge. Description logic reasoning machine can deduce relation be contained between two concepts. We adopt Protégé system to construct component domain knowledge ontology，create OWL document of component domain ontology. Annotea was adopted for Semantic remark tool to call in component domain knowledge ontology, after remark to create embodies component and requesting OWL document. Then carry out description logic consequence with RACER system [9] [10].

$$Action \subseteq Component$$
$$Objects \subseteq Component$$
$$Hardware \subseteq Component$$
$$Software \subseteq Component$$
$$ChatServer\ Component = Action \cup Objects \cup Hardware \cup Software$$
$$ChatServer = Component \cap Objects$$
$$\exists \text{ has Act.Login}$$

Figure 5 The several axiom of ontology component knowledge library

5 Conclusion

We give an example for chat server component, According to the facet component description network architecture form, to construct a simple component knowledge library. Then to detrude description logic axiom based on the OWL ontology component,

Annotea was adopted for Semantic remark tool to call in component domain knowledge ontology, carry out description logic consequence with RACER system. And gained the description logic axiom is consistent from theory of component description logics and fact inference.

References

[1] T Berners–Lee, J Hendler, O Lassila, "The semantic Web", Scientific American , 284(5) ,2001 ,pp.34～ 43

[2] M Klettke, H Meyer, "XML and object relational database systems—Enhancing structural mappings based on statistics", In:Proc of the 3rd Int' l Work shop on the Web and Databases (WebDB 2000), Dallas, Texas, 2000,pp.63～68

[3] Staab S, Erdmann M, Maedche A, Decker S, "An extensible approach for modeling ontology in RDF(S)" ,In: Grütter R, ed. Proc. Of the 1st Workshop on the Semantic Web at the 4th European Conf. on Digital Libraries. Hershey: IGI Publishing, 2000, pp.18～20

[4] Patel-Schneider P, Simeon J, "The yin/yang Web: A unified model for XML syntax and RDF semantics", IEEE Trans. on Knowledge and Data Engineering, 15(4),2003, pp. 797−812

[5] Wang Yuanfeng ,Zhang Yong, Ren Hongmin, et al, "Retrieving Components Based on Faceted Classification", Journal of Software, 13(8),2002,pp.1546~1551

[6] PrietoDiaz.R, "A faceted approach to building ontology", Proceedings of IEEE International Conference on Information Reuse and Integration (IRI 2003), 2003, pp.458~465

[7] Peng Xin, Zhao Wen-yun, Xiao Jun, "Representing and Retrieving Components Based on Ontology", Journal of Nanjing University (Natural Sciences), Vol.41, Oct, 2005, pp.470~475

[8] JiaXiao Hui ,Chen De Hua, et al, "Research on Matching Model and Algorithm for Faceted Based Software Component Query", Journal of Computer Research and Development, Vol141 ,No110 Oct, 2004, pp. 1635~1638

[9] Wang Zongwei, Zhu Guojin, Zhao Langbo, Su Xiang, "Web Problem Resources Retrieval Based on Ontology and Description Logic Inferences", Computer Engineering, *Vol.32* № *18,2006,pp.225*~227

[10] Xiao Hang Wang, Da Qing Zhang, Tao Gu, et al, "Ontology based context modeling and reasoning using OWL", Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops (PERCOMW 2004), 2004, pp.18~22

Analytical Model of Enterprise Resource Planning Platform Based on J2EE *

Yixiang Ding Minghua Jiang Ming Hu

College of Computer Science, Wuhan University of Science and Engineering, Wuhan, Hubei 430073, China
Email: dyx@wuse.edu.cn

Abstract

An enterprise resource planning (ERP) system is defined as a unified information system, performing all information processing tasks of a company and realizing an integration of the whole corporation. This paper describes the architecture of the ERP platform, which is composed of a front-end web server, an application server and a backend database server, and then proposes its analytical model by using queueing network theory. A test environment is built to measure model parameters based on enterprise standard server components and Load runner tool. The measurement results show the model predicts performance measures such as response time and throughput accurately.

Keywords: Enterprise Resource Planning, Queueing network mode1, Mean-Value Analysis

1 Introduction

The enterprise resource planning (ERP) is a core group of applications that can be extended with additional applications such as customer relationship management, business intelligence, supply chain management, or other e-business applications. The result is an integrated set of packaged offerings–referred to as enterprise application solutions that can integrate an entire government entity, including such capabilities as accounting and financials, materials and capacity management, human resources, payroll and other government-specific functions. The enterprise applications are the main direction of the software industry development at all times, and hold the very important status. At the same time, they are influenced by the tidal current of the IT development. Along with the development of global economic integration, the competition between enterprises is no longer just between the individuals but the supply chains. The ERP platform is a supporting platform which can strengthens the enterprise's market reaction rate and competitive power. A good ERP platform must be able to adjust the system structure immediately according to the business events changes, in order to adapt to the new operational mechanism of the enterprises, so the Web-based collaborative ERP platform is just developed to resolve this problem. Using this system, an ERP platform which can adapt to different business events can be built just by a series of customizing and releasing processes. In the situation of the need to change the operation flow or the application module, we can adapt the system to the new operation flow only by newly customizing and modifying, but not need to do the re-development on the tier of program design of the software.

One of the most pressing problems faced by ERP platform designers is the adequate sizing of their infrastructure so that they can provide the quality-of-service required by their clients. A validated model of ERP architecture based on web services is the basis of capacity performance in different settings. In addition to capacity planning, a simple yet accurate model of Web services is also instrumental for other purposes such as overload control or resource provisioning. Overload control and performance management are research areas on their own, but their

* This work is Supported by The Young Scholars Research Fund of Wuhan University of Science &Engineering (No: 20073224) and Hubei Provincial Natural Science Foundation of China (No: 2007ABA376).

effectiveness are based on the accuracy of good performance models. Most of the existing work of ERP platform performance modeling is confined to the front-end Web server. Their model is divided into three layers, where each layer models a certain aspect of the system. The model has several parameters, some of which are known and the remaining unknown parameters are determined by simulations.

This paper describes the implementation of an ERP platform, and then analyzes its performance by using queueing model. Performance measurements indicate that the introduced model provides higher accuracy yet still simple. A simple model is important in performance modeling because it renders a smaller parameter space thus easier to estimate; while a complicated model usually contains parameters that are difficult to obtain and validate. Hence this performance model tends to be more useful in real-world applications.

2 Architecture of ERP Platform

The web-based ERP platform is developed in the J2EE environment. In order to decrease development time, we propose a platform that provide the support for transaction, container, Exception Handling by tool classes and provide some basic function such as add, delete, modify, query, page etc. Figure 1 is this platform architecture.

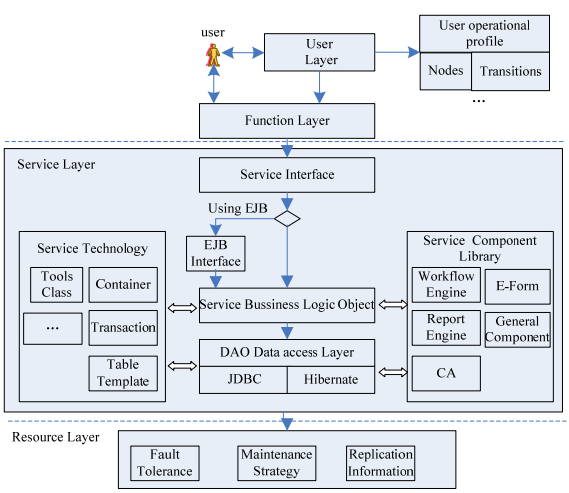


Figure 1 Architecture of ERP Platform

Database Access Layer: Data persistent can be realized by the data access object, the main job of the data access object is adding, modifying and querying the basic data, .etc. but, it is not including the complex business. At the same time, the persistent technology of this layer is varied, and these optional technologies are mature, so, we focus on how to make the different optional technologies compatible. The effect of the change of the technologies can be avoided by separating the interface and the realization.

Typically, we can use the JDBC, Hibernate to realize the data access object and form the data access layer. With the development of the related technologies, we can apply some other new technologies; even choose the technologies according to the request of the users in specially situation. The key of the data access layer is to construct the reusable object layer with technologies of independent data access layer. The technologies in the data access layer are various and the Hibernate and SQLHelper is main platform at present. Because the focus is not the complicated business, the selected technologies would less change.

Presentation layer: the presentation layer compose of JSP pages, pages components (DataGrid, menu, ComboBox, Textbox, date .etc.) and Taglib (Struts tags and component tags), which used to be operated by users and display the results. Powerful and rich components will lead to an excellent result of the page displaying and the user's operation.

Business logic layer: the main task of the business logic layer is to realize the complex business. From the database perspective, these operations usually involve a number of database tables and complex nested transaction.

From the object-oriented perspective, these business objects relies on a number of related business objects or data access object and involves the map among several tables. When realizing the business logic, we use the technology of separating the interface and realization. The interface represent the contracts of business functions for others, which designed by the designer, the coder of the client just to care about these interfaces which can be used to complete the corresponding business functions. The Service Factory

is used to established and visit the object, which constitutes the interface of the business logic for others.

Presentation layer: the presentation layer compose of JSP pages, pages components (DataGrid, menu, ComboBox, Textbox, date .etc.) and Taglib (Struts tags and component tags), which used to be operated by users and display the results. Powerful and rich components will lead to an excellent result of the page displaying and the user's operation.

3 Queueing Nework Model of ERP Platform

3.1 The description for queueing network model

Queueing theory is one of the key analytical modeling techniques used for computer system performance analysis. Figure 2 is the queueing model of this platform. Tier 0 represents user think times and is an infinite server node which requests issued by a session emanates. There are M tiers in this platform: Tier1, Tier2, Tier3, … , Tier M. After leaving Tier I, a request either returns to tier i-1 with probability pi, or proceeds to tier i+1 with probability (1- pi). Note in Tier M, PM=1.

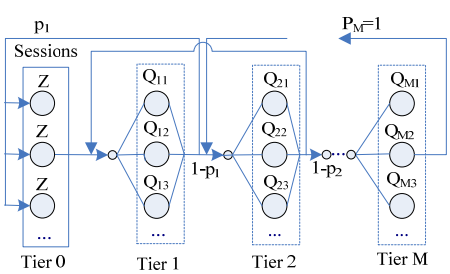


Figure 2 Queueing Model of ERP Platform

We used the Mean Value Analysis (MVA) algorithm to compute performance quantities such as mean waiting time, throughput, and the mean number of jobs at each node. The MVA algorithm form closed queueing networks was developed by [6] and can compute performance measures without computing the normalization constant. Suppose that there are N requests in the network. Let the arrival rate at node i be λi. The visit ratio is the mean fraction of visits (νi) of a request to node I relative to a reference node with $V_i = \lambda_i / \lambda$. λ_r is the total number of the requests serviced by the entire application over a duration t. To find the visit ratios for nodes the traffic equations $V_i = \sum_{j=1}^{M} V_j p_{ji}$ are solved where pji are the routing probabilities that a job goes to node i after leaving node j. Let Di be the mean service demand of each request at server i (i=1,…, K), Ri the mean response time, Q_i the mean number of requests at server i. The MVA algorithm involves iteratively solving three equations:

1. $R_i(N)=D_i(1+Qi(N-1))$.
2. $\tau = \sum_{i=1}^{N} R_i(N)$.
3. $Q_i(N) = \tau * R_i(N)$.

This equation is based on the result that the probability distribution of the number of jobs seen by an arrival to a node with N jobs in the network is the same as that when there is one less job (N-1) in the network.

3.2 Approximate solution

In order to compute the response time, the model requires several parameters as inputs: The number of application tiers M, the number of the sessions (simultaneous user connections), the average user think time $\overline{Z_c}$, the average service time $\overline{S_i}$ at each tier and the visit ratio V_i. $\overline{S_i}$ is the mean service time of a request serviced at each tier.

In this application, the Web, Java, and database servers all support extensive logging facilities and can log a variety of useful information about each serviced request. In particular, these components can log the residence time of individual requests as observed at that tier—the residence time includes the time spent by the request at this tier and all subsequent tiers that processed this request. This logging facility can be used to estimate per-tier service times. From this, we can estimate the mean service time $\overline{S_i}$ and the visit ratio V_i.

Based on the closed queuing model and the estimation of the above parameters, we can iteratively

computes the average response time of a request using the MVA (Mean-Value Analysis) algorithm. First we will define the follow notations: $\overline{R}_i$ denote the average per-request delay at *Qi*; $\overline{R}$ denote the average response time; $\overline{L}_i$ denote the average length of Q_i; $\overline{D}_i$ denote the average per-request service demand at *Qi*; τ denote the throughput.

Initialization: $\overline{R}_0 = \overline{D}_0 = \overline{Z}$; $\overline{L}_0 = 0$

Compute the service demand:

for i=1 to M do

$\overline{L}_i$ =0;

$\overline{D}_i = V_i \cdot \overline{S}_i$; /**service demand**/

end

Introduce N customers, one by one:

for n=1 to N do

for i=1 to M do

$\overline{R}_i = \overline{D}_i \cdot (1 + \overline{L}_i)$; /**average delay**/

end

$\tau = (\frac{n}{\overline{R}_0 + \sum_{i=1}^{M} \overline{R}_i})$; /**throughput**/

for i=1 to M do

$\overline{L}_i = \tau \cdot \overline{R}_i$; /**Litter's Law**/

end

$\overline{L}_0 = \tau \cdot \overline{R}_0$;

end

Compute the average response time:

$\overline{R} = \sum_{i=1}^{M} \overline{R}_i$; /**response time**/

The results presented in Figure 3 correspond to the common shape of response time performance metrics with the help of MATLAB.

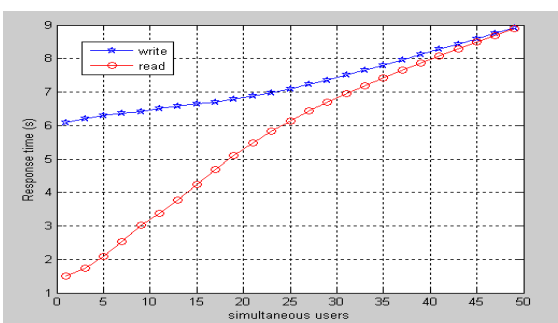


Figure 3 the Predicted Response Times with MVA

4 Numerical and Simulation Results

Table 1 is test environment used in this paper.

Table 1 Test Environment

Server CPU	Intel Pentium 4 3.0G
Server OS	Windows Server 2003 with Service Pack 3
DBMS	SQL Server 2000 Enterprise with Service Pack 4
Client OS	Windows XP Professional with Service Pack 2
Client CPU	1.6 GHz Inter(R) Core(TM) 2 Duo T5470
LAN	100 Mb/s

Load runner of Mercury Interactive is using for test tool. Mercury Load Runner enables you to test your system under controlled and peak load conditions in order to isolate and identify potential client, network, and server bottlenecks. To generate load, Load Runner runs thousands of Virtual Users that are distributed over a network. Using a minimum of hardware resources, these Virtual Users provide consistent, repeatable, and measurable load to exercise your application just as real users would. Load Runner's in-depth reports and graphs provide the information that you need to evaluate the performance of your application.

In this test, the number of the application tiers is three. We record two classes of request, a database reader and a database writer. In the user scenarios, the number of simultaneous browser connections is from 0 to 50, and we introduce the users into the application by 2 in every 2 seconds. In order to simulate the realistic usage of the application, we include the user think time as a random number between 2 and 8 seconds. After running the test, we get the following results: Figure 4 shows the transaction response time under load which is growing along with the increasing of the user numbers. Figure 5 shows the percent of the transaction response time. From the two figures, we can see that all the average response time maintains in the user's acceptable scope. There into, the response time include the user think time which is a random number, the two curves represent the write transaction and the read transaction.

Figure 4 Average Response Time and Number of Users

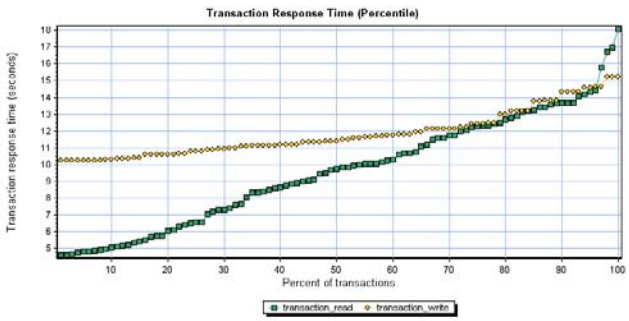


Figure 5 Transaction Response Time and Percent of Transaction

5 Conclusions

In this paper we described the platform architecture of enterprise resource plan platform in detail and presented an analytical model for enterprise resource plan platform. Our model is based on using a network of queues to represent how the tiers in a multi-tier application cooperate to process requests. A test bed is built base on widely deployed combination of Apache, Tomcat and SQL Server 200, with test tool Load runner. Validation tests show the model can predict performance measures such as response time and throughput accurately. The enhancement of the model and the validation of the enhanced model is a subject of future work.

References

[1] Zhang Zhi-Qing, Qin Ling, Yan, Hong, Zhao Tong-Ying, Wireless Communications, Networking and Mobile Computing, 2007, September 2007, pp5333~5336

[2] S. Croom, P. Romano, M. Cannakis, "Supply Chain Management: An Analytical Framework for Critical Literature Review", European Journal of Purchasing & Supply Management, Vol.6, 2000, pp.67-83

[3] G. Caprihan, Managing Software Performance in the Globally Distributed Software Development Paradigm, Global Software Engineering, October 2006, pp 83~91

[4] D. L. Eager and K. C. Sevcik, Analysis of an approximation algorithm for queueing networks, Performance Evaluation, April 1984, pp 275~284

[5] Masud A. Khandker, Performance Measurement and Analytic Modeling Techniques for Client-Server Distributed Systems, Ph.D. Thesis, University of Michigan, 1997

[6] Reiser, M., Lavenberg, S.S., "Mean Value Analysis of Closed Multichain Queueing Networks", J. of the ACM, Vol. 27, No. 2, Apr.1980, pp. 313~322

[7] F.Baccelli and P.Bremaud, Elements of Queueing Theory, Spring-Verlag, 1994

[8] R.Jain, The Art of Computer Systems Performance Analysis, John Wiley and Sons, 1991

[9] A.Willing, Performance Evaluation Techniques, Lecture Notes, Potsdam, 2004

[10] James Holmes, Struts: The Complete Reference [M], The Mc-Graw-Hill Companies, 2004

Failure Prediction Based EX_QoS Driven Adaptive Approach for Distributed Service Composition *

Yu Dai Lei Yang Bin Zhang Kening Gao

School of Information Science and Technology, Northeastern University, Shenyang, Liaoning 110004, P.R.China
Email:NEU_DaiYu@126.com

Abstract

As web services may dynamically change, composite service should repair itself if any execution problems occur, in order to successfully complete its execution. This paper proposes an adaptive approach for distributed service composition. Compared with works have been done, this paper emphasizes on finding the replacement composite service with global optimization in decentralized service composition and thus an EX_QoS model is proposed. Besides this, through failure prediction, the re-selection process will be started earlier before the invocation of failed service. This will make the re-selection process minimize the interrupting time (caused by online re-selection) of the composite service execution and improve the availability of the composite service. The experimentations show better performance of our algorithm.

Keywords: EX_QoS, Service, Composite Service, Distributed Service Composition, Failure Prediction

1 Introduction

Since web services operate autonomously within a highly variable environment, as a result of which their QoS may evolve relatively frequently, either because of internal changes or changes in their environment [1], such dynamic property of web services requires that composite service must adjust itself during runtime.

Although several works focusing on how to establish the QoS model in order to evaluate quality of services and how to do the selection in order to achieve end-to-end quality constraint have been studied, they lack of the ability to adapt to changes especially occurred during the execution of composite service. A native approach to handle the problem is to re-compose from scratch every time a change occurs. However, this may not be a feasible solution due to that the time complexity of the re-composition is high which will interrupt execution of the composite service.

Besides this, current approach for finding replacement composite service with global optimization is only suitable for centralized service composition. Currently, there exist two models for invoking composite service: centralized model [1, 2] and decentralized model [3]. For such two models, QoS for evaluating the quality of services must be computed differently. In centralized model, there exists a centralized engine. According to the routing information of composite service, such engine will route data among component services. In this model, QoS information about data transmission time for the service selection will be the time of routing data between the component service and the centralized engine. The selection of one service will not influence the selection of others. In decentralized model, on each component service, there exists a routing table generated from the routing information of the composite service. Such routing table is used for routing data among component services. In this model, QoS information about data transmission time for the service selection will be the time of routing data between the component services. That is to say, the selection of one service will influence the selection of

* This work is supported by national natural science foundation of China (No. 60773218) and natural science foundation of Liaoning Province (No. 20072031).

others. Thus, how to find replacement composite service with global optimization in distributed service composition is still another problem needed to solve.

With above problems in mind, we present a solution for adapting in distributed service composition. Firstly, an EX_QoS model for evaluating quality performance of distributed service composition is proposed. Secondly, a failure prediction approach is proposed. Through doing so, the re-selection process, which is done in simultaneously with the execution process of composite service, will make the re-selection process minimize the interrupting time of composite service execution and also improve the availability of the replacement composite service. Thirdly, a re-selected slice of composite service is generated based on proposed slice determining rules and quality affecting degree. The experimentations show better performance of our approach.

2 Related Works

In order to make the composite service recover from the failure with minimal user intervention and make the recovered composite service meet the end-to-end constraint, researchers proposes the QoS-driven adaptation approach for composite service. Based on the replacement composite service idea, researchers [4, 5] propose approaches of backing up a composite service for each component service. Then, when a component service is incurred a failure, the composite service can be easily switch to a replacement one and such self-healing process will not affect the execution performance of the composite service. In Ref. [4, 5], all the replacement composite services are backed up before the execution of the composite service. Such two approaches do not consider the QoS of services during runtime of the composite service. Thus, the replacement one will not be available sometimes. In this paper, we share their idea of backing up composite service and extend their approaches to find replacement one as soon as possible when the service is failed and replacement composite services are not available. Besides this, the above two works mainly focus on adapting approach for centralized service composition. This paper attempts to research on adaptive approach for decentralized service composition through extending current QoS by dependent quality.

One of the researching works that do the re-selection process during the composite service execution is the approach in Ref. [6]. In Ref. [6], the re-selection process will be triggered as soon as the actual QoS deviates from the initial estimates. When the failure is found, the execution of composite service will be stopped until the re-selection process is finished. Thus, this approach can be only used for runtime-unaware application. Compared with Ref. [6], we improve its idea through introducing a failure prediction to predict the availability of services, triggering re-selection process as soon as the services are not available and make the re-selection process done in simultaneously with the execution of composite service. Thus, our approach can be used in runtime-aware application.

3 Ex-QoS Model for Decentralized Service Composition

3.1 Preliminaries

In this section, we introduce some basic concepts that will be used in the remainder of the paper.

Definition 1. QoS of Atomic Service. For an atomic service s (which only contains one operation), the QoS of s can be defined as: QoS(s)=<$Q^t(s)$, $Q^p(s)$>, where:

1) Qt(s) is the response time of s which is the interval of time elapsed from the invocation to the completion of service s. Qt(s)=R/Vdepose+R/ Vtransmission, where R is the amount of request, Vdepose is the speed of deposing the request by s and Vtransmission is the transmission speed of sending request to the service and returning result to the executing engine. As mentioned in Ref. [1],Vdepose is the average speed of deposing the request and Vtransmission is the average transmission speed.

2) Qp(s) is the cost of invoking s.

The works of Zeng [1] propose a mathematical model for workflow QoS computation, described by some aggregation functions for sequence, choice, parallel and iteration structure of the workflow. In this paper, we used the aggregation functions as Ref. [1]. In order to adapt to their dynamic properties, QoS driven selection for achieving end-to-end constraint is illustrated in Ref. [7]. The aim of such selection is to maximize the fitness function of the available QoS factors; and meet the constraint specified for some of the factors. Such problem can be formulated as Eq. (1).

$$\max_{v} \; F\left(CS_v\right) \quad s.t \;\; Q^t\left(CS_v\right) \le Q_c^t \tag{1}$$

Where, $CS_v=\{(sc_1,\ s_1,\ k_1),\ldots,\ (sc_i,\ s_i,\ k_i),\ (sc_n,\ s_n,k_n)\}$, here, sc_v is the service class in abstract composite service and $s_{v,kv}$ is the service selected for service class sc_v; F is the fitness function which can be computed as Eq. (2).

$$F\left(CS_v\right)=w_t*\left(\frac{Q^t\left(CS_v\right)-u_t}{\sigma_t}\right)+w_p*\left(\frac{Q^p\left(CS_v\right)-u_p}{\sigma_p}\right) \tag{2}$$

Where wt and wp are the weights ($0<=w_t,\ w_p<=1$, $w_t+w_p=1$). σ and μ are the standard deviation and average of the QoS values for all potential composite services.

3.2 Dependent quality extended QoS model (Ex_QoS)

In decentralized service composition, it is needed to consider the dependent quality between services since such relation can affect quality of composite service.

Definition 2. Dependent Quality. For composite service *CS*, services s_1 and s_2 are two component services in *CS*. The dependent quality of s_1 towards s_2 in *CS* can be signified as $DQ(s_1,\ s_2)$, which can be computed as Eq. (3).

$$DQ\left(s_1,s_2\right)=\frac{a_{12}}{b_{12}} \tag{3}$$

Where, a_{12} is the amount of data transmitted between services s_1 and s_2; b_{12} is the bandwidth between services s_1 and s_2. We also use v_{12} to signify the data transmission speed between s_1 and s_2, it can be computed as $v_{12}=1/b_{12}$.

Then, for composite service *CS*, the dependent quality of it can be computed as Eq. (4). It illustrates that dependent quality of composite service is a sum of transmission time.

$$DQ\left(CS\right)=\frac{\sum_i\sum_j DQ\left(s_i,s_j\right)}{2} \tag{4}$$

Definition 3 Dependent Quality Extended QoS Model (EX_QoS). For composite service *CS* and a component service s, EX_QoS of s can be defined as: $<q^{(te)}(s),\ q^{(p)}(s),\ DQS(s)>$, where $q^{(te)}(s)$ is request processing time, $q^{(p)}(s)$ is cost for invoking the service, and $DQS(s)$ is a set of dependent qualities. For composite service *CS*, EX_QoS model of *CS* can be defined as: $EX\_QoS(s)=<q^{(t)}(CS),\ q^{(p)}(CS)>$, where $q^{(p)}(CS)$ is the sum of cost for invoking all services in CS and $q^{(t)}(CS)=\sum_i q^{(ti)}\left(s_i\right)+DQ\left(CS\right)$ is duration time of *CS* execution.

4 Failure Prediction Based Ex_QoS Driven Adaptive Approach

We propose an approach (Figure 1) for composite service healing itself when an execution problem is occurred. This approach uses a re-selection process to make the composite service adapt to the failures as most of the works have been done. However, compared with them, this paper uses an offline re-selection process. Since that in a composite service, the invocation of one web service will not begin until all the predecessors of this service are finished. Usually, the re-selection process is to re-select the QoS violated service and its successors. Then, the re-selecting process will not affect the process of composite service execution, when the QoS violated service needs not to invoke. Therefore, when it needs not to invoke the QoS violated service, the re-selection

process and the execution process of composite service can be done asynchronously.

Figure 1 shows the architecture of execution environment. The composite service execution and the re-selection process will be done in asynchronously. When a failure is perceived, the re-selection will be triggered and the monitor of re-selection is activated. Monitor of composite service execution can reflect when the service in the re-selection slice needs to be executed. If the composite service needs to invoke the QoS violated service and the corresponding re- selection is not finished, the composite service execution controller will be informed of such situation by monitor of execution and monitor of re- selection. In this situation, composite service execution controller will stop the execution of composite service until the controller is informed that the re- selection process is finished.

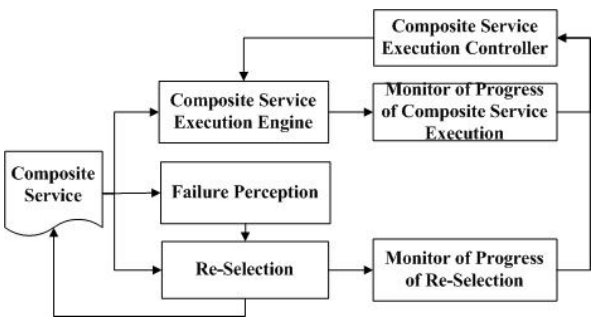


Figure 1 Offline Re-selection Execution Environment

In this framework, if a service is perceived to incur a failure much earlier before its invocation, there will have longer time to do the re-selection offline. This means that the interrupting time of composite service execution caused by the re-selection will be minimized. For doing so, a failure prediction is needed and this paper proposes a semi-Markov model based failure prediction approach. Meantime, if the re-selection is efficient, the interrupting time can also be minimized. Since that the efficiency of the re-selection process is influenced by the number of services in the re-selected slice. Minimizing the scale of the re-selection slice will improve the efficiency of the re-selection. Then, this paper proposes an algorithm for generating the re-selected slice of composite service based on the proposed slice determining rules and quality affecting degree. The details of the framework will be discussed in the following.

5 Details of the Framework

5.1 Semi-markov model based failure prediction

Web services operate autonomously within a highly variable environment (the Web). As a result, their QoS may evolve relatively frequently; either because of changes caused by service provides (e.g. service provider can minimize the price for invoking the service or improve the request processing time of the service), or because of changes caused by the network (e.g. higher network load may affect the data transmission time). Compared with changes caused by service providers, changes caused by the network may be occurred more frequently. Changes caused by the network may affect the data transmission speed and thus, affect the response time of composite service. Therefore, in this paper, we will try to predict the data transmission speed. The work of this paper is based on the following assumptions: (a) speed of processing the request is a constant value; (b) the price of requesting a service is a constant value; (c) the failures at different service and communication links are independent; (d) during the data transmission process, data transmission speed is a constant value.

We introduce semi-Markov model [8] to describe time-dependent stochastic behaviors of data transmission speed. The data transmission speed can be divided into 3 states: Qualified state, Soft Damage state and Hard Damage state. The definition of above 3 states is given as following, where We use $V(t)$ to signify the data transmission speed at time t and $ST(t)$ to signify the state at time t.

Definition 4. States of Data Transmission Speed. We use th_V_Q to signify the threshold of data

transmission speeds in Qualified state.

1) If V(t)>= th_VQ, then ST(t) =Qualified state;

2) If 0<V(t)<=th_VQ, then ST(t)=Soft Damage state;

3) If V(t)=0, then ST(t)=Hard Damage state.

In this paper, th_V_Q is the average data transmission speed Data transmission speed in soft damage state may be caused by the increasing of Network load and after certain time, the data transmission speed can automatically be improved. Data transmission speed in hard damage state may be caused by the hardware of Network, the data transmission speed will be improved only after the manually rescuing the hardware and may need more rescuing time.

Definition 5. Semi-Markov Model for Data Transmission Speed. Let Ω be the state space of data transmission speed Ω={1, 2, 3}. $Z=\{Z_t;\ t>=0\}$ is the random procedure on Ω. If the following conditions are true, we call that $Z=\{Z_t;\ t>=0\}$ is a semi-Markov process.

1) If current state is i, the next state will be entered is j with probability P_{ij} and $\sum_j P_{ij}=1$. Especially, $P_{ii}=0$;
2) Given that the next state entered will be j, the time it spends at state i until the transition occurs is a holding time t with distribution $F_{ij}(t)$.

Let $H_i(t)$ be the distribution of holding time in state i, $H_i(t)=\sum_j F_{ij}(t) * P_{ij}$. The average holding time in state i can be signified as μ_i. According to lemmas [8] of semi-Markov model, there exists stationary distribution $\pi=[\pi_1, \pi_2, \pi_3]$ and for each π_j, it can be computed as Eq.(5). Also, let P_i the steady-state occupancy probability of state i, it can be computed as Eq. (6).

$$\pi_j=\sum_{i=1}^{3}\pi_i P_{ij};\sum_{i=1}^{3}\pi_i=1 \tag{5}$$

$$P_i=\frac{\pi_i\mu_i}{\sum_j \pi_j\mu_j} \tag{6}$$

In order to predict the future state, it is required to get the context related to data transmission speed.

Definition 6. QoS-Related Context. The QoS-related context observed by observation o can be defined as $QC(o)=<t_{ob}, v, st_{ob}, s_a, s_b>$, where t_{ob} is observing time; v is the observed data transmission speed at t_{ob} between service s_a and s_b; st_{ob} is state.

In order to simplify the problem, this paper assumes that data transmission speed in each state has discrete distribution as Eq. (7) shows.

$$P(v\le V \wedge Z=i)=F_i(V)=\frac{n(v\le V)}{n} \tag{7}$$

Where, $n(v<=V)$ is the number of contexts of which the data transmission speed v is not below the data transmission speed V in state i, n is the total number of contexts in state i.

The aim of prediction can be described as: if the current state is i, current time is t and the holding time in current state is d, we need to predict the probability of the data transmission speed V_f at future time t_f above the expected speed V_e. Let j be the state V_e belongs to. To solve this problem, we will consider the following two situations:

3) State j is same to i

State j is same to i and during tf-t, no transition is occurred.

Let D_i be the random variable of time kept in state i. Under the situation when current time is t and the holding time in current state is d, if no transition is occurred during t_f-t, it means that the time kept in state i will be $d+t_f-t$. Then, the probability of data transmission speed V_f at future time t_f below the expected speed V_e is computed as Eq. (8).

$$\begin{aligned}&P((V>Ve)\wedge(D_i>d+t_f-t\mid D_i>t))\\&=P(V>V_e)*P(D_i>d+t_f-t\mid D_i)\\&=(1-F_i(V))*\\&\quad\frac{P(D_i>d\mid D_i>t_f-t+d)*P(D_i>t_f-t+d)}{P(D_i>d)}\\&=(1-F_i(V))*\frac{1-H_i(t_f-t+d)}{1-H_i(d)}\end{aligned} \tag{8}$$

Where, $P(V>V_e)$ is the probability of data transmission speed above V_e; in this paper, $H_i(x)$ is a distribution of discrete time variable x.

4) State j is same to i and during t_f-t, at least one transition is occurred.

Let D_i be the random variable of holding time in state i. Under the situation when current time is t and the duration in current state is d, if at least one transition is occurred, it means that the time kept in stat i before the transition will be shorter than d+t_f-t. Then, the probability of data transmission speed V_f at future time t_f above the expected speed V_e is computed as Eq. (9).

$$\begin{aligned}
&P\left(\left(V>V_e\right)\wedge\left(Z_{t_f}=i\right)\wedge\left(d<D_i<d+t_f-t\middle|D_i>d\right)\right)\\
&=P\left(V>V_e\right)*P\left(Z_{t_f}=i\right)*\\
&\quad P\left(d<D_i<d+t_f-t\middle|D_i>d\right)\\
&=\left(1-F_i\left(V_e\right)\right)*P_i\\
&\quad *\frac{P\left(D_i>d\middle|d<D_i<t_f-t+d\right)*P\left(d<D_i<t_f-t+d\right)}{P\left(D_i>d\right)}\\
&=\left(1-F_i\left(V_e\right)\right)*P_i*\frac{H_i\left(t_f-t+d\right)-H_i\left(d\right)}{1-H_i\left(d\right)}
\end{aligned}\tag{9}$$

Where P_i is the steady-state occupancy probability of state i.

Then, the probability under the situation that state j is similar to i, can be computed as Eq. (10).

$$\begin{aligned}
&P\left(\left(V>V_e\right)\wedge\left(D_i>d\right)\right)\\
&=P\left(\left(V>V_e\right)\wedge\left(D_i>t_f-t+d\middle|D_i>d\right)\right)\\
&\quad +P\left(\left(V>V_e\right)\wedge\left(Z_{t_f}=i\right)\wedge\left(d<D_i<d+t_f-t\middle|D_i>d\right)\right)\\
&=\left(1-F_i\left(V_e\right)\right)*\frac{1-H_i\left(t_f-t+d\right)}{1-H_i\left(d\right)}+\left(1-F_i\left(V_e\right)\right)*\\
&\quad P_i*\frac{H_i\left(t_f-t+d\right)-H_i\left(d\right)}{1-H_i\left(d\right)}
\end{aligned}\tag{10}$$

5) State j is different from i.

If state j is different from i, it means that there exist at least one transition during the duration from t to t_f. Then, the probability of data transmission speed V_f at future time t_f above the expected speed V_e is computed as Eq. (11).

$$\begin{aligned}
&P\left(\left(V>V_e\right)\wedge\left(Z_{t_f}=j\right)\wedge\left(d<D_i<d+t_f-t\middle|D_i>d\right)\right)\\
&=P\left(V>V_e\right)*P\left(Z_{t_f}=j\right)*P\left(d<D_i<d+t_f-t\middle|D_i>d\right)\\
&=\left(1-F\left(V_e\right)\right)*P_j*\frac{H_i\left(t_f-t+d\right)-H_i\left(d\right)}{1-H_i\left(d\right)}
\end{aligned}\tag{11}$$

Definition 7. QoS Failure of Services. Considering a service s in composite service CS, if it is predicted that the probability of the data transmission speed during its execution time below the predefined threshold is lower than rd, service s is assumed to incur a QoS failure that will affect the global QoS of composite service.

5.2 Triggering re-selection based on failure prediction

The algorithm presented in Algorithm 1 describes the proposed re-selection triggering approach. The basic idea is to predict the data transmission speed of each service in the composite one, and when the predicted speed below the threshold, the confident degree is also low (which means a failure will be happened at invocation time) and no replacement one can be used to rescue such failure, the re-selection will be triggered. The prediction will be based on the semi-Markov model as describe in the former section. The confident degree here is the probability of data transmission speed between the service and a reliable service (we assume that there exists such a service used only for testing the confident degree).

Algorithm 1. Re-Selection Triggering Algorithm.

servcie TriggerReselection(*CS*, *ReplaceCS*)//*CS* is a composite service; *ReplaceCS* is a replacement one.

1 begin

2 for each service s in *CS* do

3 if $\exists$ s_u, QoSPrediction(s, s_u, $T(s)$, $Q(s)$)< *ThresholdP*(s) and QoSPrediction(s, s_r, $T(s)$, $Q(s)$)<*ThresholdP*(s) then //$T(s)$ and $Q(s)$ are the invoking time and expected QoS of service s respectively; s_r is the reliable service

4 if $\neg\exists RCS\in ReplaceCS$, all services in *RCS* is predicted good then

5 return s;

6 return *null*;

7 end

5.3 Determining re-selected slice of composite service

In this paper, we give the following rules to determine the approximately re-selected slice of composite service. The rules (Figure 2) are designed according to the position of the failed service in original composite one and affecting degree between services.

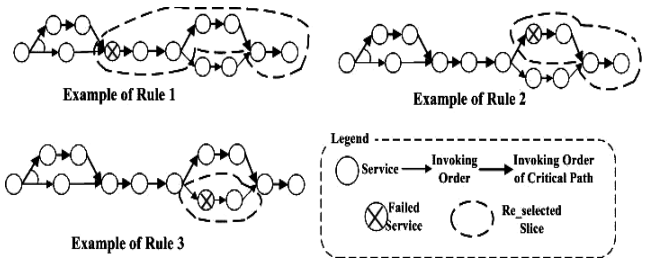


Figure 2 Rules for Determining Re_Selected Slice

Rule 1. This rule is used for the situation when the failed service is not in the parallel or fork block of the composite service. The approximately re-selected slice will be the services which are not invoked before the failed one and are in the critical path.

Rule 2. This rule is used for the situation when the failed service is in the parallel or fork block of the composite service and the failed service is in the critical path. The approximately re-selected slice will be the services which are not invoked before failed one and are in the critical path.

Rule 3. This rule is used for the situation when the failed service is in parallel or fork block of composite service and the failed service is not in the critical path. The approximately re-selected slice will be the services which are not invoked before the failed one, are in the same path as the failed one and will not be invoked after the last service of the block.

Since the re-selection is a NP hard problem, runtime of which is affected by the number of service classes, this paper tries to minimize the scale of approximately re-selected slice based on quality affecting degree.

When the transmitted data amount is bigger, the dependent quality between services is big and substituting one service will affect the execution duration of another one. $M\_Score=[ms_1,\ldots, ms_u]$ is the vector for scoring transmitted data amount, $\forall\ ms_i \in M\_Score$, $ms_i=(score, A)$ which means that when the transmitted data amount is A, it is can be scored by $score$ ($0 \le score \le 1$). $\forall\ ms_i, ms_{i+1} \in M\_Score$, $ms_i.score<ms_{i+1}.score \wedge ms_i.A<ms_{i+1}.A$. For CS, the score of transmitted data amount between two service s_i and s_j can be calculated as Eq. (12).

$$SCORE\_M(s_i,s_j)=\begin{cases}0, if\ a_{ij} \le ms_1.A\\ 1, if\ a_{ij} \ge ms_u.A\\ \dfrac{a_{ij}-ms_q.A}{ms_p.A-ms_q.A}+ms_q.score,\\ \quad if\ ms_1.A<a_{ij}<ms_u.A\end{cases} \tag{12}$$

Where, if $ms_{r2}=ms_{r1+1}$, then $ms_p=ms_{r2}$, $ms_q=ms_{r1}$; else $ms_p=ms_{r1}$, $ms_q=ms_{r2}$ (ms_{r1} and ms_{r2} are the referred nodes of m and Eq. (13) gives how to calculate them).

$$ms_{r1}=ms_r, \left|a_{ij}-ms_r.A\right|=\min_{ms_i \in M\_Score}\left\{\left|a_{ij}-ms_i.A\right|\right\},$$
$$ms_{r2}=\begin{cases}ms_{r1+1}, if \left|a_{ij}-ms_{r1+1}.A\right|<\left|a_{ij}-ms_{r1-1}.A\right|\\ ms_{r1-1}, else\end{cases} \tag{13}$$

When the bandwidth between two services are much bigger than that of bandwidth between service and any service in the same service class as the successor one, the score of the bandwidth is higher and the invoking time dependent relation is strong which makes the substitution of one service affect the execution duration of another. For composite service CS. The bandwidth between s_i and s_j is signified as b. Let sc be the service class s_i belonging to. $\forall\ s_v \in sc(v<>i)$ the bandwidth between s_v and s_j is signified as b_v, and the set of bandwidth can be formed as $B=\{b_1,\ldots,b_v\}$. The score of bandwidth between s_i and s_j can be calculated as Eq.(14).

$$SCORE\_B(s_i,s_j)=\frac{b_{\max}-b}{b_{\max}-b_{\min}} \tag{14}$$

Where, b_{max} and b_{min} are max and min values in B.

Definition 13. Quality Affecting Degree. The quality affecting degree can be computed as Eq. (15).

$$SCORE(s_k,s_l)=ws^{(M)}*SCORE\_M(s_k,s_l)$$
$$+ws^{(B)}*SCORE\_B(s_k,s_l) \tag{15}$$

Where, $ws^{(M)}$ and $ws^{(B)}$ are the weights and $ws^{(M)}+ws^{(B)}=1$.

Through calculating the quality affecting degree between services as Eq. (15), the approximately re-selected slice will be minimized. Algorithm 2 describes the proposed re-selected slice determining approach.

Algorithm 2. Re-selected slice determining algorithm.

FS DetermineSlice(*failedS*, *CS*) //*failedS* is the failed service return by Algorithm 1; *CS* is the composite service

1 begin

2 CPath=GetCriticalPath(CS);//get the critical path of CS

3 Blocks=GetBlock(CS);//get all parallel and fork blocks of CS

4 if $\neg\exists$ *Block* $\in$ *Blocks, failedS* $\in$ *Block then //using rule 1*

5 for each service $s_i \in CS$ *do*

6 if s_i *will be not invoked before failedS and* s_i *in CPath then*

7 put the service class of s_i *into FS;*

8 else if failedS in CPath then //using rule 2

9 for each service $s_i \in CS$ *do*

10 if s_i *not invoked before failedS and* s_i *in the CPath then*

11 put the service class of s_i *into FS;*

12 else if failedS not in the CPath then //using rule 3

13 for each service $s_i \in CS$ *do*

14 if s_i *will not be invoked before failedS and* s_i *in the Block and* s_i *in the same path as failedS then*

15 put the service class of s_i *into FS;*

16 for i=FS.Getlength to 2 do

17 if Score(i, i-1)<λ *then*

18 remore i from FS;

19 return FS;

20 end

5.4 re-selection for finding replacement composite service

The re-selection will be done on the re-selected slice of the composite service. The problem of such re-selection can be described as:

$$\max_{v} \; F(CS_v) \quad s.t \;\; Q^t(CS_v) \le Q_c^{t'} \tag{16}$$

Where, $CS_v=\{(sc_1, s_{1,k1}),\ldots,(sc_i, s_{i,ki}),(sc_n, s_{n,kn})\}$, here, sc_v is the service class in *FS.FS* is the set of service classes in re-selected slice; $Q_c^{t'}$ is the runtime of original re-selected slice which can be used as runtime constraint.

Such problem can also be solved by dynamic programming algorithm [10]. As the limitation of this paper, we will not describe the algorithm in detail.

6 Experimentations

Experimentation 1 is used to test performance of determining re-selected slice through minimization based on quality affecting degree. Randomly generate 200 scenarios and in each scenario, the number of services for each service class is 10, the average of quality affecting degree between services is 0.15, the number of service classes in approximately re-selected slice will be 5, 10, 20, 40 and 50. The threshold λ of quality affecting degree is 0, 0.1, 0.15 and 0.2 respectively. Figure 3 shows runtime of the proposed algorithm (a combination of re-selected slice determining algorithm and dynamic programming algorithm). Table 1 gives the quality of found replacement composite service.

Table 1 Quality of found replacement composite service

λ	0	0.1	0.15	0.2
Synthesized Quality F(x)	0.87	0.83	0.71	0.56
Price q(p)	32	34	28	35
Duration Time q(t)	124	127	135	141

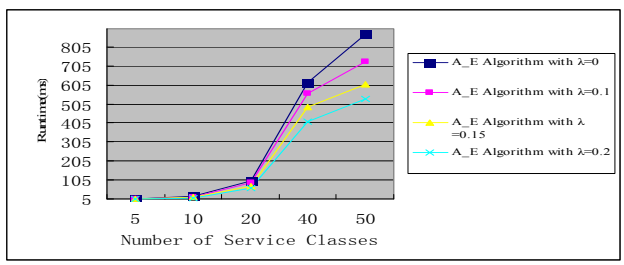


Figure 3 Runtime Comparison

Table 1 shows the relation between λ and the quality of the composite service. If λ is smaller, the quality of found replacement one will be better. Fig 3 shows that if is bigger, the runtime of re-selection will be lower. From experimentation 1, if λ can be set properly, the re-selection will be more effectively.

Table 2 Semi-Markov Based Predicted Result

	OQ=0.5s			OQ=0.1s			OQ=0.05s		
	I=10	I=60	I=180	I=10	I=60	I=180	I=10	I=60	I=180
N	300	300	300	200	200	200	150	150	150
R%	95	86	80	97	93	83	98	95	92

Experimentation 2 is used to test the effectiveness of the proposed semi-Markov model based QoS predicting approach. Simulate test set of data transmission speed according to the Gaussian distribution. The threshold of failure probability is 0.9. The size of QoS-related contexts is 100000. Compare the relation among predicted result, predicting interval and the observation interval between two neighboring contexts. Table 2 gives the result (O_Q is the observation interval between two neighboring QoS-related contexts in the test set; N is the number of predictions; R is the average accurate rate of the predictions which can be computed as (number of predictions-number of predictions that is right)/number of predictions; I is the predicting interval and the unit of I is second).

Table 2 shows that if the observation interval between two neighboring contexts is smaller and the predicting interval is shorter, the prediction will be more accurate. When the observation interval is short enough, although predicting interval is a little bigger, the accurate of prediction will be better also. Thus, through minimizing observation interval, the accuracy of prediction result can be improved

Experimentation 3 is to test the interrupting time caused by re-selection process. Randomly generate 10 scenarios, compare the interrupting time, the result of which is shown in Figure 4.

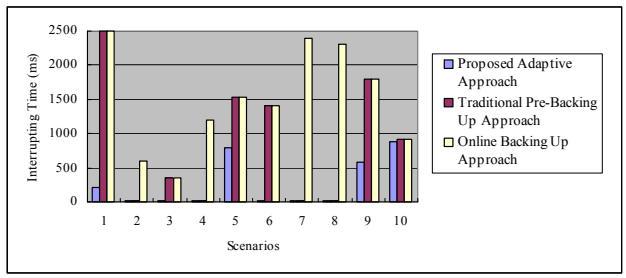


Figure 4 Comparison of Interrupting Time

Figure 4 shows that the interrupting time of the proposed approach is always the least one among the three approaches. This is because that the proposed approach is a complementary of the traditional pre-backing up approach. This will make that when the replacement one is not available and service is failed, re-selection process will start as soon as possible through failure prediction. Thus, the re-selection process will occupy as few as possible execution time of composite service.

Experimentation 4 is to test the effectiveness of the proposed EX_QoS model in decentralized service composition. We use satisfied degree of execution duration to evaluate the effectiveness of the proposed EX_QoS. Eq. (17) gives how to compute the satisfied degree of execution duration. Give 200 scenarios where the average amount of data transmission between service classes in the composite service can be 0, 10, 50, 100 or 200 bit. Compare the satisfied degree of execution duration of the composite services backed up according to the QoS model proposed by [1] and the EX_QoS model. Figure 5 shows the experimentation result.

$$SD\left(CS, cons^{(t)}\right) = \begin{cases} \dfrac{q^{(t)}(CS)}{cons^{(t)}}, q^{(t)}(CS) < cons^{(t)} \\ 1, else \end{cases} \quad (7)$$

Where, $q^{(t)}(CS)$ will be calculated as Definition 3.

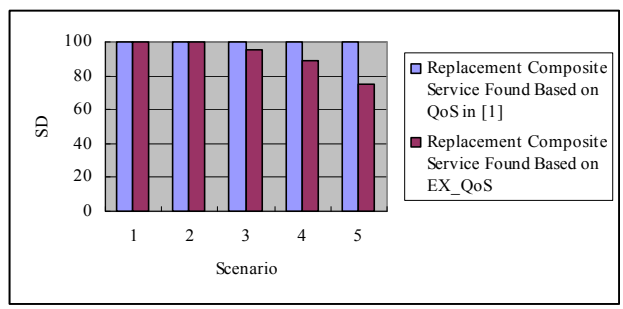


Figure 5 Relation between SD and average amount of data

From Figure 5, SD of the composite service backed up based on EX_QoS is higher than the one based on QoS in Ref [1]. It is because that EX_QoS considers the quailty dependent relation between services. EX_QoS is effective in selection in decentralized service composition.

7 Conclusions

In this paper, based on the proposed EX_QoS, we present a framework for adaptive service composition and the corresponding adaptive algorithms. The experimentations show that proposed approach can improve the availability of the replacement composite service, minimize the interrupting time and be effectiveness in decentralized service composition. This paper only considers one of the QoS dependent relations—invoking time dependent relation. In the future, we will better our QoS dependent model.

References

[1] L Z Zeng, B Benatallah. "QoS-Aware Middleware for Web Services Composition". IEEE Transactions on Software Engineering, 30(5), 2004, pp. 311-327

[2] F Casati, S Ilnicki, L Z Jin. "Adaptive and Dynamic Service Composition in eFlow". Advanced Information Systems Engineering, USA, 2000, pp. 13-31

[3] B Benatallah, Q Z Sheng, M Dumas. "The Self-Serv Environment for Web Services Composition". IEEE Internet Computing, 7(1), 2003, pp. 40-48

[4] T Yu, K J Lin. "Adaptive algorithms for Finding Replacement Services in Autonomic Distributed Business Processes". 7th International Symposium on Autonomous Decentralized Systems, China, 2005, pp. 427-434

[5] C Girish, D Koustuv, K Arun, M Sumit, S Biplav. "Adaptation in Web Service Composition and Execution". IEEE International Conference on Web Services, USA, 2006, pp. 549-557

[6] G Canfora, M D Penta, R Esposito, M L Villani. "QoS-Aware Replanning of Composite Web Services". IEEE International Conference on Web Services, USA, 2005, pp. 121-129

[7] T Yu, Y Zhang, K J Lin. "Efficient Algorithms for Web Services Selection with End-to-End QoS Constraints". ACM Transactions on the Web, 2007, 1(1)

[8] M Malhotra, A Reibman. "Selecting and Implementing Phase Approximations for Semi-Markov Models". Communication Statistics-Stochastic Models, 9(4), 1993, pp. 473–506

[9] Y Altinok, D Kolcak. "An Application of the Semi-Markov Model for Earthquake Occurrences in North Anatolia Turkey". Journal of the Balkan Geophysical Society, 2(4), 1999, pp. 90–99

[10] T Yu, K J Lin. "Service Selection Algorithms for Web Services with End-to-end QoS Constraints". Journal of Information Systems and e-Business Management, 2005, 3(2), pp. 103-126

Implementation and Design of Grid Video Education System Based on Web Services*

Xinyi Wu

College of Computer Science, Wuhan University of Science and Engineering, Wuhan, Hubei, 430073, P.R China
Email: lily_wu2007@yahoo.com.cn

Abstract

An implementation frame of remote video education system is given in this article, based on Web Services system of Grid, according to analysis of current remote video system on Internet and new technology of Grid Computing, to improve the teaching quality of remote video system and to connect different constructed video systems. The paper first gives an introduction of current situation of video education based on IP Network, and then gives the introduction of Web services、SOAP and so on. Finally, gives the implementation method and uses the Zipf' theorem to adapt the video scheduling algorithm to promote performance which makes the designing some far reaching significance for Grid Video System.

Keywords: Grid Technology, Web Service, SOAP, Video Education, System Interaction

1 Introduction

With the development of Continuing Education in colleges and Computer Network technology, the educational mode for remote network education has occurred the remote education mode can combine remote educational resources in different regions together to make the educational resources shared. And the mode can also supply vivid teaching videos to improve teaching effect.

At the same time, it is because the heterogeneous architecture of the whole internet, the educational resources can not be shared in the old remote educational system. With the fast development of computer network technologies, Grid technology has occurred which has become the hot points and leading edge of computer research. It breaks through the bottle-neck of share architecture in traditional internet resources to impulse the development of remote educational systems in modern times. With the application of Grid technology, it will change the current situation of remote education. It has great significance to do research in.

In this article, first the author will give the detailed introduction of technologies of Grid service architecture based on Web Service [4], then the author will give out the details of implementation of the Grid video educational system platform based on Web Service.

2 Introduction of Grid Technology

The Grid technology occurred in the beginnings of the twentieth century, and it has become the research trend in network resources. The concept of Grid is originally draw out form the concept of electronic networks .It means that the user can get the computing and information resources when the user is connected to Internet ,just like putting the electronic pin into the electrical outlet without knowing where the computing and information resources come from. It is obvious that Grid will change the mode、structure and methods which have been adapted by Internet users totally. Grid differs

* Project Supported By Hubei Education Foundation under Grant NO.200717005.

from traditional Internet in which constructed in OGSA (Open Grid Service Architecture).And it can carries out great scale computing task in collaboration dynamically.

The goal of Grid system[3] is to connect the heterogeneous computing resources in deferent places by high speed networks, to resolve very large scale application problems, to supply the discrete remote information sharing. The Grid System is new type of basic establishment constructed in Internet, like the NSFNET which is the original model of Internet. Grid has attracted much attention of industrial and academic researchers in China and abroad, and the research of Grid will be a scientific work full of significance. At the same time, the research of video educational system based on Grid will be a new hot point of research.

In the problem field of remote video Grid discussed in this article, the author will focus on the implementation details of construction video grid platform on Internet to make it possible for remote high quality video education.

3 Current Situation of Video Educational System Based on IP

We need to considerate the different video system interfaces in current Internet when we try to connect the different video systems in WAN using Grid Video technology. It is necessary to construct the Grid Video System on the interactive part of video system among different area video systems when we design the Grid Video Educational System based on Web Service. The goal is to protect the invested money of current remote video systems and to make full use of them. The different Interactive Video Systems constructed on Internet currently are divided into three types, and there will be the introduction separately:

A. Interactive Video System based on H.323 [6]:H.323 protocol is designed for the existing running multimedia system in LAN. It has made it possible for real time multimedia communication and meeting on network based on packets switching to come into reality. And it also makes it possible to manipulate the different terminals from different manufacturers in deferent networks interactively, to communicate each other between other products competing to H.32x.

B. Interactive Video System based on SIP [7]: SIP (Session Initiation Protocol, SIP，in RFC 2543). SIP is a signal control protocol proposed by IETF, which is a important part of the whole multimedia data communication and control structure proposed by IETF.SIP is used to video data transporting and interactivity, belongs to signal protocol based on text. SIP has good characters such as independent from the Transport Layer, can be implemented in UDP; having abroad application field; easy to be implemented by the OOP language such as JAVA and PERL, having flexible structure to extend easily; having good interfaces for other protocol such as DNS、RTP、RSVP、RTSP、SAP、SDP and so on, not needing to add other new services.

C. Video Meeting System based on IETF's MMUSIC Frame and IP Multicast technology, it is developed by IETF's MMUSIC special group to multicast the video data packets using UDP or TCP protocol to the point which has connection established or video data subscribed.

It can be founded from comparison among the three current Interactive Video Systems that they are all make the own users group to join the meeting by constructing virtual meeting video room. But the three systems mentioned above have their own structures and implementation conditions which has made it difficult to communicate each other between the three video systems. There are many isolated Interactive Video islands in Internet.

There is a new video service model occurred in network service –Web Service, which can supply standard interfaces and communication channel to combine the multi heterogeneous-architecture application programs. It is possible to connect isolated Interactive Video islands by Grid video Platform, to supply Interactive Video services in heterogeneous combined video systems. The new structure can make the different users from different video systems to join the same remote video and voice meeting .Users can join the

combined meeting by any client terminal from the mentioned three video systems, such as H.323 devices, SIP devices, and bone devices and so on, from the video islands in Internet. And we will give the implementation details of technology in Web Service and Grid Video Platform based on Web Service architecture.

4 Implementation and Design of Grid Video Education System

4.1 Introduction to web service

The Web Service [6] Frame has proposed a new computing mode, with the goal to solve the Interactive manipulation between applications supplied by heterogeneous-architecture platforms. The usage of OOP technology and Middleware technology has been expanded to Internet after Web Service has been invented, and it brings new revolutionary changes to the architecture of software and mode of service, supplies a very good solution for the problem of Interactive manipulation between different types of objects and Middleware.

Web Service has implemented the three necessary characters of Web Computing: Encapsulation, Loose Coupling, and Dynamic Biding. Web Service technology is open web technology based on XML, and it can be used to constructed new distributed application platform. Web Service technology is a kind of self-including, modulable program, supplying many functions for enterprises and individuals from network access. Its interface setting and service bindings can be described and founded by XML Middlewares, and directly communicate with XML Middleware by Internet protocols. The structure of Web Service can be found in the Figure 1.

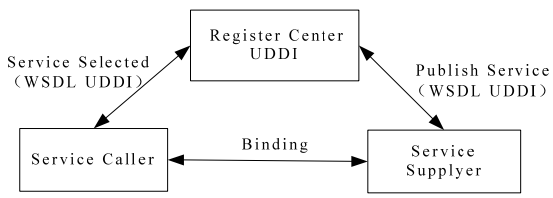


Figure 1 the Structure of Web Service

There are three components in the architecture of Web Service: Service Supplier, Service Caller, and UDDI. Web Service is implemented by Service Supplier and registered into the directory of UDDI. Service Caller first looks for the web service which will be used in the directory of UDDI, and attain the WSDL of the service, call the service from the binding of Service Supplier. Finally the Service Caller communicates with Service Supplier directly to attain the service.

4.2 Grid architecture based on web service

There are many projects abroad concentrated on Grid research, such as Globus, Legion and Web Service which has been carried into commercial use. All of them have great contribution to research of Grid architecture. On the GlobusWorld meeting held in Jan 20th, 2004, Ian Foster first propose the concept of WSRF (Web Service Resource Framework), a new service frame of OGSI [1]**.** And OGSA based on Web Services is the latest Grid architecture, which includes two critical technologies, such as Grid technology and Web Services technology. The main goal of OGSA is processing the service as the core. In the architecture of OGSA, every thing is abstracted to service, such as computer, programs, data, and equipments and so on, and then the abstracted services are registered in the LDAP[1] tree in resource server, making it to be one child node of the Global Grid Network. The architecture of OGSA Service can be founded in the Figure 2.

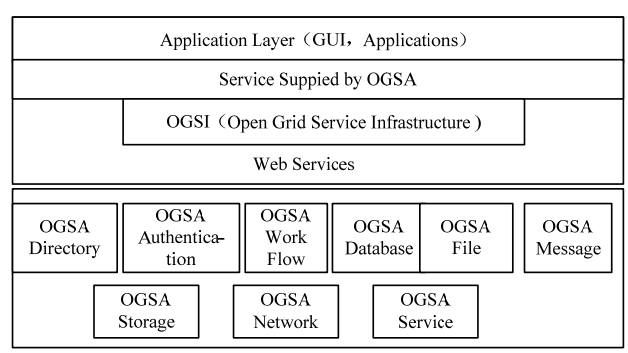


Figure 2 The Architecture of OGSA Service

Globus Toolkit has been used as the Grid Service Container in implementation of OGSA, and the latest

version of Globus Toolkit is GT4[10], which has added many new APIs to support WSRF.

4.3 Summary of the grid video system based on web service

The remote Grid video educational system platform based on Web Service introduced in this article is constructed by a video center and several local and external groups, which is designed into layered mode and layered managed. The whole system is divided into two layers, the first layer is the kernel data management layer which consists of video center to control groups in the second layer and supply interactive video manipulation between groups. The second layer consists of local and external groups, every group include a Agent and serial user point, and the Agent in charge other user points, to add them into remote interactive video and receive and send multimedia data flow. Those groups, which contain video Center's computers, are called Local group, the other groups are called External group.

Local group and External group supply the same Web Services GUI to users in Jap. And the video Center treats them as the same. Local group and External group communicate with each other in SOAP [8] message (It can be founded in Figure 3).They communicate with video center in SOAP message too. The new group registered itself in LDAP tree in video with the video can be supplied by it, and accessed by other groups. The whole system constructed in web Service and encrypted SOAP message which can be implemented in HTTP packet by XML to transport across different types of networks.

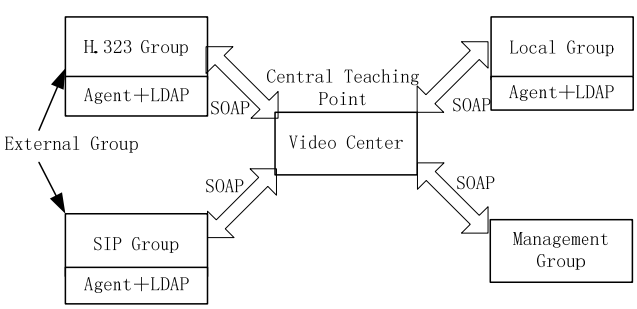


Figure 3 Architecture of Grid Video System

4.4 Soft architecture and implementation of system node

In implementation of the system, the interactive operating between video center and Local group and External group, is abstracted into service according web service frame, including Video Interactive Establishing, Naming Service, Directory Service, Network Probe, Voice/Video service and so on. The soft architecture of system node can be be found in the Figure 4.

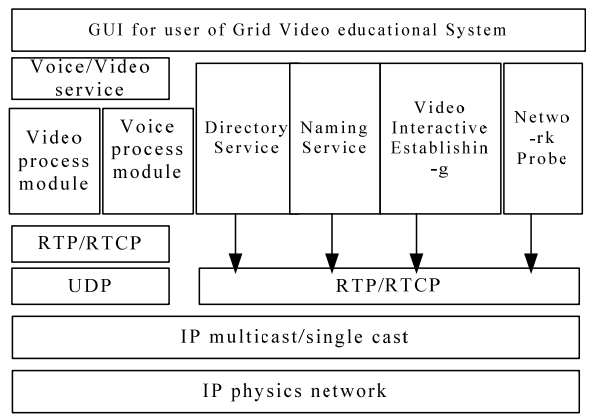


Figure 4 The Soft Architecture of Target System

4.5 Adoption of video scheduling algorithm for video access

In order to promote the access efficiency of every educational video and reduce the response time of video, the Zipf' theorem is introduced to adapt the video scheduling algorithm in the video center in Figure 3. Zipf' theorem can be described as follows:

Among the N educational videos (N>5000), the access rate of the video ranked k of the most popular videos is approximately C/k (Chervenak, 1994), with the condition Eq. (1) (Zipf' 1949). And we can get Eq. (2). After making every side of the equation the logarithm (base 10) we get Eq. (3).

$$C = 1/(1+1/2+1/3+\ldots+1/N) \tag{1}$$

$$P(k) = C/k^{\alpha} \tag{2}$$

$$\lg P(k) = \lg C - \alpha \lg k \tag{3}$$

K: the rank of access frequency of a movie in per

unit time;

P(k): the access frequency of movie ranked k in per unit time.

After inputting the access data of every video, we get the diagram (*X*-*k* , *Y*- the access frequency of every video in per unit time) by Matlab 6 in Figure 5 in which the curve is approximately a line ,and we get the value of α ($\alpha \approx 0.58$) to stand for statistics property of video access.

Using the value of α to adapt the video scheduling algorithm in the video center will promote performance of video center, and we can get a new value of α after regulating the video scheduling algorithm in the new cycle and at last the value of α varies around a contain value with narrow scope in the contain case .

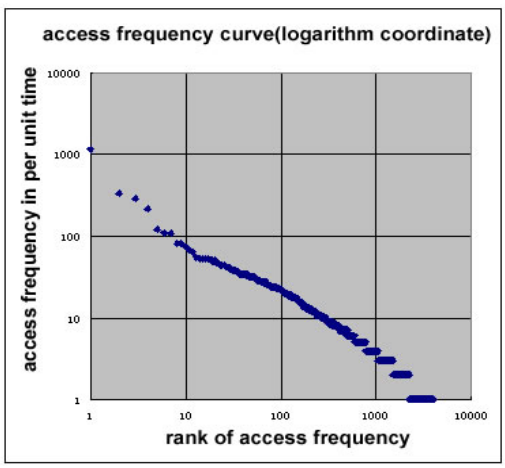


Figure 5 Relation of Access Frequency and Rank

5 Conclusions

With the widely use of Grid technology and fast development of remote educational system, the demand to Grid video educational system platform based on Web Service has been raised. At the same time, Grid will be the original model of next generation Internet drawing our attention. So the Grid video educational system platform based on Web Service has much value both in thesis and real project, and the implementation of it will supply good example for remote education in colleges, to make full use of current video devices and supply precise experience for the vivid development of remote education in college.

Refrences

[1] Ian Foster, "Computing of Grid," Publishing House of Electronics Industry, October 2004

[2] Jin Hai, Zhou De Qing et al, "The implementiation of grid system based on Web. Internet and cluster computing center," College of Computer Science.Huazhong University of Science and Technology, Mini-Micro Systems, December 2003

[3] Jin Hai, Li Qi Sheng et al, "Implementiation of resource broker in haohan grid system," College of Computer Science.Huazhong university of Science and Technology, Computer Applications and Software, January 2004

[4] Li Lusong, Li Jing et al, "Design and implementation of videoconference system based on Web services," State Key Lab. of Software Development Environment, Beihang University, Journal Huazhong Univ. of Sci. & Tech (Nature Science Edition) , Vol. 31, October 2003

[5] Du Zhihui, Chen Yu, Liu Peng, Li Lisan, "Grid Computing[M]," Tsinghua University Press, 2002

[6] Bi Minna, Wang Qingyang, X et al, "The Video conference and its Application based on H.323," Micro Computer Information, Vol. 22, 2006

[7] Liu Xiaorong, "The Simple Research and Comparison between SIP and H.323 Agreements," College of Information Engineering, East China Institute of Technology, Fuzhou, Science Mosaic, April 2006

[8] Scott Seely, "SOAP：Cross Platform Web Service Development Using XML", Publishing House of Engine Industry, April 2002

[9] Zhang Xinsheng, Zheng Jianbing, "Research of IP video meeting technology based in software switching," Wuhan University of Technology, Computer and Information Technology, June 2005

[10] GT4 Early Access WSRF, http://www-128.ibm.com/developerworks/grid/library/gr-gt4early/

The Development of E-Commerce System Based on Model-Driven Architecture

Xiaojun Li

College of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou, Zhejiang, 310035, P.R.China
Email：lixj@mail.zjgsu.edu.cn

Abstract

Traditional E-Commerce systems development methods are designed to develop systems which are a specific and static set of requirements. These methods result in systems that are sluggish in their response to dynamic conditions and changing requirements, expensive to maintain over extended periods of time, and prone to system failure. A model-driven approach extends traditional methods and is well-suited to a systems development environment characterized by rapidly changing conditions and requirements. This paper describes model-driven architecture and methodology used to develop E-Commerce system. Finally, a flower store system was developed by MDA.

Keywords: E-Commerce System, Development Method, Model-Driven Architecture, MDA

1 Introduction

The definition of ‘e-commerce’ or ‘electronic commerce’ given by the Electronic Commerce Association1 is: ‘electronic commerce covers any form of business or administrative transaction or information exchange that is executed using any information and communication technology’ [1]. Usually, we limit our approach to covering commercial activities conducted on the Internet. E-commerce offers opportunities to dramatically improve the way that businesses interact with both their customers and their suppliers, that is, to make business negotiations faster, cheaper, more personalized, and/or more agile. An e-commerce system is a set of resources that is organized to provide e-commerce services, such as registering, searching and buying products via internet etc. Traditional E-Commerce systems development methods are designed to develop systems which are a specific and static set of requirements. These methods result in systems that are sluggish in their response to dynamic conditions and changing requirements, expensive to maintain over extended periods of time, and prone to system failure.

In recent years, a Model-Driven Architecture (MDA) extends traditional methods and is well-suited to a system development environment characterized by rapidly changing conditions and requirements. MDA refers to a set of approaches in which code is automatically or semi-automatically generated from more abstract models, and which employs standard specification languages for describing those models and the transformations between them. This paper describes model- driven architecture and methodology used to develop the flower store system.

2 Overview of Model-Driven Architecture

MDA defined by OMG provides an open, vendor-neutral approach to the challenge of business and technology change. MDA is widely regarded as the next great leap in systems and software development enabling companies to manage more complex applications. MDA aims to bridge the gap between models and code and specifies a way of generating executable code for multiple platforms from one single Platform Independent Model (PIM). PIM is built by using UML and the other associated

OMG modeling standards, can be realized through the MDA on virtually any platform, open or proprietary, including Web Services, .NET, CORBA, J2EE, and others[2]. Therefore, MDA allows designers and developers to concentrate on implementing the specified requirements instead of spending valuable time adapting the system to a specific platform and environment. Since systems are modeled independent of the target platform, the ability to reuse at the design level is significantly enhanced.

Figure 1 shows the life cycle of MDA, including capturing requirements, analyzing requirements, designing, coding, testing and deploying. It has no great dissimilarity with traditional life cycle. The main difference lies in developing work piece, including PIM, PSM (Platform Specific Model) and code. PIM is a one-or-one mapping of the mental model into a more formal language such as UML. The developer can keep focus on the business logic, not the underlying technology, and the PIM can be reused later, it is not bound to any existing platform. The developer can model the PIM of applications by UML tools such as Rational Rose, ArgoUML, MagicDraw etc. The next step is to transform the PIM into the PSM which is relevant to specific platform, such as EJB, Spring, Web Services and Struts. Then, PSM can be marshaled into the code that would actually be written manually. Now, more and more tools can analyze the given PIM and construct a PSM with which templates are used to produce code, such as Rational Rose, AndroMDA, ArcStyler etc. That is to say, PIM can be transformed into code automatically by tools that embody MDA technologies. Therefore, the system development with MDA has significant improvements in software quality and time to value.

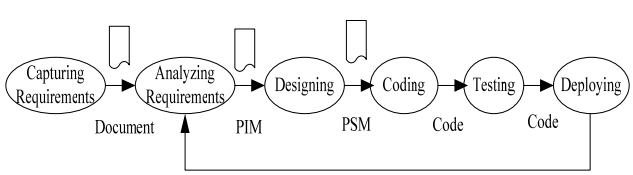


Figure 1 The Life Cycle of MDA

3 Architecture of E-Commerce System Based on MDA

Modern e-commerce systems are built using several components connected with one another, each providing a specific functionality. Components that perform similar types of functions are generally grouped into layers, such as presentation layer, business layer, data access layer and data stores. These layers are further organized as a stack in which components in a higher layer use the services of components in layer below. MDA tools take as its input a business model specified in UML and generate significant portions of the layers needed to build a system. In this paper, AndroMDA which is a famous open sourced MDA tool is adopted to develop a flower store system. AndroMDA can generate a layered java application according to PIM. Figure 2 shows various layers to Java technologies supported by AndroMDA.

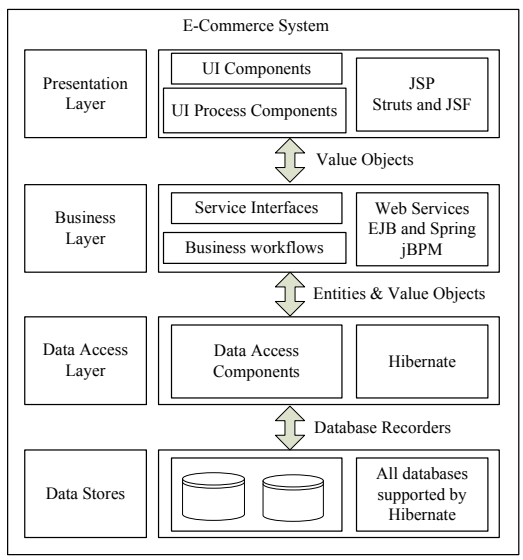


Figure 2 The Architecture of E-Commerce System Based On MDA

Presentation Layer: The presentation layer contains components such as web pages needed to interact with the user of the e-commerce system. AndroMDA offers Struts and JSF to build web based presentation layers.

Business Layer: The business layer encapsulates the core business functionality of the e-commerce

system. The business components are generally front-ended by a service interface that acts as a façade to hide the complexity of the business logic. The business layer generated by AndroMDA consists primarily of services that are configured using the Spring Framework. AndroMDA creates blank methods in implementation classes where business logic can be added.

Data Access Layer: The data access layer provides a simple API for accessing and manipulating data. The components in this layer abstract the semantics of the underlying data access technology thus allowing the business layer to focus on the business logic. Each component typically provides methods to perform Create, Read, Update, and Delete operations for a specific business entity. AndroMDA uses the popular object-oriented mapping tool called Hibernate to generate the data access layer for applications. It generates data access objects (DAOs) for entities defined in the UML model.

Data Stores: The e-commerce systems store their data in one or more data stores. Databases and file systems are two very common types of data stores. Since AndroMDA generated applications use Hibernate to access the data, the developer can use any of the databases supported by Hibernate.

In the architecture of e-commerce system based on MDA, different data is propagated between various layers. From the bottom up, the data access layers fetches records stored in relational databases and transforms them into objects that represent entities in the business domain. The data access layer passes the business entities to the business layer which performs business logic using these entities. The business layer packages necessary information into called "value objects" and transfers these value objects to the presentation layer. The presentation layer displays these value objects in the web pages.

4 Modeling the PIM of E-Commerce System

In this paper, the flower store was developed based on MDA. The flower store is a typical e-commerce system, presenting users with various views of products and services for sale; taking and acknowledging orders; processing credit cards; and managing user logins, shipping information, and shopping sessions. The flower store also includes administration functions, including inventory and order management. Figure 3 shows the use case diagram of flower store.

System development with the MDA starts with a platform independent model (PIM) of a system's business functionality and behavior, constructed using a modeling language such as UML. A PIM could represent a logical data model and consist of a number of entity classes, each with a number of persistent attributes. On the other hand, at PIM level, the business rule can be represented by behavior diagram (such as state diagram, activity diagram) or interactive diagram (such as sequence diagram, cooperation diagram).

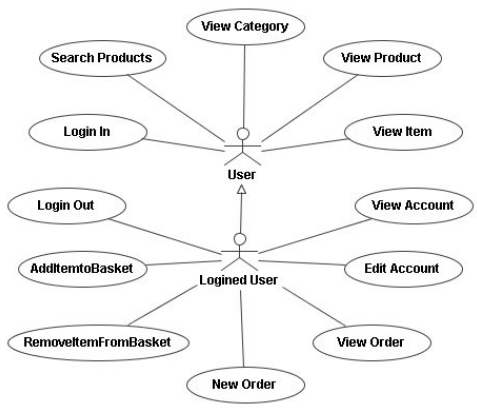


Figure 3 The Use Case Diagram of Flower Store

4.1 Entities modeling

An entity represents concepts in a problem domain. We use entities to model things in the real world such as products, purchase orders, etc. An entity contains an identity to guarantee uniqueness. At a PIM level, EC system developers mark or specify "which" entities need to be saved and "which" entity properties are used for identity. Then, the "how" the entities are implemented will be specified during the transformation step. Developers can then focus on business functionality and defer a persistence method decision and specific implementation choices later.

In flower store, there are some entities such as user, order, lineitem, item, product, category, etc. An order is an entity that can be identified by ordered. An order may contain many items. An item entity is relative to an product entity. Figure 4 shows the entities diagram of flower store.

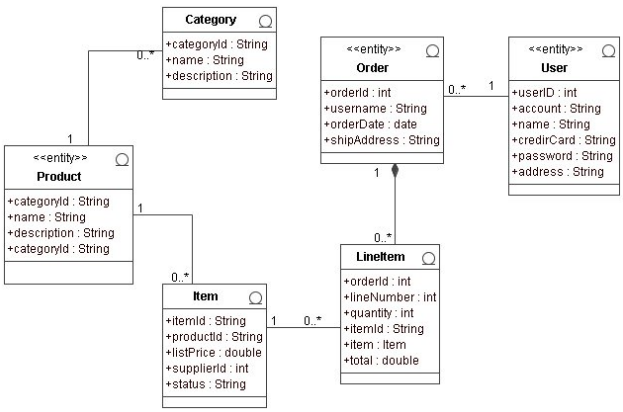


Figure 4 The Class Diagram of Entities Model

This model could be transformed through automation into a UML data model that captures the same underlying entities in the form of database tables.

4.2 Businges process modeling

A business process is a collection of activities designed to produce a specific output for a particular customer or market. A process is a specific ordering of work activities across time and place, with a beginning, an end, and clearly defined inputs and outputs: a structure for action. In UML, business process modeling can be created by behavior diagram (such as state diagram, activity diagram) or interactive diagram (such as sequence diagram, cooperation diagram). In our flower store development, the activity diagram is introduced to model business process.

An activity diagram is typically used for business process modeling, for modeling the logic captured by a single use case or usage scenario, or for modeling the detailed logic of a business rule. This diagram allows the developer to express the way he wants his application to behave, this is expressed by means of states and transitions. Therefore, an activity diagram is a state machine. In general, an activity diagram is composed by initial states, action states, transitions, and final states. The following section shows how to model "AddItemtoBasket" use case's business rule by activity diagram. Figure 5 shows the activity diagram of "AddItemtoBasket" use case's business rule.

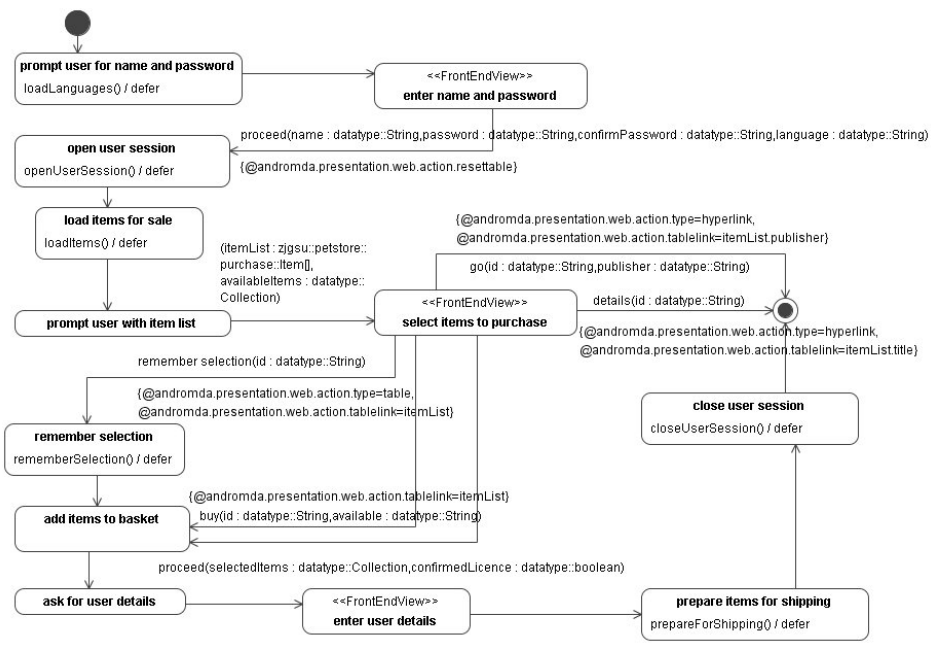


Figure 5 The Activity Diagram of Business Process

(1) Initial States

At PIM level, each use-case needs an initial state which denotes the starting point of the use-case. An initial state can have no incoming transitions and only one outgoing transition. In activity diagram, it is displayed as a solid black disc.

(2) Action States

An action state represents execution of an atomic action, usually the invocation of an action. An action state is displayed as a rectangle with rounded corners. The developer can use action states for both server-side states as client-side states. In activity diagram created by ArgoUML tool, client-side action states are tagged with the <<FrontEndView>> stereotype, they represent a JSP and may have multiple outgoing transitions. Server-side action states do not need any stereotype but can only have a single outgoing transition. In Fig5, "load items for sale" action sate is a server-side action state, and "select items to purchase" is a client-side action state.

(3) Transitions

Transitions are used to interconnect the different states in the activity diagram. They make up the actual process logic. Transition is displayed as a narrow line. A few tagged values exist for transitions coming out of <<FrontEndView>> action states. In our system, {@ andromda.presentaion.web. action. resettable} tagged value exist for transition coming out of "enter name and password" action state. If set this value to true if you want to be able to reset forms to their initial values. There are two tagged values exist for transition coming out of "select items to purchase". {@andromda.presentation. web.action.type=table} indicates that the type of trigger causing the action is table. And {@andromda.presentation. web.action.tablelink=itemList} denotes that this action applies on the information shown in a tabular format, "itemList" is the name of the table. On the other hand, transition exiting client-side action states have event parameters which represent the form fields. Such transition represents a call to the server from a webpage, usually by submitting a form.

(4) Final States

An activity diagram may have as many final states as the developer want in it. A final state represents the end of the use-case and the flow into the next use-case. Final states are displayed as bulls-eyes in activity diagrams.

5 Code Generation

After the PIM of E-Commerce system was created, the next step is to transform the PIM into program code. The MDA way of doing this is to gradually refine the model into a platform specific model (PSM). In our example, we use AndroMDA tool to generate code. AndroMDA tool has a generic code generation engine which is a platform that hosts code modules (called cartridges) that do the actual code generation. Therefore, we can generate PIM to program code by some simple commands. Reference 2 shows the detail operation manual about code generation with AndroMDA tool.

6 Conclusions

In this paper, we have summarized the key elements of E-Commerce System development based on model-driven architecture, illustrating the method with the development of a flower store.

There are many challenges that face the development of E-Commerce system today. These include increasing complexity, dynamic conditions and changing requirements, solutions that are hard to use and shortened development cycles. Therefore, we should select the right models for others to use and make them accessible for easy use. Model-driven development can be used to meet the challenges.

References

[1] G. S. Francisco, V.G. Rafael, "An integrated approach for developing e-commerce applications", Expert Systems with Applications, vol.28, 2005, pp.223-235

[2] AndroMDA documentation, http://galaxy.andromda.org/

[3] L. Balmelli, D.Brown, M.Cantor, "Model-driven systems development", IBM Systems Journal, vol.45, 2006,

pp.569-585

[4] P. Chowdhary, K.Bhaskran, N.S.Caswell etc, "Model Driven Development for Business Performance Management", IBM Systems Journal, vol.45, 2006, pp.587-605.2

[5] Model Driven Architecture (MDA) FAQ, Object Management Group, http://www.omg.org/mda/faq_mda

[6] P.Amaya, C.Gonzalez, J.M.Murillo etc, "Towards a Subject-Oriented Model-Driven Framework", Electronic Notes in Theoretical Computer Science, vol.163, 2006, pp.31-44

[7] V.Kulkarni, S.Reddy, "Separation of Concerns in Model-Driven Development", IEEE Software, vol.20, 2003, pp.64-69

[8] A.Zarras, "Applying Model-Driven Architecture to achieve distribution transparencies", Information and Software Technology, vol.48, 2006, pp.498-516

[9] T.Elrad,R.E.Filman,A.Bader, "Aspect-oriented programming: introduction", Communications of the ACM ,vol.44 , 2001, pp. 29–32

[10] I. Ray, R. France, N. Li, G. Georg, "An aspect-based approach to modeling access control concerns", Information and Software Technology, vol.46, 2004, pp. 575–587

Research of Digital Forestry Grid Based on Web services*

Fan Li[1] Xu Zhang[1] Yan Chen[1] Guang Deng[1] Pinghui Yan[1] Yong Shan[2]

1 Institute of Forest Resource Information Technique, CAF, Beijing, P.R.China
Email: lifan@caf.ac.cn

2 Daniel B. Warnell School of Forestry & Natural Resources Georgia GA30602 U.S.A
Email: shanyongsy@hotmail.com

Abstract

This paper discusses the significance of building forestry grid, and illuminates the characteristic of grid technology, as well as the advantage of using this technique on forestry information. We expound the targets and functions in the design of Digital Forestry Grid, on the other hand, and bring forth upon the system structure of digital forestry grid a new form of establishment and a way of how to apply the advanced methods. In addition, we study the architecture of forestry grid based on SOA system, which constructs an application environment of digital forestry grid node.

Keywords: DFG, Grid, Web services

1 Introduction

Grid is a new type of calculation platform that is built on the base of current internet environment. It implements management across different organizations and domains in case of resource sharing and services [1]. Grid integrates distributed resources in effect; provides all kinds of means for resource sharing; improves the rate of resource utilization. It also matches the requests coming from the users and ability of providing resources reasonably. It selects appropriate resource services for the users' requests, so that resources in a large sense of scale can be shared accordingly. The most significant characteristic of grid is resource integration and sharing [2][3][4][5].

Digital Forestry Grid (namely DFG in short) which is financed by the application grid research item in high performance computer and core software field of the National Hi-Tech Research and Development Program of China (863 Program), is part of the China National GRID (CNGRID).

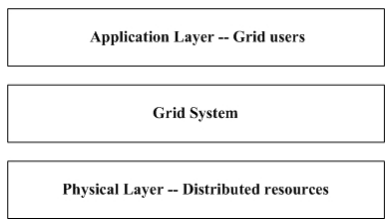


Figure 1 The three-tier architecture of the grid

2 Preparation of Manuscripts

The DFG is dedicated to countrywide forestry industry; it is the basic information establishment and the important basic platform of forestry information management system in the field of forestry

2.1 Basic character of DFG

DFG is dedicated to countrywide forestry industry; it is the basic information establishment and the important basic platform of forestry information management system in the field of forestry. The center of DFG is resource service and it is set up based on current grid standards and software, so it is a professional application and the architecture is

* This research was supported by national research and develop plan of China.

standardized. Meanwhile, it is the important portion of national grid which implements resource sharing between inside and outside of forestry together with interoperability.

With the rapid development of network technology and all sorts of skills and concepts' updated, as an application grid, it can not pursue the most fore part of technique, but set up based on an environment with relative stable skill and in the scope of grid development trend, all applications can only be available in this way. For this special requirement and characteristic of forestry, several versions of Globus Toolkit are studied first, and the bottom platform is developed accordingly. After a term of optimization and maintain, the bottom support environment of DFG has been built currently.

DFG applies the ways by packing all available services as web services, so that all services can be reached by users as web services. Web services are based on these standards like XML, they are very easy to develop and extend into a wide variety of environments, and they are also easy to deploy. We get rid of all of the problems of exchanging data between differing systems, and we don't need to worry about the detail of the processor, or how to convert the information we are sending into a neutral format because the Web service standards take care of that.

2.2 Design for DFG architecture

DFG is based on the technology of grid and make services as its center, and it implements generic data handling, space calculation, online analysis, transaction management and other such interoperability and service sharing implemented in common isomerous software system. In this way, DFG is able to provide numerical value calculation, data management (including property data, vector data, image data, etc), space analysis, data exchange, information release, network transmission and other public services.

In DFG, all services are workflows integrated and logical sequenced from different Web services in the system, and such workflows are implemented running in a control environment. The runtime control environment consists of Web services container, this container references the basic runtime environment of WSRF, it provides service-oriented basic calculation establishments for the whole DFG, and it will be deployed to each node where specific service interoperation is needed. The Web services container implements the grid core functionalities such as remote deployment, runtime management, service status spy, SOAP request handling and transfer. From the functionality point of view, the service container is an extensible Web services container; while from the format point of view, the Web services container is a set composed of basic runtime environment and some basic services which implement common system functionalities. All these services are deployed in the basic runtime environment previously, and special services can be provided after the container is started afterwards[7][8].

During the development of bottom support environment, other workflows and Web services are developed as well. For example, forestry data standard, forestry data organization and service, forestry space analysis service, forestry information system analysis service, plan for returning farmland to forest and analysis of forestry resource [9]. At the same time, validation related to support ability of low level grid software is done to confirm all organization modes for data resource in the grid, all service structures and interfaces for all types of services. For the moment, the services that have been developed are data resource service and data handling service. For data resource service, it includes all different levels of forestry resource data service, space basic data service, forestry topic space data service, field utilization service[10]. For data handling service, it includes data. The

architecture of DFG is displayed in Figure 2.

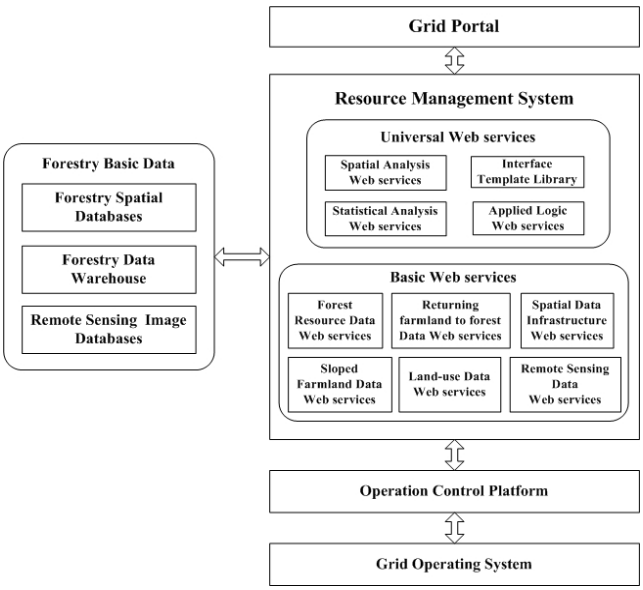


Figure 2 Design for DFG Architecture

3 Design of Resource Management in DFG

The core portion of DFG is forestry data resource and service management. In a grid, all resources are capsulated as services and delivered as services as well. The resource and services management consist of different level users management, data management and Web management[10][11].

The user management module realizes user management, role definition and level definition in all the virtual structures within the whole DFG.

The data management module realizes four features, data provision, data collection, data info delivery, and disposal of provided data [13].

Web service management module faces to users, who are web service founder, web service user and web service manager. The whole architecture of web service management module is composed by web service user, web service provider, authentication center, UDDI service center [14]. Normally, the web service user is the second developer on utilizing DFG. And web service provider develops web service based on web service standards. The structure is displayed in Figure 3.

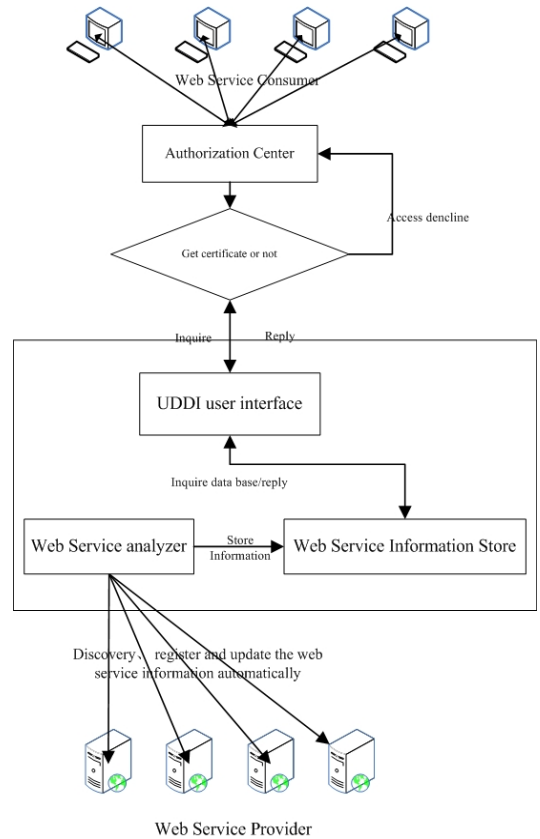


Figure 3 Architecture of web service management in DFG

4 Design of Safe Control Runtime Environment for DFG

The DFG control environment is to provide support environment for the whole grid platform, including workflow control module, security mechanism control module and semantic analysis module, and the main purpose of this control environment is to provide security and runtime mechanism guarantees to resource management in upper level. First, the security mechanism consists of digital certification in X.509 format, encryption and digital watermark, secure access log. Second, the workflow module consists of batch task scripts that can be recognized by the digital system, grid batch tasks that can be described, inquiry of task status, cancellation the running tasks. Final, the semantic

analysis module is composed of ways of constructing a forestry knowledge semantic system, building semantic analysis model, and semantic analysis [15].

The structure of interaction between grid control environment and resource management module is described below in Figure 4.

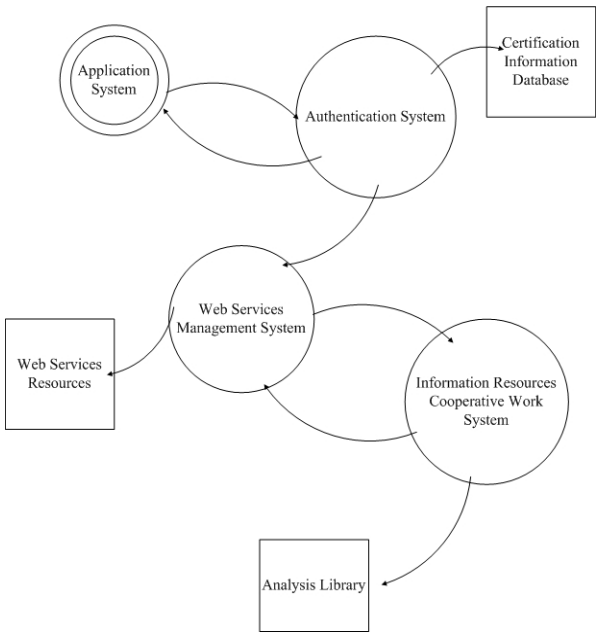


Figure 4 module interoperability operating of DFG environment

5 Design of DFG Operation System

The digital forestry grid operation system provides substrate support and services for the whole grid platform, including strategy management, router management, grid environment exception handling, etc. The feature of strategy management provides access control strategy management for virtual services, which maps access control relationship between user identity and virtual services he can access (reach service operation level). The router management provides the functionality of connection between routers, and it supports the feature that the address information of global routers that are connected can be updated term. For the exception handling module, it can position code details in both client and service exactly, and it can provide features of extensible exception definitions (registration), encapsulation and throwing out on server side, at the same time, all these extensible exceptions can be captured on client side by given exception handling mechanism[16].

6 Design of DFG Application

For the time being, the mainstream application of digital forestry grid is in the field of forestry resource management and restore-infield-to-forestry project. Digital forestry grid is able to fulfill different application targets' requirements coming from 4 levels of nation, province, city and country. So the digital forestry industry breaks through its original way by these applications and becomes necessary and mandatory operation system of project management and strategy, that it provides technical support of project programming, plan audit, project checking and other application services to different levels of management and strategy departments. The design of digital forestry application architecture is shown in Figure 5.

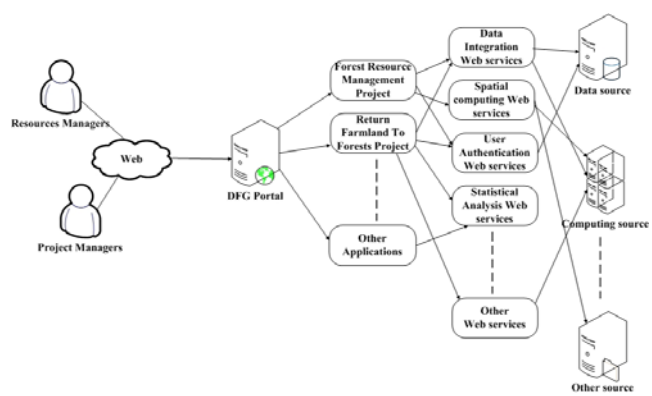


Figure 5 Architecture of DFG application

7 Conclusions

In this paper we study and analyze the advantages of using grid technique into forestry industry. Firstly, it's helpful to realize integration for forestry data resource that the forestry information level is increased. Secondly, it is good for all application developments based on web services by means of realizing large data

integration in forestry industry. Thirdly, multilevel and distribution mode for data in all levels can be realized, as well as the realization of dynamic cooperation among digital forestry industry system across different fields and all levels. Thus, maximize computation, data resource and technique resource sharing become true, and it provides information guarantee for the whole nation to make trustable and reliable decisions in the forestry industry and the resource environment construction.

References

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

[2] http://www-128.ibm.com/developerworks/cn/grid

[3] I. Foster, "The Grid: A New Infrastructure for 21st Century Science", Physics Today, Vol. 55, No.2, 2002, pp.42-47

[4] I. Foster, D. Gannon, "Open Grid Services Architecture Use Cases", http://www.gridforum.org/documents/GWD-I-E/GFD-I.029v2.pdf, 2004

[5] W. Allock, et al. Data Management and Transfer in High-Performance Computational Grid Environ-ments, Parallel Computing, 2001

[6] J. Yu, Y. Han, Service-Oriented Computing: Principles and Applications (in Chinese). Beijing: Tsinghua University Press, 2006

[7] X. Li, "Research on the Web Services Technology in Grid Computing", Computer Engineering & Science, Vol.27 (10), 2005, pp.107-110

[8] Z. Huang, "Resource Registry Meta-Service in Digital Forestry Grid", Scientia Silvae Sinicae, Vol.42, September2006, pp.45-50

[9] X. Zhang, F. Li and Y. Liu, "Digital Foresty Support Platform Developed by Application of Grid Technique and Formulation of Its Web Service Standard", Scientia Silvae Sinicae, Vol.42, September 2006, pp.14

[10] Z. Xu, W. Li, "Research on VEGA Grid Architecture", Journal of Computer Research and Development (in Chinese), Vol.39 (8), 2002, pp. 923-929

[11] I. Foster, C. Kesselman, S. Tuecke, "The Anatomy of the Grid: Enabling Scalable Virtual Organizations", Int'l.J.Supercomputer Applications,Vol.15 (3), 2001

[12] Q. Gao, "Grid: Resource Sharing Technologies for VO (Virtual Organizations)-Oriented", Computer Science (in Chinese), Vol.30 (1), 2003, pp. 1-6

[13] Y. Wang, N. Xiao, H. Ren and X. Lu, "Research on Key Technology in Data Grid", Journal of Computer Research and Development (in Chinese), Vol.39 (8), 2002, pp.943-947

[14] UDDI.org. http://www.uddi.org

[15] X. Zhang. et al, "Research and Implementation on Digital Forestry Platform", Scientia Silvae Sinicae, Vol.42, September12006, pp.37-40

[16] F. Li, X. Zhang, Y. Chen and Y. Liu, "Web Services Organizing and Presentation of the Digital Forestry Support Platform", Scientia Silvae Sinicae, Vol.42, September12006, pp.111-114

A New Network Management Architecture Based on Web and CORBA with Push Technology *

Xiaohong Wang Jingyang Wang Min Huang Huiyong Wang Liyan Zhang

Hebei University of Science and Technology, Shijiazhuang, Hebei, 050054, China
Email: ever211@163.com

Abstract

This paper analyzes the limitations of the traditional network management architecture based on Web and CORBA, which adopts the mechanism of pull information, that is, Only users apply for information on their own initiative, can it be updated. It is especially inefficient in processing trap information. The validity of the fault management will have a direct influence on the reliability of the whole network management. On the basis of the traditional network management architecture based on Web and CORBA, a new architecture is proposed which inosculates information push technology and CORBA notification service. This model changes the passive status of server and implements network management in an active style. It can push the trap information to manager actively through the TCP connection established when logging in. When the abnormity is occurred in network, trap information can be processed in time. As a result, the management efficiency of network is improved greatly.

Keywords: CORBA, Push Technology, Notification service, Network management system

1 Introduction

With the rapid development of IP and ATM technology, Network management has made a great progress. A variety of traditional networks begin to be combined with each other. The scale, complexity and heterogeneity of network have changed a lot. All of those make the network become more difficult to manage. The traditional network management system can hardly satisfy the higher requirement for network management. In order to make network system has higher performance and expansibility, it is developing to the Web based and distributed direction. Network management system based on Web and CORBA (Common Object Request Broker Architecture) will become inevitable trend for network management. It can effectively make up the shortage existed in centered network management architecture. However, As we know, Web based network management adopt HTTP protocol, which is used to pull information, that is, only users apply for information on their own initiative, can information be updated. For the trap information, this model becomes inefficient. The validity of the fault management will have a direct influence on the reliability of the whole network management. As a result, In order to help manager to find out trap information in time, network management architecture based on Web and CORBA with push technology is proposed [1].

2 Shortages of Traditional Web and CORBA Based Network Management Architecture

Web and CORBA based network management architecture can be seen from Figure 1, this model mainly consists of Java Applet, Web Server, CORBA Server and Agent.

Web Browser adopts HTTP protocol to communicate with Web Server. As we know, HTTP

* This paper is supported by CSC to Xiaohong Wang and the fund of Shijiazhuang (Grand NO. 07113431A)

protocol is used to pull information. Information can be refreshed only when Web Browser applies for. When trap happens to network, the inherent functions of HTTP protocol are hardly suitable to Real-time response to trap information and to push it to manage in time. That is one of the shortages existed in the Web based network management.

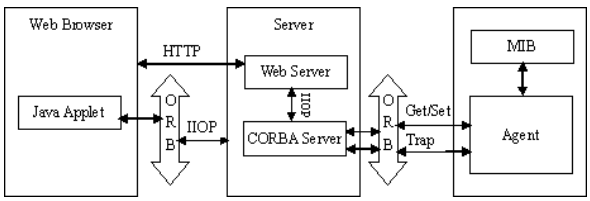


Figure 1 Web and CORBA based network management architecture

The traditional method to get fault information is that user connects with CORBA Server according to a given time interval and visits it one after another. The disadvantage of this model is that user needs to connect with server frequently if time interval is too short, that will affect efficiency of CORBA Server seriously, while the time interval is too long, user can not obtain Real-time info. Fault management plays an important role in network management. It serves not only to find out trap info. More important, it serves to push trap info to corresponding manager and have trap info processed in time to ensure network run normally[2][3].

3 Introduction of Correlative Technologies

3.1 Information push technology

An increasing expansion of network information resource makes information push technology come into being. Unlike the traditional pull information style, users only need to subscribe to information they require at their first usage of network. Hereafter, the subscription info will be pushed to users through Web automatically. Compared with the traditional pull technology, the uppermost difference between them is the status of server. Push technology make server push info to client actively, while pull technology need client to play an active role.

There are mainly two modes in the current push technology. They are automatic push and event-drive push. Automatic push is that client require sender to hand in the appointed info automatically according to promissory time. Event-drive push can also be named subscribe/ issue push, which is based on rule. The rule is established in advance by user, push manager and sender. Its main idea is that push manager judge the rule set in advance appears or not, if the rule appears, push manager will hand in the correlative info to user actively. The main differences between automatic push and event-drive push are as follows. Automatic push must have a specific referring time, while event-drive push is based on rule set in advance. Event-drive push has a middle layer for storing the relationship between info and rule. It decreases information load and increases the speed of information flow. We adopt the later one in this paper[4].

3.2 Notification service of CORBA

CORBA standard has not only defined CORBA structure, IDL criterion, but also defined some basic services, such as name service, event service, notification service and so on. Event service provides an incompact and asynchronous style for communication. Its aim is to establish a common distributed event model. It makes CORBA possess the characteristic of middle object primary, but it still exists some obvious limitations, such as, it can not ensure the reliability of communication, does not support QoS and does not provide structure event and filter mechanism. Thereupon, notification service is born. It is formed on the basis of event service as well as adding some new characteristics. It makes up shortages of event service efficiently and provides more guarantees for communication [5],[6].

A key element of the Notification Service is the Notification Channel (referred to here as the Channel) whose role is to propagate events from suppliers to consumers. Once an event has been delivered to the Channel the Channel takes responsibility for delivering

it to each subscribed consumer. This arrangement is shown in Figure 2.

The default behavior of the Notification Channel is to deliver every event it receives to every subscribed consumer. This is also the behavior of the Event Channel. However, the Notification Channel has the facility to filter events and thereby provide selective delivery. To use this facility, consumers specify which events they are interested in receiving by registering a filter expression with the Notification Channel. The Channel then applies the filter expression to each event to determine whether it should be delivered to that consumer or not. This and other features, such as Quality of Service parameters, can be used to tailor the behavior of the Channel. However, these facilities are only mentioned here to provide an overview of the capabilities of the Channel. The examples discussed in this paper do not require such advanced facilities, even though their use in these cases is conceivable.

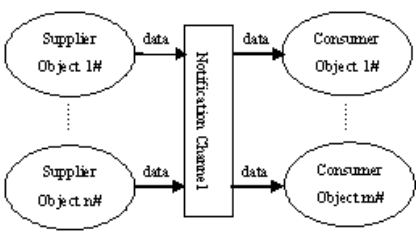


Figure 2 Notification Channel

A consequence of application elements using the Notification Channel is that they no longer communicate directly with each other but indirectly via the Channel. There are many benefits arising from this decoupling, including the following:

1) Supplier elements can deliver events at different rates to which consumer elements process them. And therefore, they can produce events at a different rate as well. In this respect, the Channel acts as a buffer, accommodating and leveling-out peaks in an application's processing activity.
2) The absence or unavailability of consumer elements does not prevent supplier elements from delivering events. In this respect, the channel allows an application to continue functioning even when parts of that application are unavailable.
3) A supplier can send an event to every consumer by creating a single event and delivering it to the Channel. In this capacity, the Channel acts as a broadcast medium for the application. If filtering is used in the Notification Channel, then the Channel acts as a multicast medium.
4) The identity of consumers is not needed by suppliers in order to reach them; only the identity of the Channel is needed by consumers and suppliers. Because of this, suppliers and/or consumers can be introduced to a system without requiring reconfiguration of existing suppliers or consumers in order to accommodate them. This has enormous benefit for large distributed applications [2].

Notification service inherits characteristic of event service, there are two modes in the notification service. They are push technology and pull technology. Also, it supports two kinds of objects namely producer and consumer. Furthermore, in order to make multi-producer can communicate with multi-consumer asynchronously, CORBA still introduces the mechanism of Notification channels. Producer and consumer can be connected through Notification Channel. They do not communicate with each other directly but obtaining a proxy from Notification channel to carry out communication. In the push technology, producer pushes event to consumer actively. While in the pull model, the process is reverse [7][8].

The important improvement on event service is that notification service provides filter mechanism for event. It is implemented by filter object. In notification service, filter object can be associated with Notification channel object, admin object and proxy object to implement different granularities on different levels.

In the paper, we adopt push technology of notification service. Compared with pull technology, push technology can avoid using buffer. When event appears, producer can inform all consumers connected with it. In this way, producer need not to visit consumer one after another, so it is more efficient.

4 Network Management Architecture Based on Web and CORBA with Push Technology

In order to make up shortages of traditional network management system based on Web and CORBA and improve management efficiency, this paper puts forward network management architecture based on Web and CORBA with push technology which inosculates information push technology and CORBA notification service, the model can be seen from Figure 3.

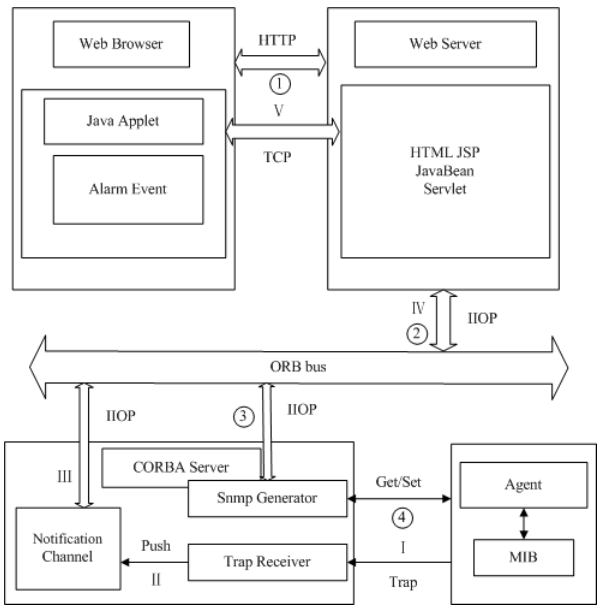


Figure 3 Network management architecture based on Web and CORBA with push technology

The structure mainly consists of Java Applet, Web Server, CORBA Server and Agent. Web Browser send management command to Web server by HTTP protocol, then Web server send operation information to CORBA server which will carry out corresponding operation. The processes of obtaining Real-time trap info are as follows. A TCP connection will be established with Web server when user logs on. Different user has different connection. When device sends out trap info, firstly, agent sends it to producer of notification service. Whereafter producer pushes the trap info to consumer through Notification channels. According to some conditions given by user, finally, the selected info will be pushed to each Web browser who subscribes to it. The trap will be showed and processed by Applet. The detailed explanations are as follows.

① Web Browser sends management command to Web server by HTTP protocol.

②③ Web server sends operation information to SNMP Generator of CORBA Sever by IIOP protocol, which translates operation information to the corresponding SNMP command.

④ SNMP Generator pushes the SNMP command to appointed Agent.

Ⅰ The trap info is sent to trap Receive of CORBA Server, which translates the trap info and then push it to Notification Channel.

Ⅱ Trap receive sends the trap info to producer of CORBA notification service.

Ⅲ and Ⅳ Producer pushes the trap info to consumer through Notification channel.

Ⅴ According to some conditions given by user, the selected info will be pushed to each Web browser through TCP connection.

5 Advantages of the Model

Compared with traditional Web/CORBA based network management architecture, information push technology can effectively make up shortages of HTTP protocol, it changes the passive status of server and implements network management in an active style, the model of pushing trap info in an active way can alleviate burden of Web server. More important, it can send trap info to manager in time. When there is something wrong with the network, trap info can be responded and processed in time.

Notification service can effectively decrease the burden of trap-gather server. In this paper, trap information is sent by notification service, users need not to connect with server directly. They communicate with server by registering to Notification channel. After server obtains trap info, server firstly sends it to Notification channel through producer of event. Then the trap info is sent to event consumer through Notification channel. In this way, server only needs to

transmit data to Notification channel, the burden of server decreased greatly. Furthermore, Notification channel provides more flexible mechanism for user and trap-gather server, server is transparent for users, we only need to accept trap info according to information format defined by interface. The transparent communication of them can solve problems of platform independent and interacting of them.

6 Conclusions

This paper puts forward a new network management architecture, which applies information push technology and notification service to traditional Web/CORBA based network management architecture. It can implement network management in an active way. However, the push style is based on TCP connection established between browser and Web server, user must keep the connection all the time. When the amount of users is large, the burden of Web server will increase. That will decrease the efficiency of Web server. The next study is to solve that problem.

References

[1] Jiang Ye,Zhao Yunhe, and Chai Zhi, "The Designation and Realization of the Communication Network Management System," Computer Engineering, vol.7 , 2006, pp. 54-58

[2] Object Management Group, "Notification Service Specification," Version1. 1, 2004 ,10

[3] Lu Xin and Peng Laixia, "Research and Application in Network Management of CORBA," Modern Electronic Technology, vol.10, 2006, pp. 47-49

[4] Wang Changlin, Yan Zhihui, and Li Xiaofei, "A Study of Web Chat System Based on HTTP Push Technology," JOURNAL OF NANCHANG UNIVERSITY, Vol.27, No.1, 2005, pp.94-98

[5] Zheng Xianrong and Chen Qiang, "Study and Design of Notification Service Integration Based on CORBA ComponentModel," Computer Application Research, vol.8, 2005., pp. 47-48

[6] Din Yan, Guo Changguo and Wang Huaimin, "Design and Implementation of Real-time Notification Service," Computer Engineering, vol.5, 2005, pp. 94-95

[7] Chen Jian and Li Maoqing, "The Application of CORBA Notification Service in Network Management System," Journal of SHANXI University of Science & Technology, vol.4, 2005, pp. 85-88

[8] Ding Yan, Dou Lei and Wang Huaimin, "Implementation of Event Filtering in OMG Notification Service," Computer Engineering & Science, vol.6, 2003, pp. 57-60

[9] Qin Ke and Yang Gelan. "CORBA Technology Introduction," SHANXI Science & Tecnology, vol.1, 2006, pp. 22-23

[10] Michi Henning and Steve Vinoski, "High Level Programming Based On C++ CORBA," Tsinghua press, 2000.11, pp.660-688

A Web Communication Model on the Basis of Anycast Technology in IPv6

Xiaonan Wang[1,2] Huanyan Qian[3]

1 Computer Science and Engineering School, Changshu Institute of Technology, Changshu, Jiangsu 215500, China

2 Computer Science and Technology School, Nanjing University of Science and Technology
Nanjing, Jiangsu, 210094, China
Email: wxn_2001@163.com

3 Computer Science and Technology School, Nanjing University of Science and Technology
Nanjing, Jiangsu 210094, China
Email: hyqian@mail.njust.edu.cn

Abstract

With the rapid development of web service the voice of improving web service quality and performance is increasingly vehement. This paper proposes a web communication model on the basis of anycast technology and this paper can provide for the clients the web service with better quality and shorter response time. In this paper the web communication model is deeply analyzed and discussed, and in IPv6 simulation by implementing WWW service its feasibility and validity are proved.

Keywords: IPv6, Anycast, Unicast, Router, Web

1 Introduction

Web service is a kind of application-integration mode which is built on the basis of open protocols, allows the existence of programmed elements on web site, and can calculate and deal with distributed information and provide some service[1]. Web service is a kind of independent and modularized application which can be described, released, located and called through network. This kind of coarse-grain integration mode is so suitable for Internet information service platform that in recent years web application service has rapidly expanded. Web service can both provide information and ensure the normal operation of various kinds of applications and service components so it exists in many fields as a primary form of offering network information service[2].

With web service's spreading into governments and enterprises and its comprehensively applying to various kinds of fields the voice of improving web service quality and performance is increasingly vehement.

2 Web Communication Model

This paper proposes a web communication model on the basis of anycast technology. The following words give a detailed discussion on and analysis of this model.

In this model each web service is identified by an anycast address, the web servers offering the same web service constitute an anycast group which is identified by the anycast address and each web server of an anycast group is called an anycast member of that group. Thus, a web-service-request message can be routed to the web server (anycast member) with the shortest distance[3]. Here, the values of distances between clients and web servers are calculated according to measurement units the current routing protocol specifies[4], which generally include hop count, server load capability and current available bandwidth, etc. But in fact the clients have no interest in those parameters and what they really care is the duration of service, namely, the time interval between

client's sending service request and receiving service response, which is called TRT(total response time)[5]. The shorter TRT is, better client considers anycast service to be. But in some situations, the value of TRT is proportional to the number of bytes transmitted. Therefore, the model selects RTT(round-trip time) as distance measurement unit. RTT[6] is an integrative parameter since it reflects and suggests both the status of server's current load and some attributes of the entire network and the established connections, including the current bandwidth and hop counts from client to anycast members(web servers), and so on. Figure 1 shows the WWW service performances in IPv6 simulation when hop count and RTT are used as measurement unit respectively.

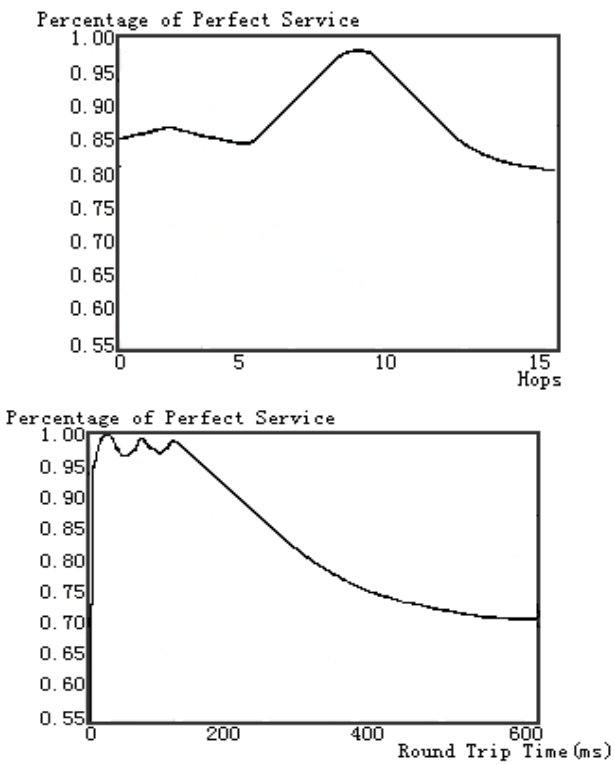


Figure 1 Anycast Service Performance

It can be inferred from that above figure that hop is not a good measurement unit, by comparison, RTT is a good choice for shorter RTT of web server is, the better service quality is.

According to the selected measurement unit the web communication model introduces the following architecture, as is shown in figure 2.

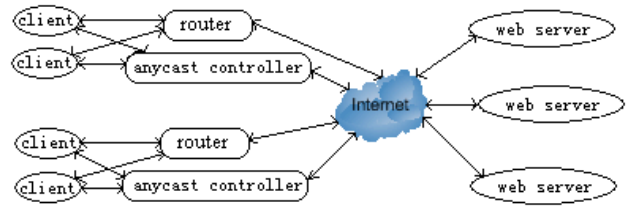


Figure 2 Anycast Architecture

In the above architecture of the web communica tion model anycast controller is added into each local network which is used to maintain the information on the anycast members (web servers) located in the local and neighbor network. In addition, in order to differentiate between Anycast service type and other service types the suffix of an anycast domain name is set to Any, for example, www.njust.edu.any. The following words gives a detailed description of the process of a client's requesting for web service: (1)a client requests DNS[7] server to parse the domain name and acquires the anycast service type by checking the suffix of domain name; (2)after DNS server receives the request sent by the client it first checks the suffix of domain name to learn the type of the domain name, if it is anycast domain name DNS server returns anycast address identifying that web service[8] to the client, or returns unicast address; (3)if the returned address is anycast address the client sends to anycast controller a message which includes the received anycast address acquired in (2) and requests anycast controller to parse that anycast address into the corresponding unicast address, or the client deals with that address in the normal way ; (4) after receiving the message sent by client anycast controller first searches the local database for the received anycast address and locates the web server with the shortest RTT, and then returns that server's unicast address to the client, or if no corresponding entry is found in the database anycast controller sends to its neighbor anycast controller a query message for the information on that anycast address, and then according to the response messages returned by the neighbor anycast controllers selects one web server with the shortest RTT and then returns its unicast address to the client and updates its local database with those response messages; (5) client utilizes the received unicast address to directly establish

a connection with the web server; (6) after the web service is finished client calculates the RTT, TRT and BW (bandwidth) of this web service and encapsulates these parameters into a message and sends it to local anycast controller; (7) anycast controller updates its local database with these received data.

3 Implementation of Communication Model

This model aims to accomplish web service through anycast technology and it is compatible with existing network application and protocols. The following words give a detailed description of and discussion on implementation of this model.

3.1 DNS server

In this model a function of parsing anycast domain name is added into DNS application. Due to the layered structure of domain name this model sets the suffix of domain name to "any", for example, www.njust.edu.any, to differentiate between anycast domain name and other domain names. In this model DNS server can parse anycast domain name and return the corresponding ancast address to the client. The detailed process of DNS server's parsing anycast address is the same as the one of its parsing other kinds of domain names so this paper gives no redundant descriptions

3.2 Client

In this model, into client application are added three new functions: (1) client can differentiate between anycast web service and other kinds of services by checking the domain name; (2) if it is anycast web service client can send to anycast controller a message which includes the anycast address and requests anycast controller to parse that anycast address into the corresponding unicast address, and in the meanwhile can receive the unicast address of web server with the shortest RTT returned by anycast controller by which client can directly establish a connection with that web server; (3) after the service is finished client can calculate the RTT, TRT and BW(bandwidth) of this web service and encapsulate these parameters into a message and send it to the local anycast controller, if that web server is unreachable the client can send an error message to the local anycast controller and then repeat (2).

Due to the layered structure of domain this model sets the suffix of anycast domain name to Any, thus, by checking the domain name client can differentiate between anycast service and other kinds of services. In addition, in this model client still needs to know the unicast addresses of both the local router and the local anycast controller, and extracts the unicast address of the web server with the shortest RTT from the message returned by anycast controllers. At last, client can count the value of RTT according to the time stamp of the transmitted data packet, the value of TRT and the total number of transmitted bytes, and calculate bandwidth of the current network, and then encapsulate these parameters into a message and send it to the local anycast controller. If that web server is unreachable the client can send an error message to the local anycast controller and then repeat sending a parsing-anycast-address request packet to the local anycat controller in order to avoid being forced to abort web service by reason of web server's being off-line or breakdown and to most ensure the good quality of web service.

3.3 Anycast controller

In this model into anycast controller are added four functions: (1) anycast controller can receive a parsing-anycast-address request message from the client; (2) according to the anycast address encapsulated in the request message anycast controller can search out the web server with the shortest RTT and return its unicast address to client; (3) anycast controller can receive a feedback packet which records the relevant information on the last anycast web service or an error message returned by client and in term of certain algorithms deal with the received packet to update the local database; (4)

anycast controller can receive a joining request from a web server and authenticate that its identify to check if it is permitted to become an anycast member, then update the local database with the information on the new anycast member.

In this model anycast controller maintains a database to record the parameters on web servers and its data structure is as follows:

Table 1 Parameters of web server

Anycast Address	Unicast Address	RTT	TRT	BW
Anycast Addri	Unicast Addri1	RTTi1	TRTi1	BWi1
	Unicast Addri2	RTTi2	TRTi2	BWi2
	…			
	Unicast Addrin	RTTin	TRTin	BWin
…	…	…	…	…

After anycast controller receives a parsing-anycast-address request packet sent by a client according to the anycast address encapsulated in the received packet it first searches the local database for the received anycast address and locates the web server with the shortest RTT, and then returns its unicast address to the client. If there are more than two entries whose RTT values are equal the values of TRT and BW in these entries are further compared to obtain the best web server. But if no corresponding entry is found in the database anycast controller sends to its neighbor anycast controller a query message for the information on that anycast address.

In this model a query message contains two fields: one is path attribute which records all the networks a query message crosses in order to prevent the query message from looping, and the other is TTL[9] which is used to control the routing scope of a quest message. The value of TTL is initialized to the maximum number of network hops a query message can traverse and gets decreased by 1 with each hop. The entire process of anycast controller's querying for the information on anycast address is described as follows: (1) anycast controller encapsulates the anycast address into a query message and sends it to its neighbor anycast controller in a multicast way, and then start up a timer; (2) after a neighbor anycast controller receives a query message it first searches its local database for the entries on that anycast address. If some relevant entries are found out neighbor anycast controller sends the contents of these entries to the source anycast controller, or anycast controller decreases TTL by 1 and appends its unicast address into path attribute, and then checks if the value of TTL is equal to zero and routing path forms loop. If neither, the anycast controller again transmits that query message to its neighbor controllers in a multicast way; (3) After sending a quest message, the source anycast controller starts up a timer and waits for the response messages. When the timer expires the anycast controller checks all the received response messages to select one web server with the shortest RTT and then sends its unicast address to client, in the meanwhile updates its local database with these response messages. In some extreme situations if no response messages are returned the anycast controller will return an error message to the client.

In general it is efficient to set TTL to 2 or 3.

After web service is finished or failed the client must send a resulting message to anycast controller reporting the relevant parameters on this web service. If the message is an error one the anycast controller will mark the web server the error message identifies as an unavailable one, or anycast controller updates its local database with the relevant parameters in the message.

The detail of updating database is described as follows:

$$RTT=\alpha RTT_{old}+(1-\alpha)RTT_{new}$$

Here, RTT_{old} represents the current RTT value in the database; RTT_{new} the value of RTT in the last web service returned by the client; α is a constant whose value depends on the stability of the current network. In our experimental environment, this constant is set to the same value as TRT.

$$TRT=\alpha TRT_{old}+(1-\alpha)TRT_{new}$$

In the above formula, TRT_{old} represents the current TRT value in the database; TRT_{new} the value of TRT in the last web service returned by the client; α is a constant whose value depends on the stability of the current network. In the experimental environment, this constant is set to 0.25.

The total time of one service includes duration of establishing connection, transmitting data and closing connection so the value of bandwidth in this web service can be calculated according to the following formula:

$BW = S/(TRT - 3.5 \times RTT)$;

Here, TRT represents the total duration from client's sending request message to receiving response message; RTT round trip time; S the total number of bytes transmitted.

In this model all the anycast controllers are the members of a multicast group. When an anycast controller receives a joining-anycast-group request message from a web server it first authenticates its identification and then adds a new entry on it into the local database where the value of each field is set to the initialization, namely, optimal value, and at last sends a notification message in multicast way to all the anycast controllers which will update their local database with the information in the notification message. In addition, the information exchange between anycast controllers may be achieved in the multicast way.

3.4 web server

If a web server wants to become a member of an anycast group it must know the unicast address of local anycast controller and the anycast address of that anycast group. The process of a web server's requesting to join an anycact group is described as follows: (1) a web server sends to the local anycast controller a joining-anycast-group request message which includes the anycast address of anycast group it wants to join; (2) after the local anycast controller receives a joining-anycast-group request message it first authenticates the web server's identification and then adds a new entry on it into the local database where the value of each field is set to the initialization, namely, optimal value; (3) the local anycast controller sends a notification message in multicast way to all the anycast controllers which will update their local database with the information in the notification message.

The above process can be achieved by creating new types of BGP and IGMP messages[10]. To avoid malicious attacks, the information interactions between web servers and anycast controllers should take some security measures.

4 Performance

From the client perspective the shorter TRT is, the better the service quality is. So the performance analysis refers to the comparison between web service performed in this model and the one fulfilled in normal way and it is accomplished by comparing the values of TRT of performing the same WWW service in the above two ways, as is shown in the following formula and figure.

$$R = TRT_{Normalr} / TRT$$

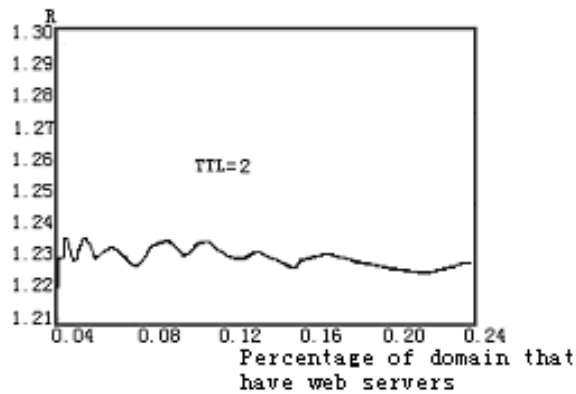


Figure 3 Performance Analysis

Here, R represents the ratio of the TRT value of performing the WWW service in this model to the one of performing the same service in the normal way; TRTNormal is the TRT value of performing WWW service in normal way; TRT is the TRT value of fulfilling the same service in this model.

From the above figures, it can be inferred that the value of R trends to 1.235. This experimental result indicates the entire performance of web service performed in this model is better than the ones in the normal way.

This model accomplishes the web service with better quality and shorter response time by anycast technology which can transmit the web service request sent by client to the best web server to transact. In

addition, this model is still a scalable and low-consumption one and evenly disperses the requests from the clients over the web servers of one anycast group in term of their current statuses, and also offers secure mechanism to avoid client's being forced to abort web service by reason of web servers' temporarily being off-line or breakdown, all of which most ensures the good quality of anycast service.

This model is compatible with existing network applications and protocols. In this model information interaction between anycast controllers and clients is only fulfilled in the local network so it has no influence on the performance of backbone networks. Although the query messages initiated by anycast controllers may consume some network resources it hardly affects the performance of backbone networks since they are only utilized in some extreme situations.

5 Conclusion

Any cast is a new characteristic of IPV6 and can support various kinds of services. This paper proposes a web communication model on the basis of anycast technology and this model can provide for the clients the web service with better quality and shorter response time. In IPv6 simulation, the experimental result indicates the entire performance of web service performed in this model is better than the ones in the normal way but as a new kind of web communication model in this model there may exist many problems and they need further study and analysis.

References

[1] B. Wu and J. Wu, k-Anycast Routing Schemes for Mobile Ad Hoc Networks, Proc. of IEEE IPDPS, April 2006

[2] R. Hinden and S. Deering, IP version 6 addressing architecture, RFC3513, April 2003

[3] S. Weber and L. Cheng, A Survey of Anycast in IPv6 Networks. IEEE Communications Magazine, pp. 127-132, 2004

[4] S. Doi, S. Ata, H. Kitamura, M. Murata, and H. Miyahara, Protocol design for anycast communication in IPv6 network, in Proceedings of 2003 IEEE Paci.c Rim Conference on Communications, Computers and Signal Processing (PACRIM'03), (Victoria), pp. 470-473, Aug. 2003

[5] J. Wang, Y. Zheng, C. Leung, and W. Jia, A-DSR: A DSR-Based Anycast Protocol for IPv6 Flow in Mobile Ad Hoc Networks. Proc. of IEEE Vehicular Technology Conference: Symposium on Data Base Management in Wireless Network Environments, 2003

[6] S. Doi, S. Ata, H. Kitamura, and M. Murata, IPv6 Anycast for Simple and Effective Communications, IEEE Communications magazine, vol. 42, no. 5, pp. 163-171, May 2004

[7] S Deering, R Hinden. Internet Protocol Version 6 (Ipv6) specification, RFC 2460,1998

[8] Jurrichiro itojun Hagino, K Ettikan. An analysis of Ipv6 anycast Internet Draft. Internet Engineering Task Force, 2001

[9] D Katabi, J Wroclawski. A framework for scalable global IP-Anycast(GIA). In: Proc of SIGCOMM, New York: ACM Press,2000, 3-15

[10] Miguel Castro, Peter Druschel, Anne-Marie Kermarrec, Antomy rowstron. Scalable application-level anycast for highly dynamic groups, Prentice Hall, 2003

A Study and Design of Integrated Information Platform Based on SOA

Wanping Wu

Department of Computer Science, Wuhan University of Science and Engineering, Wuhan, Hubei, 430073, China
E-mail: wwp@wuse.edu.cn

Abstract

This paper presents a method to implement an integrated information platform that eliminates the hassles of information isolation. Build on the top of campus network, the platform maximizes the information sharing, unifies user authentication, centralizes databases management and integrates the workflows of applications. Our method utilizes the EAI mid-ware and the Web Services to build the SOA. A central database is designed, which connects all the sub systems such as Student Management Information System (MIS), Faculty/Staff MIS, Library Catalog System, and Service MIS, while preserving the independence and privacy of all these sub-systems.

Keyword: SOA, Web Service, Integrated Platform, EAI, Information Isolation

1 Introduction

Currently, the information platforms in higher education institutes have been set up to a large degree. However, most of their constructions lack integral design as well common interfaces between different modules. Since no standards are available or has been followed when building the application modules of the platform, it is common that many sub-division systems or modules are isolated and hard to interoperate even within campuses, which motivates us to design and implement an integral information platform prototypes for higher education as well as enterprises.

We present an approach to build an integral information platform using the Service-Oriented Architecture(SOA) based on the Web services. Providing a way to effectively integrate and fuse application sub-systems, our platform is highly re-usable and scalable in account management, computing resource management and access control.

2 Integrated Platform

Information technology has been rapidly applied in the higher education institutes of China during 1990s. A common phenomenon is that once a division or department adopts one software system and improves work efficiency, other divisions follow up to build or adopt systems for their own divisions. Due to the time lag, the difference in application domains and special-purpose design, each software system works individually without scalable interconnection with other systems. Considering the current whole campus information technology as a single system, we have felt that it is hard to share information, hard to integrate and hard to manage these individual sub-systems. The following phenomena are widely spread:

(1) Inconsistent coding specification. Different division use different coding methods.

(2) A variety of database formats. Some early systems use FoxBASE and some use ACESS, while contemporary systems use SQL or Oracle.

(3) Different building tools. E.g., some systems were built using J2EE while some others used simple Excel.

(4) Poor synchronization. E.g., the records of faculty, staff and student are redundantly saved in different systems. It is costly to update these records simultaneously.

(5) Security, privacy and non-technical issues are arising.

To solve above problems, it is necessary to integrate all the information resources and logically organize them using all methodologies and tools available. The benefits of integrations include:

(1) Reduce investment in capital equipment and cost in maintenance. For example, unifying the management of different modules such as system administration, internet monitoring, security enforcement, database management and customer services can significantly enhance work efficiency and reduce costs. Every group can concentrate in their own sub-systems without hassles in interconnection with other sub-systems.

(2) Security management. When data center management is available, only the data center is accessible virtually, while all sub-systems are working independently. Therefore, only the data center needs the security control personnel. Security is safeguarded within a smaller group and thus more reliable.

(3) Effective management of data communication and data storage. Each application system is a part of the entire integral system. How to design the database structure in an effective way convenient for communication is one important task of the data center management. Data backup locally and remotely is one way to guarantee data security. The backup can be done quickly in the data center[1].

Our methodology includes four major steps. First, we carry out data integration, which is the fundamental of the whole integration. Without data integration, it is not possible to realize data sharing, let alone the application integration and account access integration. Data of sub-systems, such as the student information management system, the education administration management system, library catalog system and the dormitory management system, must be analyzed thothroughly to build an overall data blueprint. The intra- and inter- relations within theses sub-systems must be understood clearly to enable effective communication and sharing.

The second is application integration which removes the information islands and provides customers with integrated and personalized service.

The third is content integration that establishes a comprehensive information portal and centralizes control of accounts for integrated access entrance.

Finally, process integration, which is the advanced stage of information integration. It uses workflow, messaging and collaborative techniques, to achieve inter-system integration, enabling different departments to collaborate in a unified office online.

3 Integrated Information Platform based on SOA

SOA is a component model, which links different application modules (called services) together through fine-tuned interfaces. Essentially, SOA is a coarse-grained, loosely coupled service structure. Services communicate through simple yet accurate channels, independent of the underlying programming interface and communication model. SOA organizes three roles: service providers, service consumers and service registration center. The service provider is responsible for the concrete realization of services, the release of its services to the registration center and executing services requests. Service consumer is the initiator of the service requests. It queries services in the registration center and then bonds/calls the provider according to service description. Registration center provides registration services to provide service classification and search functions to better serve consumers[3]. SOA architecture model is as shown in Figure 1.

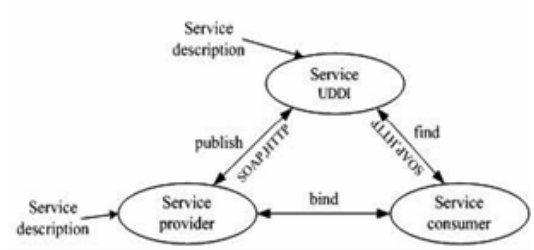


Figure 1 The architecture of SOA

The integration characteristics:

(1) SOA provides an ideal framework for service re-use and supports the assembly of service modules during workflow. Without altering the structure and

functions of the underlying modules, SOA packages them in the Web Services.

(2) SOA focuses on business process integration and stresses the need to have business as the center for the application of the standard service. A high-level service bus connects data services, business process management services, and other functions. It can change as the needs of business evolve, extending the EAI solutions from application integration to service integration.

(3) Integration granularity and service granularity are related. Enterprises need to integrate internal business processes through the flexible control services portfolio. Thus fine granular interface is a more suitable choice[2].

The structure of Web-based services is the best way to achieve the SOA. Web Services is a framework for achieving SOA technology, built on open standards and independent platform with the protocols of the distribution. XML is used for data description and exchange between service providers and service users. SOAP provides communication protocols, WSDL defines interfaces and UDDI registers Web services. These features enable Web Services to be the best way to achieve SOA.

Our integration methodology uses hierarchical model. First, we define the coarse-grained model of four major service levels, including the application link layer, application integration layer, business presentation layer and user interaction layer, as shown in Figure 2. Second, each level is composed of granular services. General functions are described below:

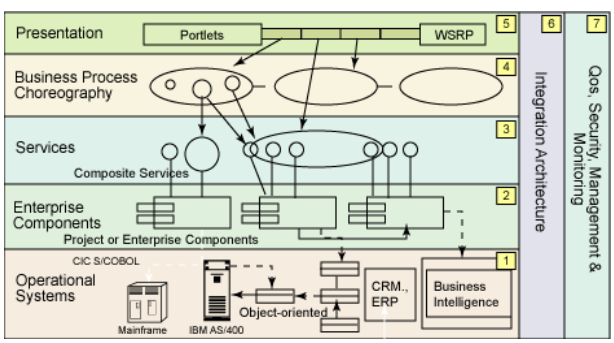


Figure 2 The hierarchical model

Application Link Layer: the bottom layer in the EAI architecture. It solves the system interface and data connectivity issues between Application Integration Server and the sub-systems. This layer also includes service adapter and message agent for the sake of communication. The adapter packages logic units, registers their services and publishes them. The message agent plays a request/response role in the middle of the client and system server[5].

Application integration layer: it is aimed to solve the data conversion problem of integral systems through the establishment of a unified data model to achieve information exchange of different information systems. At the same time, it provides various types of application services for the storage and directory for search. Message management completes the function of message queue and routing.

Business presentation layer: It connects different application systems together, coordinates their work and provides the workflow management functions, including event processing, process design, monitoring and planning.

User interaction layers: the user interface is to provide a unified information service entrance, through internal and external information which are relatively decentralized, thus ensuring the user to access customizing resources through the integral portal.

4 Design of Integrated Information Platform for Higher Education

Using the platform for higher education as example, we describe our integration approach based on the SOA. Our approach is also applicable to enterprises. The framework is built using J2EE and the central database system is built on top of Oracle 10g. The management tools will secure user accounts and authenticate unique account access. The platform is essentially distributed, consisting of the application service management center and service-oriented servers. The access to databases is via mid-wares. A central database is designed, which connects all the sub systems such as Student Management Information System (MIS), Faculty/Staff MIS, Library Catalog System, and Service MIS, while preserving the independence and privacy of all these sub-systems.

5 Conclusion

SOA is service-oriented design pattern for software integration. It is software and platform independent. Based the SOA, the presented integrated platform is aimed to solve the information isolation problem and interconnects existing systems and new systems such as web services and management modules. Implementing the SOS technology, we realized the integration of all systems from different divisions of a university and significantly enhanced resource sharing and software re-use. The integration approach has broader impacts and is also applicable to medium to large enterprises.

Reference

[1] MarkColan. Service-Oriented Architecture Expands the Vision of Web Service [EB/OL] . http://www.128.ibm.com/ developerworks/webservices/library/ws-soaintro.html, 2004

[2] Richard Monson-Haefel. J2EE Web Services-The Ultimate Guide[M]. Tsinghua University Press. 2005,4

[3] WU Xiao, LV Shuang , MA Xinqiang. Research on enterprise application integration scheme based on SOA[J]. Information Technology. 2007,4

[4] Sandhu R, Coyne E, Feinstein H, et al. Role - based Access Control Models[J]. In IEEE Computer, 1996, 29 (2) :38- 47

[5] LI Wei. The Next Generation Software Architecture--SOA[DB/OL]. http://dev2dev.bea.com.cn

[6] Hammer K. Web Services and Enterprise Integration [J]. EAI Journal. 2001.3

[7] BROWN A, JOHNSTON S, KELLY K. Using Service-Oriented Architecture and Component - Based Development to Build Web Service Applications [DB/OL]. http://my. donews.com/ zoblog/ 2006/04/

[8] XU Weibing. Research of Power System Information Integration Based on SOA[J]. China Instrumentation, 2007.6

[9] Meng Haitao, Yin Xu. The study of digital campus network based on SOA[J]. CHINA SCIENCE AND TECHNOLOGY INFORMATION. 2007,16

[10] Cai Tingyou. Research and Implementation of Enterprise Application Integration Based on SOA[J]. Microcomputer Information. 2007,5

[11] Dr HaoHe. What is service-oriented architecture [EB/OL]. http://webservices.xml.com/pub/a/ws /2003 /09/30/soa.html. 2003,9

[12] PapazoglouM P,Georgakopoulos D.Service-oriented computing[J]. Communication of ACM,2003,46(10):25-28

Tele-robotic over Internet Based on Multi-agents System

Adil Sayouti Fatima Qrichi Aniba Hicham Medromi Mustapha Radoui

Equipe Architecture des Systèmes, ENSEM BP 8118, Oasis, Casablanca, Morocco
Email: sayouti@gmail.com, qrichi_f@yahoo.fr, hmedromi@menara.ma, m.radoui@ensem-uh2c.ac.ma

Abstract

For Internet-based telerobotic systems (Internet robots), the most challenging and distinct difficulties are associated with Internet transmission delays, delay jitter and not-guaranteed bandwidth availability, which might lead to dramatic performance degradation or even instability. The solutions, proposed to face the limitations of the communication channel, are founded on the autonomy and the intelligence based on multi agents systems granted to the robot in order to interact with its environment and to collaborate with the remote user. In a first part of this paper, we present the main interests of such a remote control and we describe some existing applications. In a second part, we describe and compare our approach to the classic one. In the third part, we present our control architecture. An illustration of our approach is given in an application of control of an autonomous mobile robot. I am anxious to thank Maroc Telecom partner of our project.

Keywords: Tele-robotic, Internet, Multi-agents Systems, Control Architecture

1 Introduction

The control of current robotic systems in manufacturing industry and the service sector has remained separate and independent. In other words, these robotic systems are isolated from one another by different environments and have no effective way to communicate. This has made the current robotic systems expensive and requiring a long developing cycle, which has in turn seriously hampered the day-to-day deployment of robot technology. Therefore it is crucial to develop an integrated network environment for robotic systems based on today's Internet technology. With the rapid growth of the Internet, more and more intelligent devices or systems have been embedded into it for service, security and entertainment, including distributed computer systems, surveillance cameras, telescopes, manipulators and mobile robots. Although the notion of Internet robotics or web-based robotics is relatively new and still in its infancy, it has captured the huge interest of many researchers worldwide. Except for operating in hazardous environments that are traditional telerobotic areas, Internet robotics has opened up a completely new range of real-world applications, for example in the following fields [1]:

- Tele-teaching: a lot of universities are using robots to teach the basics of electrical engineering. The profitability of these robots is of course really poor because they are only used a few weeks a year. Why not developing common centers where students may have access to real robots without being close to them? One of the problems of e-learning is to make practical experiments. Why not using Internet technologies to let distant students to manipulate real systems?

- Tele-maintenance: when a company is shipping systems all over the world, it needs to send technicians when one of its systems has some failures. With Internet technologies, it is now possible to make some remote diagnostics, to solve and repair some problems, to prepare the right equipment to send etc.

- Tele-expertise: some specific operations on robotic systems can only be made by expert. In a close future, it will become possible for experts to operate from their office a machine located somewhere in the World, just using classic web technologies.

- Tele-production: the remote access possibilities and taking control will make work easier for remote users and will allow the performances of more tasks in the future.

The use of Internet will reduce the costs of these activities. The increase of Internet abilities in term of speed and bandwidth in the future, let us also think that the quality of the remote control and the comfort of the user will also increase. But, when developing such applications, we have to think that these activities rely all the time on an unpredictable network and that we have to build them taking into account this parameter.

During the Nineties, several projects appeared of robotic systems control, using Internet as communication network [2] with various objectives: the Mercury project to prove the feasibility by Goldberg and al. [3], the Australian telerobot for the interaction with the user by Taylor and al. [4], Rhino by Burgard and al.[5], Xavier by Simmons [6], Puma-Paint by Stein [7], mobile robotics KhepOnTheWeb by Saucy and al. [8], increased reality by Otmane Ariti [9], etc.

From the study of these experiments on Internet [10], a common frame can be described about the operational aspects of a remote control application (figure 1). The user, through his Internet navigator, addresses a request to a Web server (step 1) and downloads an application on his work station such as for example an applet Java (step 2). A connection is then established towards the server in charge of the management of the robot to control (step 3). The user is then able to take the remote control of it. In parallel to step 3, other connections are also established towards multi-media servers broadcasting signals (video, sound) of the system to be controlled.

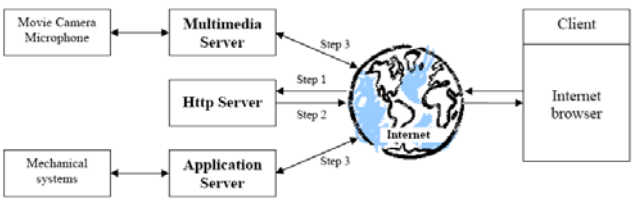


Figure 1 System Architecture Used in Tele-robotic.

This paper is presented as follows: in section 2, we describe the usual agent's model and we show its limits in term of interaction management. Then we will present the interaction oriented approach. In section 4, we describe our control architecture based on multi agents systems. ln section 5, we present an application of control of an autonomous mobile robot as an illustrative example. Finally, some conclusions are presented in section 6.

2 Classical Approach

Interactions between agents within a multi-agents System (MAS) are largely recognized like one of the essential mechanisms to ensure collaboration by the distribution of tasks, the resource sharing, the coordination of actions and the resolution of conflicts. It is thus interesting to be able to manage these interactions through all stages of MAS design and execution [11].

Although many models of agents were proposed [12], an agent is mainly made up of:

- Roles: what an agent must do.

- Competences: they can be internal relating to the way in which the agent ensures its roles, or external (social) concerning its relationship with others to perform its task.

- Interactions: they allow the agent to communicate with its environment and/or with the other agents.

The internal architecture of an agent is illustrated in Figure 2. One of the principal properties of the agent in a multi-agents system is its capacity to be in interaction with the others. These interactions are generally defined like any form of action carried out within a society of agents which modifies the behavior of another agent. They give to the agents the possibility of taking part in the satisfaction of a global goal.

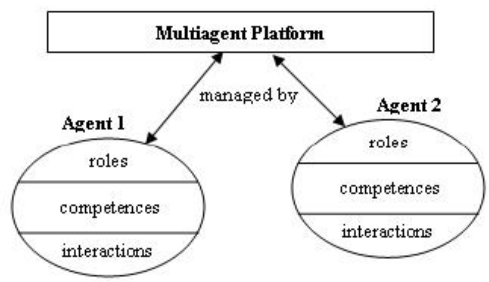


Figure 2 Agent's internal architecture

This participation allows the system to evolve to one of its objectives and to have an intelligent behavior.

These interactions are mainly based on the communication. A communication can be defined as a local action of an agent towards other agents. The questions covered by a communication model can be summarized by the following interrogation: who communicates what, to which, when, why and how? Several works were interested to answer this interrogation. These works can be classified in two groups: works on the inter-agents communication languages [13][14][15] and works on the interaction protocols [16][17][18].

3 Interaction Oriented Approach

One idea which has been proved reliable in the software components field is the separation between the interactions and the components [19]. This manner to define the interactions opened the way to new possibilities of abstraction and expression of the interactions. Thus, an interaction is not specific any more to only one component but can be reified into an entity as well as the other components of the system. This entity is viewed as a shared resource that the components can consult, use or instanciate. This approach separates the interaction-related behaviors and functionalities from the algorithmic parts of the agents (see Figure 3). From now on, interactions will be managed by an entity called “Interaction Manager”.

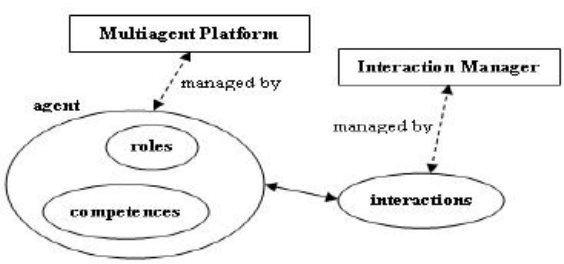


Figure 3 Agent’s new architecture

Interactions are the basic elements of our approach with the following properties:

- The interactions are defined in a formal and declarative way independently of the communicating agents. This formal description will be represented as classes in the applicative language called “interaction patterns”.

- An interaction pattern is an abstraction of the interaction concept. It is defined on the agent class level and can be instanciated. It is the counterpart to classes in object oriented languages whereas interactions instances are the counterpart to objects. In other words, the interactions are objects created by instanciating the interaction patterns.

- An interaction pattern is implementation-independent and an interaction instance can connect heterogeneous agents across different platforms. For instance, an interaction can connect a JADE agent with a MadKit agent.

- The instanciation of an interaction pattern is independent of the instanciation of the interacting agents’ classes. An interaction can be dynamically created between agents that need to interact and, it is destroyed when they don't need to interact any more.

4 Proposed Control Architecture

Today’s internet technology provides a convenient way for us to develop an integrated network environment for the diversified applications of different robotic systems. To be

successful in real-world applications, Internet-based robots require a high degree of autonomy and local intelligence to deal with the restricted bandwidth and arbitrary transmission delay of the Internet.

When turning a robot on, the problem of its autonomy is quickly addressed. However, several types of autonomies can be considered: energetic autonomy, the behaviour autonomy or smart autonomy. The designer has to choose the way he will give autonomy to his robot. He has mainly two orientations: “reactive” capacities or “deliberative” capacities [20]. These two capacities are complementary to let a robot perform a task autonomously. The designer must built a coherent assembly of various functions achieving these capacities. This is the role of the control architecture of the robot. To design an autonomous robot implies to design a

control architecture, with its elements, its definitions and/or its rules.

From the study of the different control architectures, we propose a hybrid control architecture, called EAAS for EAS Architecture for Autonomous system [21], including a deliberative part (Actions Selection Agent) and a reactive part. It is made up of two parts, each using distinct method to solve problems (Figure 4). The deliberative part which uses methods of artificial intelligence contains a path planner, a navigator and a pilot. The reactive part is based on direct link between the sensors (Perception Agent) and the effectors (Action Agent).

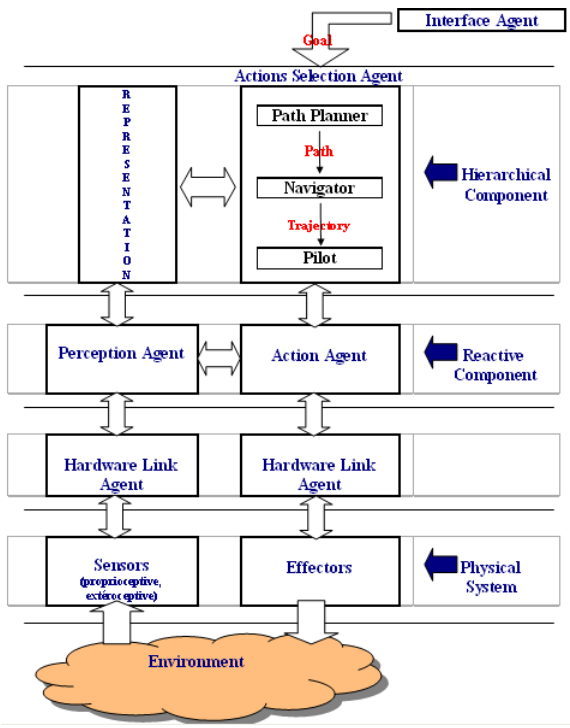


Figure 4 EAAS Architecture

Fundamental capacities of our architecture encompass autonomy, intelligence, modularity, encapsulation, scability and parallel execution. The communication between agents is realized by messages. Object oriented language is therefore absolutely suited for programming agents (we chose java). We use threads to obtain parallelism (each agent is represented by a thread in the overall process).

EAAS architecture consists in five agents: interface agent, actions selection agent, perception agent, action agent and hardware link agent. The interface agent is the high level of our control architecture. It must generate a succession of goal, or missions for the actions selection agent, according to the general mission of the robot. It is the "ultimate" robot autonomy concept: the robot generates itself its own attitudes and its own actions by using its own decisions. The perception agent manages the processing of incoming data (the sensor measurements) and create representations of the environment. The actions selection agent must choose the robot behavior according to all information available and necessary to this choice: the fixed goal, representations and the robot localization. The action agent consists of a set of behaviors controlling the robot effectors. The hardware link agent is an interface between the software architecture and real robot. Changing the real robot require the use of a specific agent but no change in the overall architecture.

5 Experiment Of Eaas Architecture

To illustrate and validate our architecture we developed an application of remote control over Internet of an autonomous mobile robot that is able to know and avoid the obstacles in order to overrun the environment. This application represents for us a demonstration and remote test platforms. The web interface is designed with the intention of making it easy for users to control the mobile robot. A simple interface is designed to provide as much information as possible for remote control. This web interface consists of several Java Applets as shown in Figure 5. The user can directly control the mobile robot by clicking the start button on the control panel. The image display applet shows the visual feedback in a continuous jpeg image. The forum service allows users to send messages to each other, private or broadcast in order to interchange their ideas over the remote control subject. The remote user is invited to test the connection using the statistical or the dynamical way, before or during taking the control by clicking on the buttons labelled statistic or dynamic test respectively. This user interface allows students to

undergo a distance learning with the opportunity to test their ability on line.

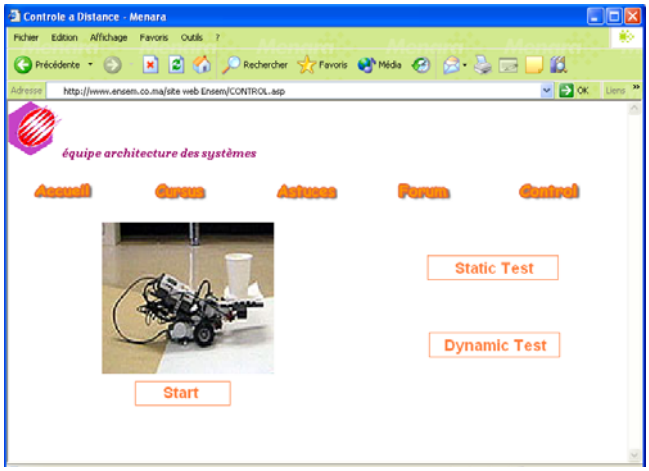


Figure 5 Web Interface

With this simple web interface, one user can control the mobile robot from the web browser with the visual feed back. The other users only have the visual feedback at the same time, and have to wait in queue until the first user logout at this stage.

6 Conclusion

In this paper, we have presented our control architecture based on multi agents approach to be able to manage the lack of quality of services of Internet in the context of remote control.

Our work and the works presented show that the remote control of robotic systems over an unpredictable network such as Internet is feasible and will be developed in the close future for tele-teaching, tele-maintenance, tele-expertise or tele- production.

References

[1] Le parc, P., Ogor, P., Vareille, J. & Marce, L. (2002). Web based remote control of the mechanical systems. IEEE International Conference on Software Telecommunications and Computer Networks, Split, Croatie, 2002

[2] K. Goldberg and R. Siegwart, Beyond Webcams : an introduction to online robots. The MIT Press, 2001

[3] K. Goldberg, S. Gentner, C. Sutter and J. Wiegley. The mercury project: a feasibility study for internet robotics. IEEE Robotics & Automantion Magazine, pages 35--40, March 2000

[4] K. Taylor and B. Dalton. Issues in internet telerobotics. FSR'97 International Conference on Field and Service Robots, 8-10 Décembre 1997

[5] D. Schulz, W. Burgard and A. B. Cremers. Predictive simulation of autonomous robots for tele-operation systems using the world wide web.In IEEE/RSJ International Conference on Intelligent Robots and System, Victoria, B.C., Canada, October 1998

[6] R. Simmons. Xavier : An autonomous mobile robot on the web. In In International Workshop On Intelligent Robots and Systems (IROS), Victoria, Canada, 1998

[7] M.R. Stein. Painting on the world wide web : the pumapaint project. In Proceeding of the IEEE IROS'98 Workshop on Robots on the Web, pages Victoria, Canada, October 1998

[8] P. Saucy and F. Mondada. Khepontheweb: Open access to a mobile robot on the internet. IEEE robotics and automation magazine, pages 41-47, March 2000

[9] S. Otmane, M. Mallen, A. Kheddar and F. Chavand. Active virtual guide as an apparatus for augmented reality based telemanipulation system on the internet. In IEEE Computer Society "33rd Annual Simulation Symposium ANSS 2000", pages 185-191, Wyndham City Center Hotel, Washington, D.C., USA, April 2000

[10] Le Parc, P., Vareille, J. & Marcé, L. (2005). Long distance remote control over Internet: a reliability challenge.. ln:XVI Workshop on Supervising and Diagnostic of Machining Systems. Karpacs, Poland

[11] A. Sayouti, F. Qrichi Aniba., H. Medromi,, A. Lakhouili. “Applying the Interaction-Oriented Approach to Remote Control of Robotic Systems”. Conferência Ibérica de Sistemas e Tecnologias de Informação (CISTI). Porto. Portugal. 2007

[12] Y. Secq. “RIO: Rôles, Interactions et Organisations, une méthodologie pour les systèmes multi-agents ouverts”. PhD report, University of Sciences and Technologies of Lille, France, 2003

[13] T. Finin, R. Fritzson, D. McKay, R. McEntire. “KQML as an Agent Communication Language”. Proceedings of the 3rd International Conference on Information and Knowledge Management CIKM'94, 1994

[14] Foundation for Intelligent Physical Agents. “Agent

Communication Language ". FIPA 97 Specification, Part 2, 1997

[15] P. Noriega, C. Sierra. Auctions and multiagent systems. "Intelligent Information Agents". Springer, p. 153-175, 1999

[16] Foundation for Intelligent Physical Agents. "Agent Communication Language". FIPA 99 Specification Draft, 1999

[17] M. Barbuceanu, M.S. Fox. " COOL : A Language for describing coordination in multi-agent systems". Proceedings of ICMAS'95, 1995

[18] K. Kuwabara, T. Ishida, N. Osato. "Agentalk: describing multiagent coordination protocols with inheritance". Proceedings of 7th Int. Conf. On Tools for Artificial Intelligence (ICTAI-95), 1995

[19] S. Khalfaoui, W. Lejouad Chaari, A.M. Pinna Dery. "Interactions entre composants pour environnements Multi-Agents". Proceedings of Journée Multi-agents et Composants (JMAC'2004), Paris, France, November 2004

[20] Cyril Novales, Gilles Mourioux, Gerard Poisson. A multi-level architecture controlling robots from autonomy to teleoperation. First National Workshop on Control Architectures of Robots. Montpellier. April 6,7 2006

[21] Sayouti, A, Qrichi Aniba, F., Medromi, H. (2008). Remote Control Based on Multi Agents Systems. First International Conference on Research Challenges in Information Science (RCIS), Marrakech, Morocco

Research and Application of the Web-Based Group Collaborative Learning System*

Yuyan Jiang

School of Management Science and Engineering Anhui University of Technology, P.R.China, 243002

Email: yuyan_j@yahoo.com.cn

Abstract

With the further development of the theory and technology of the CSCW (Computer Supported Collaborative Work), the idea of CSCL (Computer Supported Collaborative Learning) has been introduced to the modern education of computer networks. This paper presents the concept of the web-based "Group Collaborative Learning System". The system forms virtual groups for collaborative learning through cooperation environment of CSCW, and emphasizes the instant interaction and the collaborative study between members of learning. In addition, this paper discusses group cooperative mechanism of web-based CSCW applications and presents a collaborative learning based application model.

Keywords: CSCW; CSCL; Cooperative mechanism; CSCW application model

1 The Concept of the Web-Based Group Collaborative Learning

With the rapid evolution of WEB technology and its popularity in usage, many tutoring systems of different styles are established on WWW. The examples include Collaborative Online Research and Learning (CORAL), Students Assessment and Evaluation system (Albatross), Agent-Assistant Artificial Intelligence in Education and Active/Cooperative Learning (LISA). The introduction of those systems has greatly advanced the development of INTERNET based tutoring application. Nevertheless most of the systems only offer static web pages of teaching materials and lack the interaction between students, as well as between teachers and students which makes students' learning very isolated. Such a system has not truly taken advantage of the Internet and the teaching result is very limited and the students are not very enthusiastic about it. Therefore, one of the ways to solve these problems is to set up the web-based group collaborative learning system which can take full advantage of the internet learning functionalities break through the limits of the traditional teaching methods in practice and in space, improve the students' self-study abilities, and improve learning for students through discuss problems and collaborate with other students on the same network.

With the advance of the society and development of technologies, various jobs are becoming more and more complicated. A task once can be finished by a single person or a few number of people. But now more often than not, the accomplishment of a similar task needs the collaborative wisdom of many more people. So in nowadays society, working in collaboration becomes increasingly important. The most important issue in the realm of working in collaboration is to improve the overall working efficiency. Therefore, with the advance of the computer and communication technology, Computer Supported Cooperative Work (CSCW) [1] is born accordingly. The concept of CSCW

* [The project of the fund] social sciences research project of education department of Anhui Province (Serial number: 2006skj164)Brief introduction of author: JIANG Yuyan (1966-)，Female, XuanCheng Anhui , associate professor, the main research direction is the management information system , CSCW theory and application.

was originally proposed by Irene Greif of MIT and Paul Cashman of DEC in 1984. Its proposition and realization will fundamentally change the traditional way people work and their life styles.

The concept of CSCL (Computer Supported Collaborative Learning) comes from CSCW. The main meaning of CSCW is to achieve the collaborative work with the assistance of computers. Its emphasis on people, computer, and the interaction between them reflects both the notion of real computer-based systems as well as their psychological, social and organizational effects. We can say that CSCW is a multidisciplinary realm. CSCL is evolved based on the foundation of CSCW both in theory and in technology, with infusion of the theories of collaborative learning. It is currently one of the important subjects in the computer network development and research field. In other words, CSCL is the utilization of CSCW in computer supported teachings.

2 The Relevant Researches of the CSCL Systems

2.1 Learning through Collaborative Visualization (CoVis)

Learning system through Collaborative Visualization (CoVis)[2] is an interesting research project, which sought to transform science learning to better resemble the authentic practices of science, is to make a community of students, teachers, and researchers all working together. The goal of this project is to establish a electronic community in science learning, Its design concept is the “community of practice”, participating students studied atmospheric and environmental sciences through inquiry-based activities in a computer network based visual learning environment. Students can use the same visualization software tools to collect and deposit information from anywhere on internet to achieve the remote, real-time collaboration learning. The CoVis Project provided students with a range of collaboration and communication tools including online discussion, E-mail and video teleconferencing.

2.2 Computer Supported Intentional Learning Environment (CSILE)

Computer Supported Intentional Learning Environment[3] is an educational system based on the concept of collaborative database. CSILE’s main purpose is to promote students think and reflect their thought process, and it is not designed for any specific subject of knowledge. Students can retrieve the current and previous thoughts of other students from the CSILE database and they can also store their own thoughts for the reference of others. CSILE is an asynchronous discourse tool which focus is on knowledge-building contexts rather than on knowledge telling. In the process of "knowledge building", students can make great efforts to broaden their knowledge by knowing what they lack and then borrow others’ idea and assistance.

2.3 The National Center for Supercomputing Applications (NCSA)

NCSA (The National Center for Supercomputing Applications)’s Habanero project is also very innovative. It accomplishes cross-platform application sharing by interchanging information encapsulated in Java objects. The Habanero framework or API is designed to give developers the tools they need to create collaborative Java applications. The framework provides the necessary methods that make it possible to create or transform existing applications and applets into collaborative applications. The Habanero environment enables each participant to create, join, leave, and browse sessions. The detailed information about a session such as its schedule, agenda, list of the current collaborative tools on and who is allowed to participate, etc, is defined by its initiator. A user can participate the conversation by joining the session. There is no inherent limit in the number of tools per session, nor is there a limit on the type of tools that may be shared.

There are many other examples, for instance, a collaborative learning system helping the professional

growth of middle school science teachers called LabNet[4]; TENet online learning systems serving elementary school teachers, education officials and experts; 'Wired for Learning', an innovative collaboration system which helps school teachers, students and parents to communicate on-line[5]; OWLink project which develops a community-of-practice among teachers and students that emphasizes the use of technology to change instruction from the teacher-centered model to one in which emphasizes the collaborative nature of learning[6]; the teachers and students working together "Mission To Mars" courses[7], etc.

3 The Management Mechanism Management Collaborative Learning System

The existence of the collaborative mechanism is the key for a collaborative learning environment. It is the foundation for an effective group operation. But currently, most relevant research in this field emphasizes more on providing the man-machine interface on the network rather than providing a thorough definition for the patterns and rules of collaboration which CSCW emphasizes. In fact, the key for the success of network especially web-based CSCW application is to have an effective collaboration mechanism for the users to communicate and coordinate with each other. The collaboration mechanism is actually to provide supporting tools or control mechanism for a group collaboration, such as to support and maintain a distributed collaborative system.. Some current available tools including 2D Chat room or Microsoft's Net meeting do not have a mechanism to control participants' rights and obligations. Thus they can neither simulate the possible problems nor the resolutions of the problems in the real world collaboration. So, in the design of network based CSCW applications, we must consider participants' rights and responsibilities in the communication activities in order to achieve the goal of real collaboration. This research project will, according to the requirements of the collaborative learning system, propose some rules and mechanism for this kind of CSCW applications, and also provide an application model. When using a web-based collaboration system, the following factors need to be considered: Each user is autonomous and at the same time supports other related objects' autonomous activities; All objects are correlated to each other; Each object can monitor and response to any interested event simultaneously; The object will be notified of any event through the notification mechanism; Tasks of the collaboration can be delegated to other objects when needed; There are some inherent constraint to strengthen and restrict the behaviors of objects; Dynamic behavior is supported which allow group's behavior change in the course of a collaboration; When conflicts occur between objects of the collaboration, a role like a "teacher" can step up to coordinate and make reconciliation;

Due to the connectivity and dependencies of the objects in the collaborative system, the related people or objects can be notified by an event trigger mechanism on a real-time basis; The multiple agents of the collaborative system can deal with any conflicts through delegation, communication, cooperation and negotiation; To ensure the smooth execution of the web-based collaboration, some explicit guidelines need to be established. All participants will have to follow these rules when communicate with each other. . Although implementing these rules can slow down the activities like conferences, and some participants may even feel the restrict of freedom, they are necessary to control and enhance the effectiveness of the group activities. This research concludes that to achieve an effective collaborative learning system, the following functionality characters need to be established:

a. Object association. The system should record and maintain the mutual dependencies among its members. For example, Course members include many teachers and students. The relationship between teachers and students or students and their follow students need to be consistent. Such relationship includes dependency, inclusion and exclusion, etc.

b. Automatic notification. When a certain event happens, the system can automatically notify relevant objects in time. When an object is changed by a certain

user, all participants who requested the related objects will be notified and they will make changes according to connected-diagram. For example, when the storage catalogue of the popular tools for a certain subject is changed, users in charge of related teaching materials or teaching in the subjects will be notified.

c. When objects need to be adjusted, the delegation mechanism can give optimal reallocation of the different tasks. For example, if one of the collaboration members cannot participate in the activity, the absent person's tasks can be transferred to other suitable person or systems by the delegation mechanism.

d. If there are problems in interaction between different participants, negotiation is needed. The course of negotiation is the course to establish and identify an agreement. For example, when giving a quiz in classroom, students may think the time is to short to finish all the questions. Then they can use the negotiation mechanism to negotiate with the teacher and discuss if the time can be extended.

e. When some predefined events are triggered, the event monitoring mechanism can make corresponding response. For example: if a student's marks in exam is below 70, the system will automatically go back to the procedure of reviewing lessons.

f. each user, The learning processes are recorded thoroughly for each user. The purpose is for later analysis and to make teaching adjustment. At the same time, all the activities of the participants can be tracked including when a user entered the system, what activities he did, etc. This will help the system to understand members' usage pattern.

g. Constraint management. Every participation in a collaborative system has different right and obligation. To differentiate these rights and obligations, some rules have to apply for each participant. For example, when in the middle of a learning discussion, any participant cannot leave or join in without the permission of the meeting coordinator.

h. Harmony and decision. Sometimes conflicts happen due to the different needs and expectations of different users in the system. At this time, a coordinator is needed to deal with the conflicts. Usually, the basis of coordination is the constrain parameters stored in database. If the requests or choices violate the rules defined by the parameters, those requests or choices will be denied. For the remaining ones conform to the rules, a best one will be chosen according to some optimization algorithm. Another way would be notify the users to discuss for a new round of selection and finally be selected by the coordinator.

4 Systematic Realization and Application

4.1 The structure of a collaborative learning system

Based on the above stated collaborative mechanism, a system model for the collaborative learning system is presented in the following.

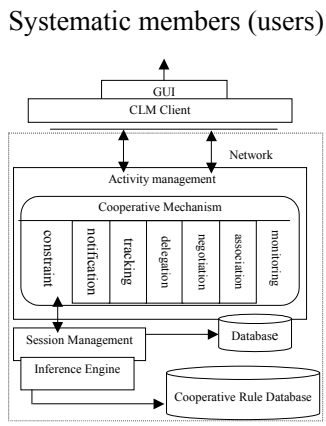


Figure 1 Diagram of the structure of a group collaborative learning system

The model consists two main parts: CLM Client, CUI unit and CLM system. The CLM client is in charge of connecting to database and provoking information communication of the collaborative mechanism through Internet and CLM server machines. GUI unit offers graphical user interface for users to interact with the system in the web-based teaching environment. The users can also use other communication tools to talk with other users.

CLM system includes several parts: Activity Management, Session Management, Cooperate Mechanism, Database System, Inference Engine, Regular Cooperation Database, Database。

a. Activity Management. Activity management is

responsible for the launch of the server when an activity is allowed to start and shut down the server when the activity ends. Once the activity begins, the activity server will gradually accept the participants to join in. The behaviors of all participants in the activity will be supervised. Activity management will also manage and arrange conversations, in other words, it can accept the request of a new conversation from client. In addition, activity management communicates with the database management system for data read and write.

b. Session Management. Session management is responsible for managing and supervising the conversation started via the activity. It monitors all activities of all participants in the conversation, will store the records into the database for history tracking. Session management is also responsible to deal with any events like withdrawal during the middle of conversation or the notification of warning messages.

c. Cooperative Mechanism. Cooperative mechani sm is the core of the whole system and is what we summed up above.

d. Database Management System. Database management system provides an interface between the application and the database. It interacts with the activity management to accept or reject the client's requests.

e. Inference Engine. When the participant triggers the collaborative mechanism, Inference Engine will infer based on the collaboration rules stored in database. After inference, it will send the results to the system and the system will send the appropriate feedback to the participants.

f. Cooperative Rules Database. Cooperative rules database stores the rules and constraints of the collaborative mechanism.

g. Database. Database stores data for conversation management and its corresponding activities.

4.2 Systematic Procedure and Conversation Procedure

a. Systematic Execution. After the CLM system is started, it will wait for the connection from the client. After the user logged in, the conversation and activities can be executed accordingly. When the conversation and activities end, the users can decide if they want to disconnect. If so, the system will stop the execution. There is an arrow head button for the user to go back and forth in conversation. This is called a nested-activity or nested-session. The so-called nested-activity means that an activity contains a lot of sub-activities . These activities can go on at the same time. A nested-activity can include dummy activities. Nest ed-session is also the same.

b. Conversation execution procedure. The partici pants may start conversations from the system activity. Only after searching through all the existing conversa tions first, the participant can decide whether to open up a new conversation. If no new conversion is initiated, the participant will join in the existing conversation. After the open up of a new conversation, the conversation agenda and rules such as the number of participants or the main topic of the conversation will then be defined. Then the conversation can start and participants can join in the conversation to discuss and work together. If any event was triggered during the conversation, the collaborative mechanism will do the corresponding response, for example, due to the interactive relationship among the objects, user can join in other relate conversations simultaneously. Before the end of the conversation, users can resume conversation at any time. , This is the nested-session conversation state above.

4.3 Developing instrument and Setting-up of the model system

This system is the Chinese version built on the Windows NT operating system platform. It uses Microsoft Internet Information technologies (IIS) on server side. In client site, it uses Microsoft Access 2000 database management system and ASP for application coding. As for GUI, it uses HTML, Dynamic HTML, VB Script, Java Script. All web page prototypes were first automatically generated using Dreamweaver and then used MS Visual InterDev 6.0 to add the dynamic programming parts. MS Visual InterDev 6. 0 is a very

convenient visualized IDE for writing applications. It consists HTML and Script editors, database development tool and allows developers to integrate other software components to the developing system.

5 Conclusion

The emergence of the CSCL concept not only has promoted the research on CSCW itself but also promoted the rapid development of the network based collaboration work especially the web-based collaboration learning systems. The key to build a successful collaborative learning system is a group collaborative mechanism. This article discussed many aspects of the application such as technologies, system design, system functionalities, system development tools, ect. It attempts to use the CSCW concept to build an online simulation system in collaborative learning for the purpose of learning discussion and experience sharing.

References

[1] Grudin J. Computer—supported cooperative work：history and focus. Computer，1994；27（5）：19—26

[2] Pea R.The collaborative visualization project[J]. Communica tions of the ACM, 1993 ; 36（5）: 60-63

[3] Scardamalia M , Bereiter C ,Lamon M.The CSILE project: Trying to bring the classroom into world 3[C]. In K McGilly(Ed.).Classroom lessons: Integrating cognitive theory and classroom practice, Cambridge, MA: MIT Press, 1994:201-228

[4] Spitzer W.Wedding K.LabNet:an international electronic community for professional development[J]. Computers and Education, 1995 ;24（3）: 247-250

[5] Kuang L Grueneberg, K & Lam R.Education on the net Constructing collaborative learning communities [C] .In Proceedings of The 6th International Conference on Computers in Education, Beijing, China, 1998: 543-545

[6] Miller L M.Teacher′s users of an electronic environment : A case study of project OWLink[C].In Proceedings of The 6th International Conference on Computers in Education, Beijing, China, 1998: 459-465

[7] Brown A L,Campione J C.Guided discovery in a community of learners[C].In: K McGilly(Ed.),Classroom lessons: Integrating cognitive theory and classroom practice, Cambridge, MA: MIT Press, 1994:229-270

A Wiki-based Study on Web-based Course of Principle of Database System

Yanhong Xie

Department of Computer Science and Technology, Dezhou University, Dezhou, Shandong 253023, China
Email: dzxieyh@163.com

Abstract

Currently contents of web-based courses still need to be updated by teachers and the process of construction cannot be recorded. This paper is aimed at this. In light of wiki, a web-based course of *Principle of Database System* is designed. This web-based course is simple to use with well-defined themes, open for cooperation and effectively demonstrates the process of accumulation and integration. What's more, this course is maintained by members of the community, highlighting co-operation, co-creation and equality. Thus it can encourage participations, advance with times and proceed in an all-round way. After introduction of the fundamental differences between web-based course and wiki system, the paper deals with the basic functions of wiki web-based course and analyzes the designs and realization of its key functions.

Keywords: wiki, Web-based Course, Principle of Database System, CuteEditor, ASP.NET, SQL Server 2005

1 Introduction

The design and development of web-based course is not only the new field, new concept of the course construction, but also the core and focus of modern long-distance education project. Current web-based courses mostly remain unchanged after completion and lack flexibility and timeliness. Considering the above mentioned deficiencies, the literature [1] puts forward the idea of self-help web-based course to resolve above problems efficiently, but this type of web-based course has still the following shortcomings: (1) Course contents can only be amended by teachers, while the others have to accept them passively. (2) Once the contents are revised, the original version will be lost, and the process of course's construction can not be record. So the author designed and realized the wiki web-based course of *Principle of Database System* using ASP.NET 2.0 and SQL SERVER 2005 by referencing the core idea of wiki. The wiki web-based course is maintained by members of the community, stressed the co-operation, co-creation and equality of them, which will be helpful to increase the senses of trust, belonging, sharing and reciprocity of the members and improve the enthusiasm of participation of them effectively, so the timeliness and completeness of the course will be guaranteed [2].

2 About Wiki

The concept's inventor of wiki is Ward Cunningham. The term comes from the Hawaiian "wee kee wee kee", and its original meaning is "quick quick". Wikipedia for the definition of wiki is: Wiki refers to a hypertext system. This hypertext system to support community-oriented collaborative writing, but also includes a group to support this writing aids. The wiki text can be browsed, created and modified with less cost than HTML text; at the same time Wiki system also supports community-oriented collaborative writing, offering the necessary assistance for it; Finally, Wiki writers can naturally constitute a community with Wiki system as a simple communication tool. Compared with other hypertext systems, wiki system is easy to use and open; it can help us to share the knowledge of certain fields in a community[3][4][5].

3 Function Design of Wiki Web-Based Course

3.1 Basic differences between web-based course and wiki

Wiki is a collaborative writing tool for people. Wiki system can be maintained by various people even any visitors. Everyone can express their views or expand or explore the common theme. Therefore, in order to guarantee the accuracy of the contents of page, some norms about technique and operational rules such as reserving every change version of website, page locking, IP ban, edit rules are formulated. But the preciseness of techniques and the speciality of oriented community determine that some limitations and improvement must be made according to the actual situation when wiki idea is applied in web-based course[6]:

1) The category of user is different with wiki and they should be divided into three types: teachers, students and those who concern the course. To ensure the safety of the course content, registration applies of students and those who concern the course need the teacher's approval. Only after approval, can they have some certain rights of revising the course content.

2) Functions of user prohibition and user tracking. Most users use the public computers on campus, and so it is difficult to fix IP address. Therefore, when a user has some destructive behaves, "user prohibition" is adopt to cancel the user's updating right instead of "ban IP", and its login time and IP address are recorded.

3) Teachers have update right of some page content such as course outline, teachers brief introduction, schoolwork etc, but students and those who concern the course have right of read-only.

4) Students and those who concern the course can set the classification level of new content created by themselves. For example, some user's schoolwork and test can be set as not be modified or cannot be visible to others.

5) Online test is an important function of web-based course, but wiki system does not have this function

3.2 Basic function design of wiki web-based course

Principle of Database System is a web-based course integrated the idea of wiki, but is not all the same with wiki, and its main function designs as follows[7]:

(1) User management

The course's user is divided into three types: teachers, students and those who concern the course. Teachers are the highest level of rights in the course, who can add, delete users of teachers and students in batches, examine and approve the users who register the course, forbid some malicious users. The users of students and those who concern the course must register the course if they want to manage the course's content. Only after approval of the teachers, can they manage the course within a certain range.

(2) Category management of the course columns

According to the characteristics of web-based course, the course provide some types of course columns such as letter depiction, video playing, picture showing, courseware presentation, real-time exchange, test evaluation and personal knowledge collecting etc.

When users create new columns, the course will present different page style according to the user's selection to adapt to the needs of different character of subject and teaching design. Columns of real-time intercommunion type columns provide some intercommunion ways such as traditional BBS, chat room etc, test and evaluation type columns provide input and revision of variety question types such as multiple-choice, multi-select and error correction etc as well as generation of test paper.

(3) Page edition

Teachers and some students and those who concern the course who own the right to modify the pages can use this function. When they find some unreasonable columns, inadequate description or have some new ideas, they can not only amend the contents of the course, but also create their own programs to enrich the content of course. Of course, teachers can also lock some pages such as brief introduction of teacher, assignments that need not to be modified to prevent the pages being edited.

(4) Version comparison

The course saves some information such as every edit content, editors and edit time etc into database, thus, user can easily find out any history version, and can compare the differences between the two versions. Even if the malicious user make the page in mess，even delete the entire page, teachers can easily select the most satisfying version from the original version to restore.

(5) Personal contribution management

The various types of user can browse, modify and delete contents of pages that created or edited by them, and they can set attributes of those pages such as whether the page contents are permitted to update by other users and what kinds of users have the rights to modify them etc.

(6) Other ancillary functions

For the convenience of users, the system also provides other ancillary functions such as the latest developments, recent changes, the focus tracking, recommended columns, label management etc.

3.3 The basic characteristics of wiki web-based course

(1) The open and safety of the course contents. The users of the web-based courses at all levels have different privileges, but all users have create, modify and delete privileges of the majority columns, and the change of the web-based course's page can be observed by visitor at any time, and this can help users to understand the news and hot spots of the course. Of course, the course also has more comprehensive protection mechanisms, such as user prohibition registration approval, lock page, version comparison. So wiki web-based course has a certain security measures when it is open for users.

(2) The equality of the courses users. The maintaining of the web-based course's content is completed by individual users spontaneously, and the equality of users is stressed. The main duties of the teacher users converted into the supervision and guidance from full-time maintenance originally, and the situation of the monopoly of teachers to update and maintain the courses is broken. So the participation passion of the users is stimulated

(3) Co-operation of course users and co-creation of contents

Wiki web-based course is a collaborative system and it hopes the user have the spirit of co-operation and take an active part in building the course to realize co-creation. At the same time the user can start their own way of thinking and release a personal opinion that based on respect and learn from others working, so that we can seek the different strategies of resolving the problem.

(4) Effectiveness of course contents accumulation

If the user is not satisfied with their published content, they can modify them at any time, of course the others can express their own opinion. The version of each modify is recorded and the difference of the two versions is compared. This will enable users to publish their own opinion comprehensively and coherently by referencing to the others views. So, the process of knowledge accumulation and integration in course-learning is showed dynamically and it can benefit the teachers' guidance and monitoring.

4 Technology Analysis of Wiki Web-Based Course

The course development uses: 1) Microsoft Visual Studio 2005 Professional Edition, with the .NET Framework 2.0. 2) Microsoft SQL Server 2005 Enterprise Edition. 3) The source code within this course is implemented in visual C#. In addition to some data tables those saved some basic information such as user, questions, student paper, BBS and chat room etc, the data tables are designed mainly such as column classify, page links, history resources, labels etc. Based on the above data tables, the above main functions of course are realized. The technical analysis of the design and implementation of its important functions as follows[8][9]:

4.1 User management

User management of course is implemented with login control of ASP.NET 2.0 which is composed of two embedded systems: member management and role management. User management of the course is realized flexibly and effectively by using and extending login control. Its realization steps as follows:

(1) Run aspnet_regsql tool in %systemroot% \ Microsoft.NET \ Framework \ v2.x directory. According to the wizard tips, data tables and stored procedure are saved in designed server and database.

(2) Revise the configuration file Web.Config to set AspNetSqlProvider as the default provider procedure of member identity. Code as follows:

```
<configuration>
    <system.web>
<membership
DefaultProvider="AspNetSqlProvider"/>
    </ System.web>
</ Configuration>
```

(3) Rewriting Membership class. The Membership user management of ASP.NET 2.0 is too simple, and some attributes properties that are closely related to the user have no way to add to the data table. Therefore expanding Membership is must to be done. The course adopt the method of rewriting the provider procedure of user attributes, and named the rewrite class of them as CustomProfileProvider that stored in CustomProfile-Provider.cs file in App_Code directory of current project.

(4) Revise Web.Config, customize user attributes, and configure provider procedure of user attributes.

(5) Make use of role management to add three roles: Teacher, Student and those who concern the course.

(6) The user management is realized easily to make use of Login, LoginName and CreateUserWizard controls ASP.NET 2.0 and MembershipUser class.

4.2 Page edition

The course adopt CuteEditor 5.0 as edit tool of page content so as to all users of variety degrees can edit the page content that includes all kinds of multimedia such as graphics, images, audio, video, animation easily and efficiently. CuteEditor is a powerful, Web-based online WYSIWYG(What You See Is What You Get) editor that supporting a variety of media upload, download and edit. Its function button and upload files type and size can be set in configuration file and type, upload position and operation privileges of user multimedia resources can be modified according to the actual needs. Make use of the editor's three view mode: *Normal* graphics view, *HTML* code view and *Preview* preview view comprehensively, the user could design some standardized, clean and colorful pages. The steps of making use of CuteEditor to edit the page as follows:

(1) Copy CuteEditor.dll and CuteEditor.lic to \bin folder of the project and modify Web.config file as follows:

```
<appSettings>
    <add key="DictionaryFolder"
value="bin"/>
</appSettings>
```

(2) Copy CuteSoft_Client folder in CuteEditor folder to the root directory of the application, and create a upload folder in it.

(3) Open the Visual Studio 2005, right-click and select the "options" of the toolbar, then click "browse" button of .NET Framework component and select CuteEditor.dll in bin folder. After above steps, Editor icon appears on the tool bar and just drags the Editor icon to the page if the user wants to use it.

(4) If the user wants to insert multimedia information, click the corresponding icon to upload the multimedia files to the server and then choose them to insert, whose operation method is similar with Office Word.

4.3 Version comparison

In order to realize the version compares function, the course designs a historical resources table, and the main fields are designed as follows: SrcID (historical resources version identity), SrcUserID (historical resources editor's ID), SrcTypeID (historical resources

columns classification), SrcContent (historical resources content), SrcEditTime (the time edit the historical resources). Version compares function in accordance with the following steps:

(1) Select the knowledge of historical version, and click the "version" button.

(2) All the history version and its revised version of the knowledge are listed in the form of "current-modify", as well as some other information such as updating time, updating user etc is displayed. The first record is the last modified version, so its "modify" content is empty.

(3) Chose one of the historical version of the "current - modify" option, click the "compare history version" button, the content of current version and its former version are displayed. The aim of two versions are listed at the same time is to compare the similarities and differences between the two versions.

(4) Select the title of a certain history version, its specific content is listed.

5 Conclusion

The web-based course based on Wiki has the characters of simple usage, clear theme, open cooperation and it can effectively demonstrate the accumulation and integration process of the course. It can not only effectively highlight the dominant position of the students and inspire students to learn, but also can help teachers guide and monitor. For teachers and students, it provides a collaboration platform of full exchange.

Although such courses have the function of user prohibition to ensure the safety of courses, it could not put an end to actions of malice vandalism. So the members of the courses should have higher quality and it suits to college and university that students are more self-discipline.

References

[1] Tong Ying, Wang Zhi-jun, Xie Yan-hong, Wang Xue. Design and Implement of Self-help Web-based course. Modern Educational Technology, Vol.17, No.9, September 2007, pp. 59-61:51

[2] Ding Qiao-rong, Zhang Lei, Zhang Cong-shan. Study of conform the Virtual learning community by wiki. China Modern Educational Equipment, No.9, September 2007, pp: 123-124

[3] Lai Xiao-yun. Roles of Wiki in Web-based Research Learning. Modern Distance Education, No.6, July 2005, pp: 40-50

[4] Wikipedia.http://baike.baidu.com/view/1245.htm

[5] iTwiki. http://wiki.ccw.com.cn

[6] Wang Jian. Application of Wiki in teaching. China Modern Educational Equipment. Vol.21, No.2, April 2007, pp. 141-143

[7] LI Xue-Jun, Li Long-Shu, CHENG Hui-Xia, Xu Yi. The Design and Implement of Wiki system Based on UML. Computer Science, Vol.34, No.7, 2007, pp. 251-253:274

[8] YANG Yong, ZHANG Wei-shi, Zhang Xiu-guo, SHI Jin-yu. Design and implementation of Wiki-based software design pattern libraty. Computer Engineering and Design, Vol.28, No.16, Aug.2007, pp.3978-3980

[9] Xu Qi. Research Study in the University Base on Wiki. Modern Educational Technology, Vol.17, No.10, October 2007, pp:67-74

Research on Heterogeneous Database Query Based on XML

Yushui Geng[1] Xiangcui Kong[1] Xingang Wang[2] Aizhang Guo[2]

1 School of Information Science and Technology, Shandong Institute of Light Industry, Jinan 250353, P.R. China
E-mail:gys@sdili.edu.cn; cui5271@163.com

2 Center of Modern Education Technology, Shandong Institute of Light Industry, Jinan250353, P.R. China
E-mail: wxg@sdlil.edu.cn; guoaz@sdili.edu.cn

Abstract

Heterogeneous database system has brought many difficulties for user's query. The appearance of XML bring hope to integrate heterogeneous data. This paper designed heterogeneous database integration system's structure and query processing based on XML. We used XQuery as the global query language on querying disposal, and mainly studied mapping and query optimization technology.

Keywords: XML, Heterogeneous database, query, Xquery, mapping

1 Introduction

Data integration keeps a hot research topic in this decade. Because users usually access data via queries, and data integration systems often describe data sources as views on the global schema, query processing becomes one of the core problems of data integration. About data integration, a common data model is needed to provide a global view to users. XML is a versatile markup language, capable of labeling the information content of diverse data sources including structured and semi-structured documents, relational databases, and object repositories. XML brings hope to integrate heterogeneous data, because of the character belonged to XML which includes self-description, flexibility, the powerful ability of data description and data exchange, easy for expansion and so on, so that XML can easily interact with other data models without any defect as standard data model. This system uses xml as the middleware of data description tools, through the mapping relations, transforms each kind of heterogeneous database into the unified xml form, uses XQuery as query language for the user to provide the query method and the contact surface.

The rest of the paper is organized as follows. In Section 2, we present related works on XQuery and mapping. In Section 3, we introduce the detailed design of the system's structure and query processing. In Section 4, we have carried on the query optimization from two aspects. Finally, we summarize this paper and indicate some directions for future research work in Section 5.

2 Related Work

Problem Statement

In recent years, there have been a number of researchers focusing on the integration of XML data and heterogeneous data like relational data or Web information source described by URI [1, 2]. An increasing number of business users and software applications need to process information that is accessible via multiple diverse information systems, such as data systems, file systems, legacy applications, or Web services. There are some studies on the integration of distributed relational data [3,4], integration of relational data with semi-structured data [5,6,7], and efficient XQuery processing in a single system [8]. However, these approaches are not direct methods for the integration of heterogeneous data with XML data. The existing data integration and query technology has some insufficiencies. (1) The conversion

of overall model with the data source is very complex. (2) Query processing efficiency is low and the data transmission capacity is big. (3) It cannot solve multi-platform heterogeneous and realize the data long-distance operation.

The Introduction of Xquery

A query language that uses the structure of XML intelligently can express queries across all these kinds of data, whether physically stored in XML or viewed as XML via middleware. This query language called XQuery, which is designed to be broadly applicable across many types of XML data sources.W3C proposed the XQuery language uses xml to take the abstraction data model, may make the inquiry to the above type data pool. XQuery not only may use in the xml documents, it also has a more actual use, may use for to take based on xml heterogeneous database integrative system's query language [9]. XQuery unified SQL and the Xpath function, has provided one kind of general query processing language. A powerful feature of XQuery is the presence of FLWR expression(for-let-where-return). The for-let clause makes variables iterate over the result of an expression, the where clause allows specifying restrictions on the variables, and the return clause can construct new xml element as output of the new query. In general, an XQuery query consists of an optional list of namespaces definitions, followed by a list of function definitions, followed by a single expression [1].XQuery uses for-Let-Where-Return grammar, for expression circulation visit the node collection which returns by the way expression, the Let expression returns way expression returns single node.

Mapping

In order to realize data integration based on XML, data transformation must be taken into consideration. The key is the transformation between XML and database. In [10] the authors proposed the R2XL Mapping language which describes the relational data to the XML documents mapping. We use this language as our mapping language. By defining the mapping document will be distributed database mapping data to meet user requirements for XML documents. From the relational scheme to the xml pattern's transformation, is actually operating the xml form which defines in advance.

For the single table's mapping, it is simple in the process of transform database to xml documents. The transformation of single table will be completed as long as a table’s mapping is an element and each row in this table mapped into this element’s sub-element or the attribute. When comes to the multi-table transformation, the focus is how to resolve the table between the foreign key relationship.

By the database multi-tables to the XML documents' mapping, these rules should be followed:

(1) First, a table is designated as a "root table," and its mapping is a new element.

(2) For the root table's non-foreign key column, its mapping is a new element, simultaneously as root table element’s sub-element. For elected spatial column, establishes it for choose the sub-element.

(3) For the table which establish contact by foreign key and root table, we can use the same method above to solve it. The elements corresponding to this table will seem to be the sub-element of the elements in root table. If this table is connected the root table through the foreign key, then it takes the root table’s sub-element which is elected and duplicated (equal to * char).If the root table is connected this table through the foreign key and foreign key is empty, then it takes as the root table’s sub-element can may be elected (equal to? char).

(4) For the principal key columns of the root table which contacted through the foreign key, if it has not been connected other tables as the foreign key, those columns will be mapped into new element and as this table element’s sub-element.

3 Detailed Design of the Structure and Query Processing

The design of this system mainly uses middleware integrated based on the xml. The system’s structure as shown in Figure 1. The XML heterogeneous integration middleware is mainly composed of Wrapper and Mediator.They completed the main functions of the middleware system. The Wrapper’s function is causing

heterogeneous database to be interactive, packing heterogeneous data, realizing the operation uniformity, the data visiting transparency and the position transparency. Each Wrapper provides a virtual view of xml for Mediator, thereby shielding the various differences between data sources. Mediator is the core of the query processing. It provided a high-level xml hypothesized view to the user. The user can use XQuery to carry on the distributional inquiry to these data.

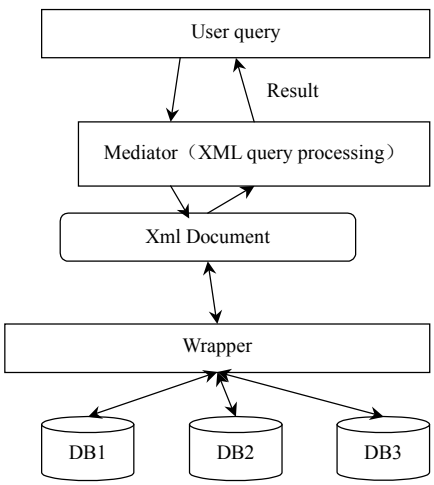


Figure 1 The System Structure

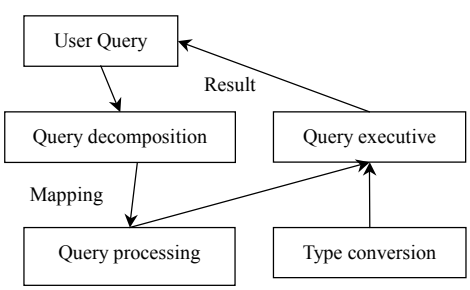


Figure 2 Mediator processing Structure

The main module and the function of Mediator are as follows:

(1) The query decomposition module: this module realizes the grammar decomposition of the XQuery query, and judges the exist of query data according to the XQuery query read-in corresponding mapping document, and decomposition the mapping document to be a mapping tree, finally transmits analyze result query to the processing module.

(2) The query processing module: according to the XPath expression of XQuery query, this module prunes the mapping tree to produce a new mapping tree, according to the new mapping tree and the condition of the user query sentence to carry on the optimization. After this query processing, only the data which meet user's request is query and be transformed.

(3) The query executive modules: according to the finally mapping tree, this module extracts data description part and produces a series of sub-query. These sub-queries carry out separately for each data pool according to its description information. It carries on the data conversion according to the mapping tree's data definition and the document's structure, then combines the query result and produces a hypothesized DOM tree. Finally it produces the goal XML documents and returns it to user system.

(4) The type conversion module: this module is composed of a series of data type transformation functional module. It is used for query executive module.It conversed the data which extracts from the database to DTD or XML Schema data type. These data type conversion module is alterable which can be defined by the system manager.

4 The Query Optimization

Prune To The Mapping Tree

The FLWR expression in the XQuery language is a very important part. The major part of the XQuery language is the expression of FLWR [11]. FOR-LET-WHERE-RETURN support iteration and may bind the middle variable to the intermediate result. The FLWR expression includes: 1 or many "FOR" (to bind node to variable to circulates traversal) clauses; 1 or many "Let" clauses (to Allocated on a value or sequence for variable); 1 choosed "Where" clause (judgment sentence, if it is true, this tuple will be retained and binds to return cluase); 1 "Return" clause (defines content which each tuple must return). The query processing module prunes the mapping trees according to the FOR clause, the LET clause and XPath clause expression, which in the mapping document the user does not need and only retains the user need the part.

Finally it produces a new mapping tree after above steps. After such processing, we raised the query efficiency.

Rewrite The Query Sentence

We may carry on the decomposition to some query sentence. For example the attribute definition's SQL sentence of the value is book‘s data attribute is: “select bid, year, title, price from Book”. When the user carry on the query, this SQL sentence will be escorted to the corresponding database to carry on the query, then the result returns middleware level. Finally according to mapping tree and where clause in XQuery sentence query condition, this SQL sentence decomposition is " select bid, year, title from Book where year>2005 ". Only then user need's data will be queried in the data source level, and the result returns for the middleware level. This reduced the data current capacity between the data source level and the middleware level greatly.

5 Summary and Forecast

This article proposed a middleware design of heterogeneous database integration based on the XML. This system put heterogeneous database data into the XML view through the mapping rule and the R2XL language. The user uses XQuery languages realizes heterogeneous database data’s transparent visiting. Its advantages as follows: (1) Using XML as the public data description language, which simplified the transformation of the data model greatly. (2)Unlike heterogeneous database integration system, it not only integrated DBMS database, but also integrate non-database systems management data. (3) The powerful universal XQuery language makes query xml global and local views in a uniform way easily. It raises the query processing efficiency and system's extendibility

This paper is the initial exploration of heterogeneous database integration and inquiry technology. There are many issues to be further researched. Although XQuery is one kind of function strengthened XML query language and can realize many kinds of data type inquiry, it cannot look like the SQL language to realized insertion, deletion and renew to the data. Therefore this system only supports the unidirectional inquiry. How to make the system supports insertion, deletion and renewal to the data could be a difficulty. I hoped that we can study a deeper level in the later job.

References

[1] Manolescu, I., Florescu, D., Kossmann, D.: Answering XML Queries over Heterogeneous Data Sources. In: Proc. VLDB, Roma, Italy, pp.150-160. September 2001

[2] Shanmugasundaram, J., Kiernan, J., Shekita, E., Fan, C., Funderburk, J.: Querying XML Views of Relational Data. In: Proc. 27th VLDB, Roma, Italy, pp.135-145 September 2001

[3] Kossmann, D.: The state of the art in distributed query processing. ACM Computing Surveys 2000, pp.422-469

[4] Suciu, D.: Distributed query evaluation on semi structured data. ACM Transaction Database System, 2002, 27(1), pp.1-62

[5] Zhang, X., Pielech, B., Rundensteiner, E.A.: Honey, I shrunk the XQuery: an XML algebra optimization approach. In: Proc. WIDM 2002, McLean, Virginia, USA, November, 2002

[6] Su, H., Rundensteiner, E.A., Mani, M.: Semantic Query Optimization in an Automata- Algebra Combined XQuery Engine over XML Streams. In: Proc. VLDB 2004, Toronto, Canada, September 2004

[7] Zhang, X., Rundensteiner, E.A.: XAT: XML Algebra for the Rainbow System. Technical Report WPI-CS-TR-02-24, Worcester Polytechnic Institute, 2002

[8] Koch, C., Scherzinger, S., Schweikardt, N., Stegmaier, B.: Flux Query: An Optimizing XQuery Processor for Streaming XML Data. In: Proc. VLDB 2004, Toronto, Canada 2004

[9] Srinivas Pandrangi. Practical Applications of XQuery, XML Conference, 2003

[10] Yao, Q., Zhao, P.: Method of publishing relational data as XML Document, Computer Engineering and Application, 2007, 43(15), pp.160-162

[11] Cheng, l., Zhu M.: The Implementation Mechanism of XQuery, Computer Engineering and Application, 2002, 24, pp. 78-80

The Application of BizEngine to Information Management System

Ying Zhang

School of Information Technology, Jiangnan University Wuxi, Jiangsu Province 214122, China

Email: wuxizy_2004@163.com

Abstract

Through the last two decades, China has achieved enormous progress both in scale and level of national taxation information construction. Compared with developed countries, our information system still lacks of standardization and unity. BizEngine is a process-centered workflow management system based on Web which aiming at satisfying the need of both administrative workflow and production workflow. We prove that BizEngine could enhance the stabilization and the security while providing a mature interface to the digital tax integrated services.

Keywords: BizEngine; workflow; information

1 Introduction

China's taxation information construction has reached a higher level through the last 20 years' effort, it has already played an important part in the nation's tax revenue. The modern management of tax collection based on multi-information system CTAIS, golden tax project, official document processing, anti-tax shelter and export drawbacks, has promoted the increase of taxation year by year through changing its management concept and improving the work efficiency. Furthermore, it has laid abundant foundation for the further development of the taxation information management. Nowadays, the taxation information technology has become a necessary condition for improving the national tax collection and administration.

Generally speaking, there are still a lot of problems unsolved in our county's tax revenue system, such as regional unbalanced development, inconsistent and unfair criterion, unified planning lacking and low level of automatic intelligence technology. According to the national condition and the needs of tax collection and management reform, combining multi-application systems has become essential.

2 Taxation System Architecture Utilizing Bizengine

Workflow of BizEngine

Workflow is the most typical and universal operation processing pattern. It's inevitable to treat with all kinds of workflows in so many different software systems. The most simple method is dealing with every conditions and processing by algorithms, but with the obvious disadvantage of rewriting the whole algorithm once the process changes even a little which makes the maintenance turn to be inconvenient. Therefore, a much more generalized workflow engine has emerged as the times required, which designs and supervises the process of workflow by definition and configuration rather than programming.

BizEngine is a process-centered workflow administration system fully based on Web and satisfies the both need of administrative workflow and production workflow. It realizes the design, operation and the supervision of the workflow, moreover, it provides a resort to connect or exchange the data with service database. BizEngine provides a series of convenience graphic tools to design and modify the process according to the reality which makes maintenance much easier. It also can unify different processing methods and integrate the framework with

authority convenient for use.

BizEngine provides mature application-oriented software components, middleware and management tools. The main features are as follows:

1. Five species of kernel components, middleware and management tools (Figure 1) which used in information integration, process integration, data integration, security integration and message integration devotes to forming a mature and steady interface for the system development.

2. Agile building ability such as visual process definition, visual data-mapping definition and visual intelligent table list definition can perfectly suit the changes of operation, which solving the confliction among universality, standard and individuality.

3. Reducing the construction risk by shortly cutting the system's developing cycle.

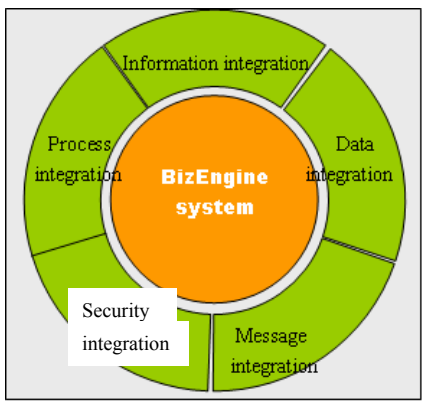


Figure 1 component of BizEngine

Digital synthetical tax management platform

This management platform consists of three parts: BizEngine application support platform, digital tax platform and system interface (Figure 2).

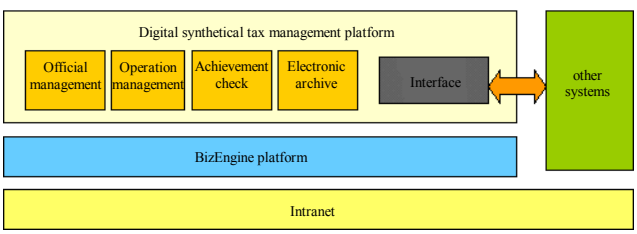


Figure 2 the constitutes of digital tax management system

- BizEngine application support platform

BizEngine application support platform provides general software components, middleware and management tools. Including:

1. Security integration

Unify the user account and identity authentication in each sub-system by providing the uniform directory services, security authority management, security log and data certification management.

2. Process integration

Insure the distribution, movement, operation and supervision of the tasks by the platform of designing, operating and supervising the workflow. Make the users design the process and the running condition easily without any professional software.

3. Message integration

Combine various communication methods such as e-mail, instant message, mobile phone short message, telephone call and fax as a whole platform with the unique access interface.

4. Data integration

Enable accessing and connecting various databases, including data-mapping between working process and operation system, synchronous construction between different systems, and individual data view.

5. Content integration

Provide custom-built and individual information platform, strong content management and search ability as also.

6. Open interface and external interface

Connect all kinds of external application through the access interfaces providing by Web service and XML.

- Digital tax platform

Digital tax platform is the logic operation implement, mainly consists administration platform, operation management platform, achievement assessment platform, electronic archives and decision-making support platform.

1. Administration platform

Construct a uniform organization and personal administration by combining all kinds of official process into a whole platform with the function of affair management, administration supervision and education.

2. Operation management platform

Incorporate different departments' operation through tax collection, login management, proving management, survey and evaluation, invoice management, anti-counterfeiting and tax inspection.

3. Achievement assessment platform

Given evaluation result automatically according to the dynamic checking standards, implement fetching data and create checking results automatically.

4. Electronic archives

Create a large electronic archives database; support the functions of classification, pigeonhole, searching and statistic.

5.Decision-making support platform

Provide statistical results and data reports to all levels of tax department.

- System interface

Forming the unique entrance and the information collection interface through the browser.

1. United registering

Convenient access to every different system once logged in through the unique entrance.

2. Information collection

Display all the common information which is personal relevant on the unique interface.

3. Contents searching

Quick search of the contents by various conditions such as keywords or time.

4. Personality

Make the Individual desktop according to everyone's favor and need.

5. Unique message

Display all the messages from different systems or modules on a unique message platform enabling users informing in various ways.

6. Automatic remind

Users can be reminded by mobile phone short message once there is latest or important information even when they are not beside the computers.

Digital tax platform can turn to be "start point" or "window of information" by severing all level tax departments or officers. All the users equipped by this platform can get the access to different systems according to theirs authorities anytime when they log in. Efficiency also can be improved by the timely informing method.

3 Problems in the development of BizEngine

3.1 Integration of data

As a unique levy management system exists in the tax system, it has become critical to make all the data integrate and communicate with each other. This problem can be solved in BizEngine including "data integration engine" which consists by "data integration middleware" and "data-mapping administration".

Data integration engine enables different kinds of databases connecting, as well as different systems. Though this engine, process and database can be operated together in order to integrate numbers of systems. The functions of the engine are abundant: definition and operation of the data source and mapping, data fetching service, data synthetic searching, database supervision and event response.

The process can be summarized as follows:

1. Integrating the workflow and the use's system in time by "general data administration", enables all the data can be dealt with in the workflow system, and then, the operation result can be saved in database. That is to say, "general data administration" could combine the workflow and the event system.

2. Data source configuration: Set all the parameters, SQL sentences, searching conditions and saving process which are needed when connection.

3. Data-mapping configuration: Set the operations for database, including R/W of single record, add and rewrite records and multi-records reading.

Take the application in nation's tax collection as an example, "BizEngine data integration engine" gets the data from "tax levy system", then writes the data into "national tax information management system", implements the check for the mixed data from different systems by copying the data into "national tax performance management system" after combining with

other data in “national tax information management system”.

3.2 Security Protection

“Digital tax management system” is a centralized application system facing all level tax department, the functions only can be inspired when authorities are grading in different level of department. The structure of a city-level tax organ can be divided into three parts: city-level department, lower level department and the internal department. Meanwhile, authorities for users of different levels should also be set according to the features of taxation, which is summarized as “highest authority tax organ”.

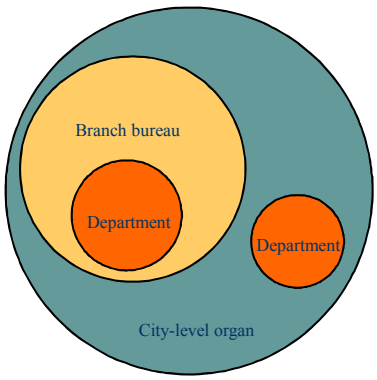


Figure 3 three parts of the structure

- BizEngine security platform

Integrated security management is one of the kernel functions of BizEngine application system. All the security functions can be found in “Integrated security platform”, meanwhile, it also can service various application systems.

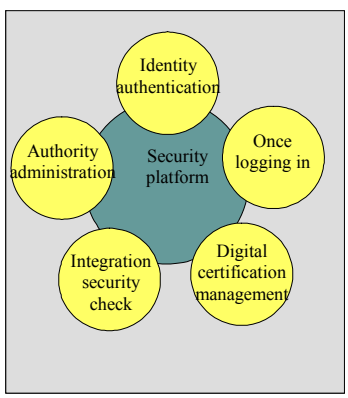


Figure 4 BizEngine security platform

The functions of “Integrated security platform” mainly including:

1. Identity authentication: authenticate the user’s log-in information, provide various kinds of validation, such as user name/password checking, digital certification, mobile phone short message checking.

2. Authority administration: provide the authority information as the users log in.

3. “Once logging-in”: only log in once makes access to all the sub-systems.

4. Integration security check: provide security log, including logging-in log, administrator log, system mistake log.

5. Digital certification management: unified distribution of the digital certification.

- Web authority management

It’s complex to manage the user’s authority in BizEngine support platform.

Each user owns two kinds authorities at one time:

1. Functional authority

Enable or disenable the access to certain function of the user. This process is accomplished by “role”. The “function” would be connected with the “user” through distributing the “function” to the “role”, then distributing the “role” to the “user”.

2. Data authority

Set the authority for pointed user, actor or department. Data authority is allocated by “access control list”. The authority of a security object to a data object is recorded in “access control list”. The connection of functional authority and data authority makes up a rigorous authority system, just like a network. The user would get his authority through network when he or she logs on the system. All his operation in the system will be restricted by this authority.

4 Conclusion

We have presented one kind of digital taxation management system based on BizEngine, and the settlement for the crucial problems in the development

of this system. Compared with the software produced by foreign countries with the alike function, BizEngine can perfectly satisfy our government's needs as it is produced just according to it. Equipped with visual tools for designing, users can easily implement all kinds of complex processes with a mouse click or a haul, but not coding. Workflow engine makes it possible for ordinary user accomplish the work which used to be done by professional coder, it also shorten the development process and reduces the cost.

Nowadays, simple server backup is still used commonly; system with high practicality which facing the crucial function has not been created yet. The middleware should open interfaces to improve the relation with other components. According to the success of various application, it can be concluded that this system which combining the experiences of different users can optimize the operation and predigest complexity efficient.

References

[1] Malcolm Chisholm, "How to Build a Business Rules Engine: Extending Application Functionality through Metadata Engineering," 1 edition, 2003

[2] Barbara C. McNurlin and Ralph H. Sprague, "Information Systems Management in Practice" 7th Edition, 2005

Mobile Telephone Learning Mode Research Based on 3G Technology

Shijue Zheng[1] Xiaoyan Chen[1] Tao Tao[2]

1 Department of Computer Science, Huazhong Normal University, Wuhan, Hubei, 430079,China

Email: zhengsj@mail.ccnu.edu.cn

2 Sanya Garrison, Sanya, Hainan, 572011

Email:xiaoyan0205@126.com

Abstract

3G stands for third generation, and is a wireless industry term for a collection of international standards and technologies aimed at increasing efficiency and improving the performance of mobile wireless networks.3G wireless services offer enhancements to current applications, including greater data speeds, increased capacity for voice and data and the advent of packet data networks versus today's switched networks. With 3 G times arrival, mobile phone study has accepted by people more and more .The paper introduce mobile learning(M-learning) and prove the feasibility of mobile telephone learning, advancing to combine 3G and mobile telephone in order to achieve a better study approach.

Keywords：M-learning; 3G technology; mobile telephone learning; study mod

1 Introduction

We are in the information society, life-long education has become the need of the modern society. In nervous and quick rhythm life, people hope that they can study at anywhere, anytime. With the appearance of 3G, M-learning becomes possibility, so learning based on mobile telephone has becoming people's attention hotspot.[1]

The availability of mobile and wireless devices is enabling different ways of communicating. Mobile communications are no longer restricted to companies that can afford large investment in hardware or specialised software. Individuals now have easy and inexpensive access to mobile telephony, and the cost of mobile access to the Internet is steadily reducing. Mobile technologies have enabled a new way of communicating, typified by young people, for whom mobile communications are part of normal daily interaction, who are 'always on' and connected to geographically-dispersed friendship groups in 'tribal' communities of interest[2].

M-learning is a new development stage of distance learning, the characteristic is to realize free learning at anytime, anywhere, anyone, any style(4A).

M-learning relies on the mature wireless moving network, the international and multimedia technology. Students and teachers use mobile equipment (such as portable type computer, PDA, mobile telephone and so on), [3]by the mobile teaching server, being to realize alternation teaching conveniently[1].

The paper discuss mobile telephone learning mode research based on 3G technology, the mobile telephone is not only a message querying and transmitting tool, but also is an important tool of quick learning., is the combinative M-learning system of communication system , blue-tooth technology and 3G portrait telephone.[4]

2 The Theory Basic of M-Learning

2.1 The definition of M-learning

M- learning is a new learning way that make use of wireless communication network technology and wireless mobile communication equipment(such as mobile telephone, personal digital assistant (PDA), Pocket PC and so on) to get education information, education resource and education service.

2.2 The basic structures of M-learning

The basic structures of M-learning are international net, mobile CERNET and mobile communication equipment. As fig.1 shown.

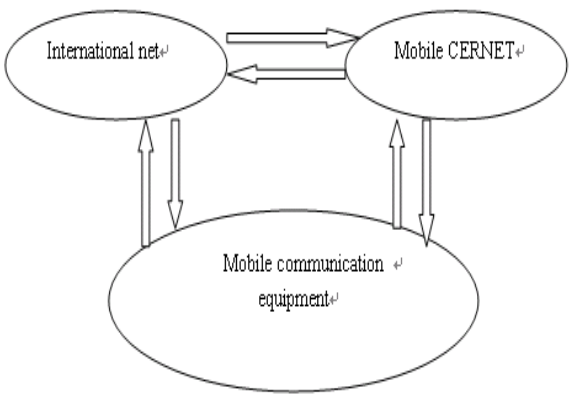


Figure 1 The basic structures of M-learning

3 M-Learning Based on Mobile Telephone

3.1 The feasibility of M-learning based on mobile telephone

3.1.1 Hardware condition

(1) The big dimension color screens technology is mature and universal

The mobile telephone does not always become important equipment, a important cause is that the screen dimension is not enough big. It is difficult to be satisfied with studying on effect in dimension and display. But with LCD price coming down, main current the mobile telephone screen adopt colour liquid crystal already broadly, now the mobile telephone hardware platform already is enough to study. CPU processes speed, memory size, multi-media effect, screen resolution ratio already reach pragmatic degree[5].

(2) High-capacity lightning memory's application and price are cheap

The mobile telephone runs dictionary, cartoon, video and so on, which needs the lightning memory consuming bigger memory space, therefore the mobile telephone must have high-capacity. But the price of 1 GB lightning memory is about 150 yuan now, has been ten times lower than 3 years ago, but the capability has been improved. Therefore, high-capacity memory is not the bottleneck which restricts the development of the mobile telephone already.

3.1.2 Software condition

(1) The kind of Medium learning software are increased

The mobile telephone platform based on learning software is very few, main concentrating on electron book, MP3, video and so on. The content of learning is very unitary, concentrating on English mainly. Fortunately ,some institutes start to concentrate efforts on the exploitation studying content and the software in mobile telephone , learning software has multi-media trend[6][7].

(2) Software is maturity and perfect

with the development of new technology , the content of learning is becoming multimedia –initialization, especially the proportion of cartoon and video will increase broadly. [8]The software suite mobile telephone is becoming mature and perfect.

3.2 The mode of M-learning based on mobile telephone

Mobile telephone has popularized in life already , people has been able to study at anywhere, anytime, the system structure of M-learning based on mobile telephone as fig.2 shown..

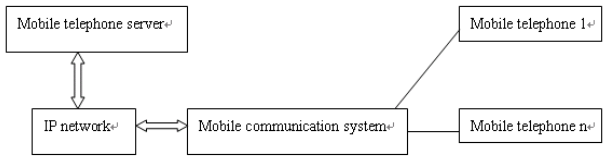


Figure 2 The system structure of M-learning based on mobile telephone

3.2.1 M-learning mode of short message server

M-learning based on short message is a simple and rapid learning way. Through short message, not only between users, but also between user and net server can transmit characters. Users transmit short messages to teaching server through mobile telephone, PDA, or other wireless equipment. Teaching server analyses user's short messages, then transforms them to data request, and analyzing data, then send them to users. [9]Make use of this characteristic, may realize communication between wireless mobile network and net, and accomplish some teaching activity. M-learning mode structure of short message server is shown as follow:

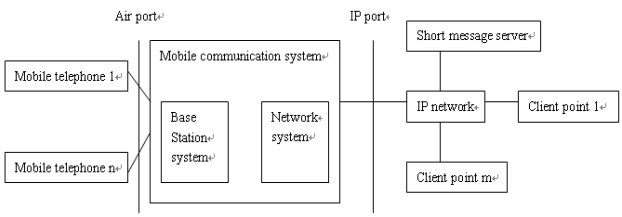


Figure 3 M-learning mode structure of short message server

3.2.2 Online real time interactive M-learning mode

Online real time interactive M-learning mode is to realize the alternation between teachers and students. Realizing common alternation between people are: mobile communication equipment, email, forum and mobile QQ. In learning, student often come across a lot of problems, now the alternation between the teachers and students is especially important, the teacher can provide various learning strategy for student, help students to know self studying style, finding the learning strategy that suite to individual need and developing. [10]The teacher needs to provide advice, auspice and encourages for students, helping students to resolve problems in learning.

3.2.3 M-learning mode of WAP browse service

WAP browsing, similar to the browser application of computer, supports XML as compatible HTML. At the same time, can use CSS, which improves the expressive power of contents. Besides, supports TCP/IP standard of Internet in protocol aspect. Generally, the page files on web server are stored with HTML form, so on line browse net information by mobile equipment, the HTML files must be converted to WML files.

With the development of communication, 3 G technologies appearing, applying 3 G technologies to mobile telephone, which take huge convenience for M-learning based on mobile telephone, its application has far-reaching significance.

4 3G Technology

3G (3 rd Generation) , Chinese meanings is the third generation mobile communicate . Third generation mobile communication system 3 G being able to provide the broadband information business , such as high speed data , image and television image and so on, the transmission rate is up to 2 M BIT/ S , bandwidth may reach more than 5 M HZ. The main distinctions between the third generation mobile communication and the former two generations are voice and speed, it is able to handle various medium forms such as image, music, video, provide various information such as teleconference, Electronic Commerce serves.

4.1 The application of 3G technology in M-learning based on mobile telephone

3 G is used for mobile telephone learning, it can as a tool to improve the attention to education, assigning more resource for education, cultivating more r talented person. 3 G provides various broadband information business will make mobile education system to provide

more convenient service for a user, its application shows as follow:

4.1.1 Providing mobile net

Mobile Internet is a network that integrate wired Internet , mobile communication net , it include wired Internet , wireless connection and nimble terminal mainly. Wired Internet is Internet; Wireless connection is the network being composed of wireless module; The nimble terminal points to the end instrument having terminal treatment ability. Compared with wired Internet, mobile Internet have made up the shortcoming that wired Internet not able to move. Mobile Internet being able to contact every person in school ,it has broken close learning space, making students get a space that can develop their individuality, classroom has been expanded by boundless field in theory. As long as places are coveraged by the communicating signal, can actualize mobile teaching. Mobile users use wireless terminal to connect Internet, visit teaching server and browse, query, real time alternation, similar to general Internet users.

4.1.2 Mobile resource

Teachers can collect the experiment, plotting, sound, portrait and large amount of data from Internet, searches out required material, making use of again, process to required class of teaching resource, build the various data base on the school education platform. Because of 3 G data transmission rates are very high, students can collect required contents by mobile network, such as the emphases, difficulties and general problems of course. Students can query various problems and not be restricted by time and place.

4.1.3 Mobile discussion and bbs

The bbs users can query the title and content of bbs by mobile telephone, can publish papers by mobile telephone, can make use of hyper text connection way to make users discuss the problems they are interested in. This discussion can be celebrated at anytime, anywhere, the exchanges between teacher and teacher, teacher and student, student and student have increased greatly. Students not only can communicate with teachers in school, but also can communicate with famous expert of home and abroad.

4.1.4 Mobile multimedia

Multimedia has gathered two kinds or more than two kinds mediums module, such as voice, data , image , video. In 3 G system, can actualize data, video multimedia communication between end users, which make “face-to-face” communication at anytime, anywhere becoming possibility. Supposed that student can see teacher’s course not in classroom, can carry out exchange through their computer or mobile telephone, teacher can guide them to study. In addition to watch the lecture, students are able to use 3G mobile telephone to handle equipment, use mobile telephone to carry out surveillance and control.

5 Conclusion

M-learning based on 3G technology has becoming a convenient rapid study way, Its advantages, becoming an important study way, and it has broad application prospect. But we must realize M-learning based on 3G technology is only an important study way, is the complement of other study way, it can not replace the traditional book and classroom study.

References

[1] Keegan D.”The future of learning:From e- Learning to m-Learning” [EB/OL], http:/ / learning.ericsson.net/ leonard- o/ thebook/ book.html

[2] Aleksander Dyeet al.”Mobile_Education- A Glance at The Future”[EB/OL], http:/ /www.nettskolen.com/ forskning/mobile_education.pdf

[3] Clark Quinn.”M- Learning:Mobile,Wireless, In- Your-Pocket”.[EB/OL].http:/ / www.linezine.com/ 2.1/ features/ cqmmwiyp.htm

[4] Paul Harris.Goin Mobile[EB/OL], http:/ /www.learningcircuits.org/2001/ jul2001/ harris.html

[5] Houser, C., Thornton, P. and Kluge, D. “Mobile learning: cell

phones and PDAs for education " ,International Conference on Computers in Education,vol.2, pp.1149- 1150, 2002

[6] Rung-Ching Chen, Ya-Ching Lee and Ren-Hao Pan" ADDING NEW CONCEPTS ON THE DOMAIN ONTOLOGY BASED ON SEMANTIC SIMILARITY", Department of Information Management, Chaoyang University of Technology168, Jifong East Road, Wufong Township, Taichung County 41349, Taiwan ROC, 2006

[7] Thornton, P. Houser, C. , "Using mobile phones in education ",The 2nd IEEE International workshop on Wireless and Mobile Technologies in Education", pp.3-10, 2004

[8] Alberto Apostolico, Ricardo Baeza-Yates and Massimo Melucci "Advances in information retrieval: an introduction to the special issue", Information Systems Vol.31, PP. 569 – 572

[9] C. Y Chang, Sheu, J. P., and Chan, T. W" Concept and design of Ad Hoc and mobile classrooms", computer assisted Learning, vol.19, pp.336-346,2003

[10] Chi-Hong Leung and Yuen-Yan Chan, "Mobile Learning: A New Paradigm in Electronic Learning ",The 3rd IEEE International Conference on Advanced Learning Technologies,pp76- 80, 2003

Application Level Multicast Routing Algorithm Based on Clone Selection Strategy

Dezhi Wang[1] Jinying Gan[2] Xinwei Cui[1] Deyu Wang[2]

1 Department of Computer and Science, North China Institute of Science and Technology

2 Department of Electrical & Information Science Engineering, North China Institute of Science and Technology
Beijing, 101601, China

Email：wangdz20017@sohu.com; wangdz@ncist.edu.cn

Abstract

The application level multicast tree is constructed on a suppositional overlay network which can be presented as a complete graph. So finding an optimization multicast routing tree with degree constraint is an optimization problem, which can be formulated as minimum average cost spanning tree with a bound problem and is NP-hard. A new multicast routing algorithm based on clone selection strategy is presented to solve the problem. It utilizes the mutation characteristic to find an optimization solution for the global solution scope. The simulation results show that the algorithm has a faster converging speed and better ability of global searching with property of stabilization and agility

Keywords：Application level; Multicast; Clone selection strategy; Routing algorithm

1 Introduction

Over the years, the multicast is becoming an important technology of the network communication, with the fast developing network application such as VOD, video meeting, computer cooperate and so on. It is an effective communication way for the computer network from one point or many points to many points. It utilizes a tree structure to deliver data from senders to a group of receivers, with packets only replicated at branching nodes in the tree [1]. Although the conception of IP multicast is simply and performance is effective, there is not global multicast foundation architecture except MBone [2]. IP multicast has not an availability scheme for group address and must update all the routers to support IP multicast in foundation network. These problems make IP multicast lack expansibility and configuration. So in recent years, the application level multicast (ALM) is presented for improving multicast application. It is soon becoming a hot research issue. The ALM adopts the Internet as the based structure to provide multicast services for end hosts and all multicast packets of ALM transmit in unicast. In the ALM, the end hosts organize an overlay network on the base physical network and multicast packets distribute in the multicast tree on the overlay network. Comparing with IP multicast, the ALM has easy deployment characteristic and solves the address problems. These increase the incentive for the network operators to enable multicast application.

To support effective multicast on the application layer, designing a better routing algorithm is essential. Currently, various projects [3-6] of the ALM are implemented for different application objects and the typical schemes are the Narada, Scatercast, Overcast and ALMI. Among of them, the Narada and Scattercast want each member of the multicast group has the least delay. And the object of the Overcast is to let the each member of the multicast group has the most bandwidth used. The ALMI try to improve the using ratio of the network with reducing cost of the system.

Because the capability of the application multicast nodes is limited, the construction of the multicast tree

must have node degree constraint. The least cost tree problem with degree constraint is an NP-hard problem. The traditional gene algorithm is effective for NP-hard problem [0]. But for the least cost tree problem with degree constrain is not very effective with the expanding of the network scope. The computing quantity quickly increasing leads computing slowly. And in some extreme condition the algorithm constringency can not arrive.

For solving those problems, we put forward a more logical model of the multicast tree and research the application level multicast algorithm based on the model in this paper. And an application level multicast routing algorithm based on clone selection strategy was advanced. The algorithm adapts to the multicast application such as the VOD and remote education, which have only one source and is sensitive to delay

2 Clone selection strategy

The Genetic Algorithm (GA) simulates the action of biologic evolution process with breed, mutation, competition and selection to achieve the function. The GA emphasizes the survival of the fittest in the groups and production of the superiority members. The classical GA has three operators: selection operator, cross operator and mutation operator. But the traditional GA strategy dose not comprise the cross operator, and depend on the selection and mutation operators to implement the genetic evolution process. Because the GA dose not has the feedback mechanism in the evolution search process, and the function of all operators is guided with the simple probability. The course of the searching for better members can present phenomena of the degeneration and precocity or slowly convergence speed. In allusion to above problems, the clone selection strategy is advanced. The fundament of the clone selection strategy is that the clone selection and mutation operators replace the selection and mutation operators of the GA strategy, and imports the feedback mechanism to delete the degenerative individuals and supply new individuals [0]. Based on the above process, the purpose of researching for the best result of the clone selection strategy is becoming more clearly, and the convergence rapidity will increase. The clone strategy can avoid the phenomena of the degeneration and precocity in the GA [0]. So it can be used to solve the complicated problems such as NP-hard problems.

The substance of the clone selection strategy is that a mutated solution colony is created nearby the best solution based on the affinity degree of the solution. This method expands the search space for the best solution, and helps the algorithm avoids the evolution precocity phenomenon and search plunging in the local least solution

3 ALM routing problem

The traditional IP multicast routing links are the real physical links from one router to another router. But the application level multicast is implemented on the suppositional overlay network, which is composed by the hosts. The links between the nodes of the overlay network are not real links. Because the nods of an overlay network are hosts, they can contact each other on the foundation network. So constructing an ALM tree on an overlay network is a problem creating a tree on a Complete Graph.

The overlay network is modeled as an undirected complete graph $G=(V,E)$, where V is the set of the source and all destination nodes of the multicast dialog, and E is the set of the suppositional links on the overlay network. So the application multicast tree is $T=(V_T,E_T)$, where $V_T \subset V$ is a set of multicast nodes and $E_T \subset E$ is a set of links in T. Let $\forall e \in E$ denote an overlay link, and $l(e)$ is the cost of the link e.Because a host has limited computing capability and bandwidth, each overlay node only connects a few another nodes. So let $d_{\max}(v) \in N \quad (v \in V)$ is a degree constraint of the node v, and $d_T(v)$ is the real connecting nodes number. It constrains a node to connect other nodes. Let $P_T(s,v)$ is a path from the source node s to the destination node v, and the cost of the path $P_T(s,v)$ is

$$Cost(s,v)=\sum_{e\in Pr(s,v)} l(e).$$

Based on the network model, the application level multicast tree problem can be described as a degree-bounded minimum average cost spanning problem (DMACS). For given undirected complete graph $G=(V,E)$ and cost function $Cost(s,v)\in R^{+}$, the minimum average cost spanning tree (MST) problem [14] without degree constraint can be formulated as follows:

$$T^{*}=Min((\sum_{v\in V} Cost(s,v))/|V|) \qquad (1)$$

By extending the MST with some punished functions, we can get a general minimum average cost spanning tree problem (GMST). The characteristic of the GMST is that it considers the degree constraint has some cost which can be computed as punished function. On the tree, the degree of a leaf is 1 and the degree of a non-leaf node is bigger than 1. On a tree with n nodes, the degree of a node is $n-1$. Let the degree weight function is $\xi=\xi(d),d=1,2,\cdots n-1$. Normally, the function ξ is a monotonously non-degressive function for the variable d. Based on GMST and the formula (1), the DMACS problem is formulated as follows:

$$T^{*}=Min((\sum_{v\in V} Cost(s,v))/|V|+\sum_{k=1}^{n}\xi(d_T(k))) \qquad (2)$$

In the function, the n is the sum of the network nodes. For ensuring the reasonability of the multicast tree, while the degree constraint is satisfied with $d_T(v)\le d_{\max}(v)$, the value of the function $\xi(d_T(k))$ is much less. Otherwise, while the degree constraint is not satisfied as $d_T(v)>d_{\max}(v)$, the value of the function is much bigger.

Obviously, the DMACS problem defined above is a multi-objective optimization problem, which is NP-hard due to the combinatorial explosion. The normal algorithm can not tackle this difficult problem. But we can use the clone selection strategy algorithm to solve the problem.

4 Clone strategy multicast routing algorithm (CSMA)

The goal of our algorithm is to find out the least cost of the application level multicast tree with a degree constraint. For a better application of the clone strategy to solve the DMACS problem, we need deeply research the coding of an antibody and adapt an optimized coding mode. The affinity degree function also should be simple and reasonable for computing. Now we will present concrete steps of the algorithm.

（1）Antibody coding and colony initialization

The multicast tree is constructed on an overlay network which can be formulated as a complete graph. The Prüfer code is just for a tree on the complete graph, so that we adopt this code as the code for the antibody code. By the Prüfer code, the tree is presented as a rank with $n-2$ numbers. Decoding is a contrary step to coding. The detailed steps of the coding are denoted as follows.

Step1: All tree nodes of the overlay network are numbered with a sequence natural number.

Step2: Select the leaf i node with the least number, and then find the only father j node connecting the i node.

Step3: The father j node is coded as the first number of the Prüfer code. And the link e_{ij} and the i node are deleted. So the tree remains $n-1$ nodes.

Step4: The iterative process finishes until the tree having only last one link.

From the process an antibody is coded with $n-2$ numbers. And the iterative process is repeated N times, so that the random initialization colony set A_0 is generated with the size N. Each antibody of the set presents an application multicast tree.

（2）Affinity degree computing

Because the solution of the DMACS is the minimal value, the formula (2) can be transformed into the maximal value problem and the formula (2) becomes the affinity degree function (3).

$$f(T)=\frac{1}{(\sum_{v\in V} Delay(s,v))/|V|+\sum_{k=1}^{n}\xi(d_T(k))} \qquad (3)$$

Based on the formula (3), the affinity degree value of an initialization antibody is computed. The antibodies are ranked with the affinity degree and the set A is gained from the set A_0, which is a set $A=\{a_1,a_2,\cdots a_N\}$. In the set A, each member has $f(a_i)\ge f(a_{i+1})$ and $i=1,2,\cdots N-1$. The function $f(*)$ is the affinity degree

function.

（3）Clone

Each antibody a_i of the colony set A is cloned with the size $k_i = Int(\beta N / i)$. So that the new antibody colony set $C = \{a_{11}, a_{12}, \cdots a_{1k_1}; a_{21}, a_{22}, \cdots a_{2k_2}; \cdots; a_{N1}, a_{N2}, \cdots a_{Nk_N}\}$ is generated with the size N_c, ($N_c = \sum_{i=1}^{N} Int(\beta N/i)$, $N_c \geq N$). The invariant β is a clone modulus, which controls the scope of the clone. The a_{ij} is a copy of a_i. From the clone scope, the antibody with higher affinity degree has a more scope, and the genes of the higher affinity degree antibody can be reserved and improve in the algorithm.

（4）Mutation

The antibodies of the colony set C are mutated with the mutation probability P_c which is got by experience. For simply operating process, the algorithm adopts the location mutation which random selects an integer $a \in [1, N]$ replacing a gene value in an antibody with the probability P_c. Then the new antibody colony set C' is generated with the size N_c.

（5）Selection

The antibodies in the older colony set A and the new mutated colony set C' are compared and selected based on the affinity degree value. Then a new generation colony set A' are composed with the size N.

（6）Antibody variety supplement

For preserving the variety of the antibody colony, the colony set B with the size $N_b (N_b \geq N)$ is random generated. From the colony set B, the antibodies with the higher affinity degree replace the lower antibodies in the set A'.

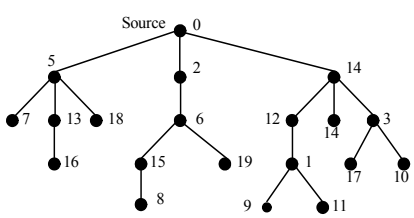


Figure1 An optimization multicast tree of 20 network nodes

（7）End iterative process

If the new generation colony set A' has the optimization antibody, the iterative process ends and put out the final solution. Otherwise, the computing returns to (2) step for another iterative process until the maximal iterative step T is arrived.

The algorithm complexity associates with the colony size N, N_c and iterative step T, so that the algorithm complexity is $O(NN_cT)$.

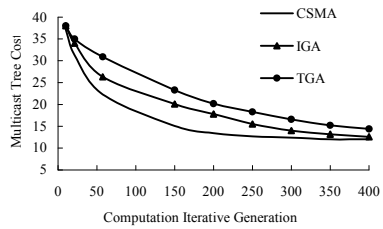


Figure2 Multicast tree cost of 20 nodes network

5 Performance evaluation

In this section, we discuss the computational experiment of the algorithm based on clone strategy for DMACS problem. For performance evaluation of the algorithm, we also simulate the traditional genetic algorithm (TGA), the improved genetic algorithm (IGA) 0 and this paper algorithm CSMA. The multicast network is denoted by a random created complete graph. The link cost of the network is a random integer in a bound, and the degree constraint also is a random integer in the bound $[2, n/2]$. The clone strategy mutation probability of the CSMA is 0.6, and the cross and mutation probability of the TGA and IGA are 0.6 and 0.01. For CSMA the clone coefficient β is 0.9, which is got by the multiple experiment results and experience. The initialization colony is lesser than the cloned colony for CSMA. For equitable experiment, the cloned colony size of the CSMA is same as the size of the TGA and IGA.

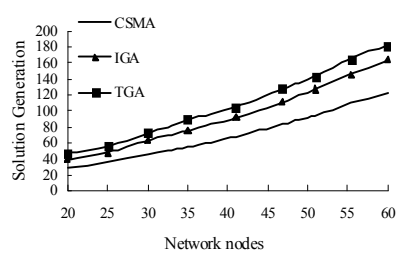


Figure3 Solution generation variation with the different network node

Table 1 The results comparison of three algorithm

Size	CSMA		IGA		TGA	
	FSG	SAG	FSG	SAG	FSG	SAG
20	22	28.76	25	34.58	50	62.54
30	34	41.52	43	52.35	76	91.34
40	52	62.38	67	79.45	98	118.58
50	81	103.14	103	121.78	137	156.78
60	115	130.62	146	171.82	189	221.45

The talbe1 shows the results of the proposed algorithm in 20 to 60 nodes multicast network with the first optimization solution generation (FSG) and optimization solution average generation (SAG). Each experiment is simulated 20 times and average solution is as the final result. From the table1, the CSMA has a better performance with the increase of the node size. Compared with IGA and TGA, the CSMA can faster find the optimization solution. And the sum of the generation with the optimization solution is much more than the other two algorithms. So the CSMA has a much faster constringency speed than the IGA and TGA, and can ensure the computing find the optimization solution in the end

The computation result is further investigated by analyzing the 20 nodes multicast network. Fig. 1 is the optimization application level multicast tree based on the CSMA. From the multicast tree structure, the nodes are reasonably connected and accord with the degree constraint of the node. So the CSMA can get an optimization solution.

Further experiments are performed to examine the astringency of the CSMA with 50 nods in the network. Fig.2 shows the relation between the optimization solution and the iterative step number. It is clear that the CSMA can faster find the optimization solution and cost of the multicast tree is less than two other algorithms. Fig.3 shows that the different algorithm iterative generations varies with the node number variety from 20 to 60.The CSMA has less generation number with the nodes of the network increasing than TGA and IGA. The CSMA has faster constringency speed. The causation of the CSMA better performance is that cloning operation expands the area of solutions near the optimization solution, so that the optimization solution can faster found and stop the iterative process

6 Conclusion

The multicast tree is constructed on an overlay network, which can be present as a complete graph. So the multicast routing problem with a degree constraint becomes a finding a minimum average cost spanning tree problem, which is an optimization problem and NP-hard. A clone selection strategy algorithm is based on the antibody cloned selection of the biology. It can solve the complicated optimization problem by reasonable clone selection operation. The algorithm CSMA can solve the application multicast routing problem with degree constraint.

In simulation, the TGA, IGA and CSMA are respectively computed for finding an optimization multicast routing tree. The results of the simulation show that the CSMA has a faster constringency speed than the genetic algorithm, and can also overcome the precocity of the genetic algorithm.

Our work is just beginning, only limited testing and simulation have done, the design needs to be validated in using. It is necessary to do more optimize for some multi-objective problem during the process of looking for the best multicast trees in the future.

References

[1] C.K. Yeo, B.S. Lee, and M.h. Er, “A framework for multicast video streaming over IP networks”, Networks Compute. Rev. Lett., Vol.26, pp. 273-289, 2003

[2] K. Almeroth. “The evolution of multicast:from the MBone to inter-domain multicast to Internet2 deployment”, IEEE Network. Rev. Lett. Vol.14, pp 10-20, 2000

[3] Jannotti J, Gifford D, Johnson K. “Overcast: Reliable Multicasting with an Overlay Network”, in Pro. Proceedings of the Fourth Symposium on Operation Systems Design and Implementation, San Diego, CA, pp. 197-212, 2000

[4] D. Pendarakis, S. Shi, D. Verma, M. Waldvogel, “ALMI: an Application Level Multicast Infrastructure”, in Pro. Proceedings of the Third Usenix Symposium on Internet Technologies and Systems (USITS), March 2001

[5] Y. D. Chawathe, “Scattercast: an architecture for Internet broadcast distribution as on infrastructure service”, PhD

Thesis. Stanford University, September 2000

[6] L.Lao, J.H.Cui, M.Gerla. "Toma: A viable solution for large-scale multicast service support", in Pro. IFIP Networking, May 2005

[7] Xiawei Z, Changjia C, Gang Z. "A Genetic Algorithm for Multicasting Routing Problem", in Pro. International Conference Communication Technology Proceedings, WCC-ICCT 2000, pp.1248-1253, 2000

[8] Leandro N De Castro, Fernando J Von Zuben. "Learning and Optimization Using the Clonal Selection Principle", IEEE Transactions on Evolutionary Computation, Rev. Lett, Vol. 6, pp 239-251, 2002

[9] M. Balazinska, E. Merlo, M. Dagenais et al. "Advanced clone-analysis to support object-oriented system refactoring", in Pro. The 7th Working Conference on Reverse Engineering, Brisbane, Australia, 2000

[10] A. T. Haghighat, K. Faez, M. Dehghan, A. Mowlaei and Y. Ghahremani. GA-based heuristic algorithms for bandwidth -delay-constrained least-cost multicast routing. Computer Communications, 2004, 27(1): 111-127

[11] Li Lao, Jun-Hong Cui, Mario Gerla. Multicast Service Overlay Design. in Proceedings of International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'05), Philadelphia, Pennsylvania, July 2005

OverView of Adaptive Mobile Network Network of 4G system*

Cheng Chuanqing[1] Cheng Chuanhui[2] Qing Xiuhua[3]

1 Department of Computer Science, Wuhan University of Science and Engineering, Wuhan, Hubei, 430074, China
Email: ccqcjl2005@126.com

2 Zhongnan University of Economics and Law, Wuhan, Hubei, 430074, China
Email: sammicch@hotmail.com

3 Departmetn of Mathematics and Physics, Wuhan University of Science and Engineering,Wuhan, Hubei, 430074, China
Email: wl3833@126.com

Abstract

Efficiency and adaptivity play a major role in the design of fourth-generation (4G) wireless systems. 4G systems should be bandwidth efficient, power efficient, and allow for low complexity transceivers. The fourth Generation mobile network include many information and is a certain path to future mobile communication system .This paper discusses the overview on adaptive mobile network of 4G systems. 4G network structure, adaptive network of 4G system, compare of 3G and 4G are discussed in this paper Research status in quo also introduced in the paper.

Keywords: mobile network adaptive 4G system

1 Introduction

The International Telecommunication Union (ITU) recently defined recommendations for mobile communication systems beyond the third-generation (3G)]. In these recommendations, data rates of up to 100 Mbps for high mobility and up to 1Gbps for low mobility or local wireless access are predicted. Systems fulfilling these requirements are usually considered as fourth- generation (4G) systems.4G systems are expected to achieve the following goals:

- Quality of service (QoS) provisioning: Packet-based services (probably using mobile IPv6) will play a major role. Low delays are needed (for example for streaming content), which results in the need for short block sizes. Parallel data streams with different QoS classes should be supported to allow for hierarchical source coding, video streaming, parallel support of different applications, etc.
- System scalability: The system should be scalable with respect to data rates (with high peak rates up to 100 Mbps for high mobility and 1 Gbps for low mobility or local wireless acces), bandwidth (about 20 MHz to 100 MHz in 5 MHz slots), and cell sizes (including pico, micro, and macro cells)
- High efficiency including bandwidth efficiency (about 4-10 bits/s/Hz) and simultaneously high power efficiency
- High-speed mobility, high velocities, fast adaptation
- Low computational complexity (iterative processing). MHz in 5 MHz slots), and cell sizes (including pico, micro, and macro cells)
- High efficiency including bandwidth efficiency (about 4-10 bits/s/Hz) and simultaneously high power efficiency
- High-speed mobility, high velocities, fast adaptation
- Low computational complexity (iterative processing).

* This work is supported by Hubei Prrovince Nature Science Foundation under Grant 2006ABA296

During the last few decades, mobile communication have developed rapidly. The 1G was based on analogy technique and deployed in the 1980s. Speech chat was the only service of 1G. 2G was based on digital signal processing techniques and regarded as a revolution from analogy to digital technology, which has gained tremendous success during 1990s with GSM as the representative. The utilization of SIM (Subscriber Identity Module) cards and support capabilities for a large number of users were 2G's main contributions2.5G extended the 2G with data service and packet switching methods, and it was regarded as 3G services for 2G networks. Under the same networks with 2G, 2.5G brought the Internet into mobile personal communications. This was a revolutionary concept leading to hybrid communications.3G is deploying a new system with new services instead of only providing higher data rate and broader bandwidth. Based on intelligent DSP techniques, various multimedia data communications services are transmitted by convergent 3G networks.

2 Some Technique

2.1 End to end QoS

Developers need to do much more work to address end-to-end QoS. They may need to modify many existing QoS schemes, including admission control, dynamic resource reservation, and QoS renegotiation to support 4G users' diverse QoS requirements. The overhead of implementing these QoS schemes at different levels requires careful evaluation.A wireless network could make its current QoS information available to all other wireless networks in either a distributed or centralized fashion so they can effectively use the available network resources. Additionally, deploying a global QoS scheme may support the diverse requirements of users with different mobility patterns. The effect of implementing a single QoS scheme across the networks instead of relying on each network's QoS scheme requires study.

2.2 Security strategy

The security strategy of current 3G system is mainly to implement safe communication among users, and is designed to special service type ,so there is security defect e.g. none of user digital signature, the production mechanism of key is less safe, too more algorithm, authentication protocol is easy to be impacted. There are following security threats in 3G

- .illegal get the sensitive data, to attack system information
- Illegal operation to sensitive data, to attack the data integration, include: change information, insert, repeat or delete.
- Interference of network service, lead the system refuse or decrease the service quality.
- To deny the action ever do.
- Illegal access of service

4G is a isomerism network and various services, current 3G security strategy is not fir for 4G network Moreover, the length of encryption/decryption key is fixed, which is not fit for different service to have different security demand .Now, some researcher have progressed aimed 4G security .Now the overall direction of 4G security system is determined basically, such as mobile IPv6, demanding of authentication, authorizing, audit and Accounting. Future 4G will produce a multiple, can re-configuration security strategy and the enhance mobile device security performance.

2.3 Adaptive 4G network

Future 4G network is a adaptive network. More advances in networks are needed to keep pace with the rapidly changing terminals and applications, as follows:

Smart antenna, software radio, together with advanced base station are the key techniques to achieve adaptability of wireless access points to diverse terminals, i.e. to make radio systems and air networks re-configurable.

- Hierarchical and ubiquitous as well as overlay cellular systems, including picocell, microcell, macrocell, and magecell ones, implement seamless network interconnection of both

symmetric and asymmetric nature, and seamless terminal handoff of both horizontal and vertical levels respectively.

- Network layer hierarchical mobility management based on Mobile IPv6 and Cellular IP brings quick and seamless handoff to terminals. The Mobile IPv6 also presents a great contribution to the adaptability of heterogeneous networks.
- Ad hoc wireless networks are a kind of self-deployed wireless networks to make networks portable and adaptable, and thus dynamically share unlicensed radio spectrum.
- Network reconfiguration can be obtained by there configuration of protocol stacks and programmability of network nodes. Thus, it can adapt dynamically to the changing channel conditions and low or high data rate users.
- Miscellaneous services can be delivered through a mixture of transmission networks including unicast, multicast, and broadcast ones. According to the service types, e.g. real-time attribute, importance, bandwidth demand, or data stream type, multiple levels of QoS can be defined for various services.
- Network resource can be dynamically allocated to cope with varying traffic load, channel condition, and service environment. Traffic conditions will be dynamically monitored and controlled via techniques such as distributed and decentralized control of network functionalities.

2.4 Hybrid network in 4G

Hybrid of communication network, computer data network, broadcasting/TV network is a final target of next generation network. So 4G is probably a hybrid wireless network of cell mobile network, WLAN, and TV network.

If it is said that NGN will be the infrastructure, 4G will be the access network of NGN, 4G will supply perfect information service just like computer data network. The computer network deploy packet-switch protocol, which help to the hybrid network .But some mobile communication concept e.g hand off ,in interference among regions.

Digital TV supplied by broadcast/TV network, can bring ordinary free TV service as well as paying TV service. Moreover, VoD and information service will be widely applied. It has much similarity to current 3G system or future 4G system.

The public load network is the base of hybrid, Voice, data, image, video and etc, so many services have public load network-internet will be the base of the hybrid network.

4G which is a All-IP network should have the ability to do any want to do, to applied in any where possible to be applied. As a mobile NGI, it can be a access network of NGN.

3 Comparison Of 3G And 4G

With the pressing demand of high rate mobile data service and mobile IP service, mobile communication system must have the ability of High data rate transport. Now there are some 3G standards which is recognized by ITU, such as WCDMA, CDMA2000, and D-SCDMA .Though the 3G have some advantages with the comparison to 2G, there are some defects which can not be satisfied by people.

(1) Insufficiency of supporting ip protocol.

(2) un-flexible of service supplying and service management

(3) Difficulty in continuously increasing bandwidth and high data rate to meet multimedia services requirements

3.1 Network structure

3G network include three parts which are core network, access network and user device. Core network is responsible for centre switch, data transport and service supplying .It can be divided into circuit switching network and packet network. Access network is responsible for connect mobile user to core

network.3G can deploy different access network to the same core network .The user device is the mobile terminal to get service, which connect to access network via standard wireless interface.

4G will deploy single global cell core network to replace the cell network of 3G.Full IP will be applied in the network, just like figure1. Core network can support different access method, just like IEEE 802.11a, WCDMA, BlueTooth, HyperLAN. At the same time, the user device have the exclusive recognizable code, which can co-operate among isomerism system via hierarchies structure. The structure can make multiple services connect to IP core network transparently and have better commonality and extensibility. The network structure is just like Figure 1.

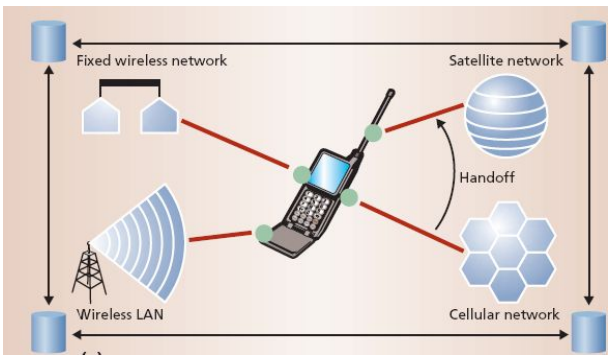


Figure 1 The network structure

3.2 Core network

4G mobile communication system is a full IP network. It has load mechanism based on IP, network maintenance and management based on IP, control of network source based on IP. application services based on IP and etc.

Core network is independent of concrete wireless access network, can supply end-to-end IP service and can be compatible with current core network and PSTN. Core network have open structure, can differ service/control/transport/. Based on IP, the wireless access method and protocol is independent from core network, protocol and link layer. IP is compatible with more than one wireless access protocol, so it is very flexible when designing the core network and does not need to take into account which method or protocol. should be used in wireless access.

3.3 Mobile terminal

The main character of 4G system lead to there is great difference between 3G and 4G. Future 4G mobile terminal should have following features:

- More strong interaction performance(more convenient to connect to network)
- More high network connectivity(mobile device can build up by ad hoc)
- Plenty of individuation. service(can support cell phone,, GPS location and etc.)
- Individuation. self- reconstruction ability(can change service demand and network condition adaptively.)
- Enhance security guarantee
- Enhance speech recognition function

More over, the form of 4G terminal is more various。A watch even probably can be a 4G terminal. But in 3G system, only cell phone can be a terminal. On the whole, 4G mobile terminal is different from 3G terminal because of the character of 4G system. 4G system can satisfy the requirement of high rate and wide bandwidth. The terminal also must assure it can gear to different air interface and different QoS label and mobility. To make compatible with different air interface, mobile terminal must have software refresh ability.

3.4 Core technique

In 4G, the core technology is OFDM but in 3G is CDMA.

Future wireless multimedia service have high demand on both data transport rate and the transport quality. So the modulation technique must have higher cell rate as well as longer code element periodic.

OFDM is the Short for Orthogonal Frequency Division Multiplexing, an FDM modulation technique for transmitting large amounts of digital data over a radio wave. OFDM works by splitting the radio signal into multiple smaller sub-signals that are then

transmitted simultaneously at different frequencies to the receiver. OFDM reduces the amount of crosstalk in signal transmissions. 802.11a WLAN, 802.16 and WiMAX technologies use OFDM.

Besides OFDM, there is W-OFDM technique, which enables data to be encoded on multiple high-speed radio frequencies concurrently. This allows for greater security, increased amounts of data being sent, and the industries most efficient use of bandwidth. W-OFDM enables the implementation of low power multipoint RF networks that minimize interference with adjacent networks. This enables independent channels to operate within the same band allowing multipoint networks and point-to-point backbone systems to be overlaid in the same frequency band.

4 Conclusion

This paper discussed the feature and core technology of 4G mobile system and introduced the research status in quo. With the progress of research, the 4G will be close with us. The 4G mobile system has high data rate, high spectrum utilization ratio, low transmitting power, supporting flexible service, so it will be the certain path to the future radio and mobile communication system

References

[1] Zhang Jian, The Development Trends of 4G Technology, GUANGDONG COMMUNICATION TECHNOLOGY, 2004

[2] L. Ping, L. Liu, K. Wu, and W. Leung, “A unified approach to multiuser detection and space-time coding with low complexity and nearly optimalperformance,” in Proc. 40th Allerton Conference on Communication, Control, and Computing, Monticelli, Illlinois, Oct. 2002

[3] Jun-zhao Sun, .Features in Future: 4G Visions from A Technical Perspective[C].IEEE Global Telecommunications Conference 2001.Vol 6

[4] V. Gazis, “Evolving Perspectives of 4th GenerationMobile Communication Systems,” IEEE PIMRC 2002,Coimbra, Portugal, Sept. 2002

[5] S. Verd´u and S. Shamai, “Spectral efficiency of CDMA with random spreading,” IEEE Trans. Inform. Theory, vol. 45, no. 2, pp. 622-640, Mar. 1999

[6] T. B. Zahariadis et al., “Global Roaming in Next-GenerationNetworks,” IEEE Commun. Mag., no. 2, Feb. 2002,pp. 145-51

[7] A. Viterbi, “Very low rate convolutional codes for maximum theoretical performance of spread-spectrum multiple-access channels,” IEEE J. Select. Areas Commun., vol. 8, no. 4, pp. 641-649, May 1990

[8] X. Ma and L. Ping, “Coded modulation using superimposed binary codes,” IEEE Trans. Inform. Theory, vol. 50, no. 12, pp. 3331-3343, Dec. 2004

[9] Li Weiwei, Comparison and Transition of Key Technologies on 3Gand 4G, GUANGDONG COMMUNICATION TECHNOLOGY, 2004

[10] Fenner W. Internet group management protocol. Version 2, RFC 2236, Xerox PARC, 1997

An Improved Clustering Algorithm for Wireless Sensor Networks

Pingping Wang[1] Shangping Dai[1] Yajing Shan[1] Ping Zhang[2]

1 Department of Computer Science, Huazhong Normal University, Wuhan, 430079, China

Email: wpp84629@yahoo.com.cn

2 Foreign Languages College, Zhongnan University of Economics and Law, Wuhan, 430073, China

Abstract

Energy efficiency is an important design issue that can prolong the effective lifetime of a network with a limited energy supply. Clustering technique has been proven to be an effective approach for reducing energy consumption. It also can increase the scalability and lifetime of the network. In this paper, we propose an improved cluster formation algorithm for wireless sensor networks according to considering the energy as an optimization parameter. Compared to the algorithm in Ref[2], the ACE-CILP algorithm increase the cluster head election mechanism, and the simulation results show that ACE-CILP algorithm achieves its intention of consuming less energy, equalizing the energy consumption of all the nodes, as well as extending the network lifetime perfectly.

Keywords: Wireless sensor nodes; Cluster; energy efficiency; network lifetime

1 Introduction

Wireless sensor networks have attracted much research attention in recent years and can be used in many different applications, including battlefield surveillance, machine failure diagnosis, biological detection, inventory tracking, home security, smart spaces, environmental monitoring, and so on [1, 3]. A wireless sensor network consists of a large number of tiny, low-power, cheap sensor nodes having sensing, data processing, and wireless communication components. It has not only the ability to sense some phenomena in the interested region but also the network features, thereby representing an improvement over the traditional sensor systems.

Distinguished from traditional wireless networks, wireless sensor networks are typically characterized by denser levels of nodes deployment, higher unreliability of sensor nodes, asymmetric data transmission, and severe energy supply, computation, and memory constraints. These unique characteristics and constraints present many challenges for networks. Since the sensor nodes are equipped with tiny, irreplaceable batteries with limited power supply, it is essential that the network be energy efficient in order to prolong the lifetime of the whole network. Therefore, energy efficiency is a major design goal in WSNs[2].

Due to the nature of the WSN, a sensor node is usually powered by batteries and hence has a very constrained energy budget. To maintain longer lifetime of the network sensors, all aspects of the network should be carefully designed to be energy efficient. Clustering sensors into groups, so that sensors communicate information only to cluster heads and then the cluster heads communicate the aggregated information to the processing center, may be a good method to save energy.

Many clustering algorithms in the literature have been proposed [3-5]. But Most of the proposed clustering algorithms aim at generating the minimum number of clusters such that any node in any cluster is at most d hops away from the cluster head. Therefore, these algorithms can't get the minimal energy. In order to consume less energy, some adaptive clustering

* This work is partially supported by "The National Society Science Foundation of P.R. China", Under Grant No. 07BYY03。

approach are proposed in literature[6], which randomly rotate the role of a cluster head among all the sensor nodes in the sensor network. The operations of these algorithms are divided into rounds where each round including set-up phase and steady transmission phase. After electing cluster head, the cluster heads collect the data from the nodes within their respective clusters and use data aggregation to the base station during the transmission phase, this cluster formation approach will cause overload for some cluster heads.

Motivated by the above mentioned issues, in this article, we propose a novel energy efficiency clustering algorithm for designing wireless sensor networks. The ACE-CILP algorithm achieves its intention of consuming less energy, equalizing the energy consumption of all the nodes, as well as extending the network lifetime perfectly.

2 Problems

The sensor nodes in a wireless sensor network are usually deployed randomly inside the region of interest or close to it. A base station (BS) connected to the Internet is engaged to give commands to all the sensor nodes and gather information from the sensor nodes. In addition to sensing, the wireless sensor nodes can process the acquired information, transmit messages to the BS, and communicate to each others. Architecture of the wireless sensor network is depicted in Figure 1.

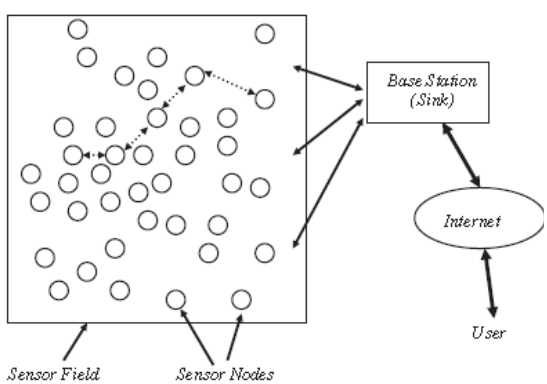


Figure 1 The architecture of a wireless sensor network

In this paper, we consider a wireless sensor network consisting of N sensor nodes and a BS, which are randomly distributed over a region. Our assumptions about the nodes and the network are as follows:

(1) Each node has the ability to transmit data directly to the BS.

(2) The BS is located far away from the monitoring field.

(3) All nodes are homogeneous and have the same capabilities.

Consider a region to be covered by N sensor nodes. We assume that the region is divided into m sub-region. We also assume that the nodes in each sub-region are organized in clusters to take advantage of possible data aggregation at the cluster head nodes. The location of sensor nodes is determined by the application requirements.

First we introduce the following notations:

(x_i, y_i): the coordinates of the I sensor node in the region, j=1, …, n.

(u,v): the coordinates of theoretical optimal cluster head node to be determined in the region.

E_j: the communication energy consumption per unit distance of j sensor node communicated with the cluster head (u, v).

Thus, we obtain the following a total energy consume:

$$MinE = \sum_{j=1}^{n} E_j[(u - x_j) + (v - y_j)^2]^{\frac{1}{2}} \qquad (1)$$

However, the communication energy consumption per unit distance is invariable, namely the value of E_j is invariable. Hence, electing cluster head and the cluster formation are critical problems in sensor network applications and can be drastically affect the network communication energy consumption.

3 The Novel Energy Efficiency Clustering Algorithm

Clustering sensors into groups, so that sensors communicate information only to cluster heads and then the cluster heads communicate the aggregated information to the processing center, may save energy.

Clustering technique has been proven to be an effective approach for organizing wireless sensor. Furthermore, the communication patterns for wireless sensor networks take one of two general forms:

- Time-driven (periodical) transmissions: Periodical transmissions from all the sensor nodes.
- Event-driven transmissions: Reports from only those sensor nodes that observe a specific event.

In this paper, we adopt the Time-driven (periodical) transmissions. A cluster based data gathering protocol consists of a series of rounds. In each round, the entire sensor network will be partitioned into different clusters. Each cluster consists of one cluster head and a number of sensor nodes. There are usually two steps in Phase (1): cluster-head election step and cluster formation step. After all the clusters are organized, the cluster head aggregates data from the sensor nodes in the cluster and then transmits information to the BS directly in Phase (2).

3.1 Algorithms for cluster head election

This section describes Algorithm ACE-C of cluster head election by counting. Suppose that the number of sensor nodes in a sensor network is N and we number the sensor nodes from 0 to N -1. Each sensor node hence can use the assigned number as a unique identifier (ID) in the sensor network. We assume that there is C clusters in each round. Using the ID's, algorithm ACE-C elects the sensor nodes as the cluster-heads in a round-robin fashion. After N/C rounds, each sensor node has been the cluster head once, and the whole process starts over from the sensor node with ID = 0. It's ID to a variable d .This variable is used to determine if it is the turn for a sensor node to be a cluster head. Initially, sensor node v sets the total number of cluster heads, t to be 0. Then, it repeatedly executes the following steps. The d is first increased by one. The algorithm then considers the remainder r of d divided by N. There are two cases:

- When r = 0, this case indicates that it is the turn for sensor node v to be a cluster head. Sensor node v then broadcasts an advertisement message to all the other sensor nodes.
- When r is not 0, this case denotes that sensor node v is not a cluster-head. It will wait a period of time for receiving an advertisement message.

After the decision, sensor node v increases the total number of cluster-heads generated, t by one. The repetition stops when t = C since there have been C cluster heads elected. Algorithm ACE-C therefore will stop and the result will be used in the next step of the clustering phase.

Algorithm of Cluster head Election by Counting

```
Input: C: the number of cluster heads in around;
       N: the total number of sensor nodes
t = 0 /* number of cluster-heads */
While t<C do
  d = (d+1) mod N
  If (d = 0) then
        v is a cluster-head
        Broadcast an advertisement message
        Increase t by 1
end if
      Wait a period of time to receive the
advertisement message
      If (a message is received) then
          Increase t by 1
      endif
  end while
End
```

3.2 Cluster formation

After the cluster heads have been elected, the next step in the clustering phase in a round is the cluster formation step. As discussed earlier, each cluster head can handle a certain number of communication channels. This constrains the number of sensor nodes that can communicate with the cluster head, thus existing in a cluster. The balanced clustering formation can overcome this shortcoming by having an upper/lower bound on the size of each cluster.

Suppose that there are C clusters in each round. For the sake of simplicity, we assume that N is a multiple of C. We want to obtain a strictly balanced solution; each cluster has to support exactly N/C sensor nodes.

We build a directed graph G, in the following way: a vertex in G denotes a sensor node or a cluster head. Moreover, for any pair of sensor node and cluster head, such as (x, H_i), we put a directed edge from x to H_i in G. Each edge has a weight equal to the message transmission energy dissipation between the two end vertices. For example, an edge connecting x and H_i has weight W(x, H_i). A source node S and a base station T are also added to G. Figure 2 shows an example of G built for a sample wireless sensor network.

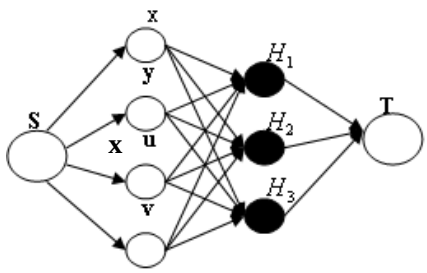


Figure 2 Directed graph Representing WSN

We use matrix M to express graph G:

$$M = \begin{bmatrix} m_{11} & m_{12} & \cdots & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & \cdots & m_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ m_{n1} & m_{n2} & \cdots & \cdots & m_{nn} \end{bmatrix} \quad (2)$$

Where M_{ij} denotes each edge in graph G, and if M_{ij}=1, sensor node i will transmit its data to its cluster head j, otherwise, M_{ij}=0 means that the node i does not transmit any data to cluster head j, in other word, sensor node i does not belong to the cluster with cluster head j.

Since each vertex corresponding to a sensor node has unit capacity, they all have to pass exactly one unit of flow. Otherwise, S can not send n unit of flow toward T. Therefore., each vertex corresponding to a cluster node has to pass exactly N/C unit of flow.

$$\sum_{j=1}^{N} M_{ij} = C \quad (3)$$

Where C denotes the number of cluster heads in the sensor network. Formulation (3) represents that there are only C connections from sensor nodes to a specific cluster head, in other word, a cluster consists of C sensor nodes and one cluster head.

Minimize the total energy consumption used:

$$E = Min \sum_{j=1}^{C} \sum_{i=1}^{N} M * W_{ij} \quad (4)$$

Where W_{ij} is the energy dissipation between sensor i and sensor head j.

4 Simulation and Results

To check the feasibility of the Improved Clustering Algorithm, we take the CCNU campus for example and give the detailed situation about the placement of sensor nodes in CCNU campus. The simulation work is realized on MATLAB. The topology of wireless sensor network in CCNU campus is shown following:

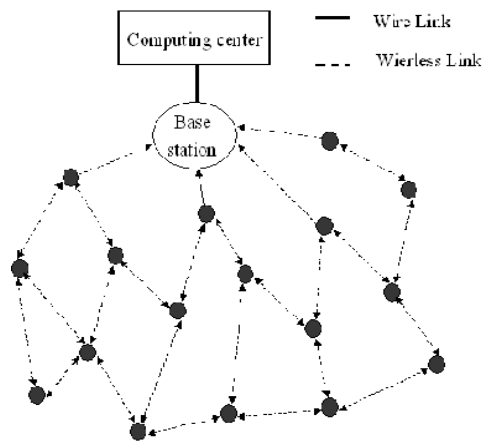


Figure 3 Topology of WSN in CCNU campus

In the experiment, we divide the campus in a 60×60 region in which the distance of every two points is d=50m by field measurement. Therefore, 60 nodes distributed uniformly in a 60×60 region of CCNU campus. Suppose the number of clusters C is 4, the cluster formation results are shown in the table 1.

Table 1 Cluster formation result

Clusters	Cluster head	Sensor node(ID)
1	6	3,9,11,18,24,27,31,35,38,41,44,50,54,59
2	21	1,4,12,14,20,25,28,34,36,43,47,53,58,60
3	17	2,5,10,15,16,23,26,30,32,37,45,49,52,56
4	42	7,8,13,19,22,29,33,39,40,46,48,51,55,57

We compare the performance of the method with

the ILP algorithm, the results in Figure 4 show the variation of the total number of nodes still alive over time. We are interested in the number of rounds when a certain number of nodes die. In ILP, the time of first node died is 103. When compared with ILP, the ACE-CILP increases the by 28%. And the time of last node died is extend. This simulation result ACE-CILP algorithm extends the network lifetime perfectly.

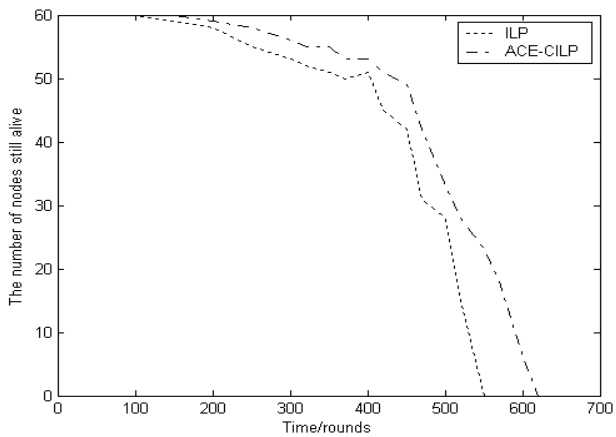


Figure 4 Number of nodes still alive over time

5 Conclusion

Since energy efficiency is a crucial factor for the performance of wireless sensor networks, using an optimal cluster formation algorithm can extend the lifetime of the sensor network because the energy dissipation is minimized. In order to achieve that goal, we propose an improved energy efficiency clustering algorithm. And the simulation results show that ACE-CILP algorithm achieves its intention of consuming less energy, equalizing the energy consumption of all the nodes, as well as extending the network lifetime perfectly.

References

[1] W. B. Heinzelman, A. P .Chandrakasan, and H.Balakrishnan, "An application-specific protocol architecture for wireless micro sensor networks," IEEE Transactions on Wireless Communications, vol.1, pp.660-670, 2002

[2] Shijin Dai, Lemin Li, Xiaorong Jing, "A Novel Cluster Formation Algorithm for Wireless Sensor Networks," ICWMMN 2006 Proceedings, pp.200-202

[3] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-Efficient Communication Protocol for Wireless microsensor networks," presented at Proceedings of the 33rd Hawaii International Conference on System Sciences, Hawaii, 2000

[4] G. J. Pottie and W. J.Kaiser, "Wireless Integrated Network Sensors," Communications of the ACM, vol. 43, pp.51-58, 2000

[5] Chuan-Ming Liua, Chuan-Hsiu Lee, Li-ChunWang, "Distributed clustering algorithms for data-gathering in wireless mobile sensor networks," J. Parallel Distrib. Compute. 67 (2007) 1187 - 1200

[6] Ameer Ahmed Abbasi , Mohamed Younis, "A survey on clustering algorithms for wireless sensor networks" Computer Communications 30 (2007) 2826-2841

[7] S. Ci, M. Guizani, "Energy map: mining wireless sensor network data, " IEEE ICC'06 (2006)

[8] Chor Ping Low, Can Fang, Jim Mee Ng, "Efficient Load-Balanced Clustering Algorithms for wireless sensor networks," Computer Communications 31 (2008) 750-759

[9] Li Li, DONG Shu-song, WEN Xiang-ming, "An energy efficient clustering routing algorithm for wireless sensor networks," The Journal of CHUPT 13(2006)71-75

[10] Rajesh Krishnan, David Starobinski, "Efficient clustering algorithms for self-organizing wireless sensor networks," Ad Hoc Networks 4 (2006) 36-59

An Elevated Trust-based Security for Mobile Ad-hoc Networks*

Yanli Pei Shijue Zheng

Department of Computer Science, Central China Normal University, Wuhan, Hubei, 430079, China
Email: ylpei905@hotmail.com

Abstract

Since mobile ad-hoc networks always applies in those situations such as emergencies, crisis management, military and healthcare, so message security is of paramount importance in it. However, because of the absence of a fixed infrastructure with designated centralized access points, implementation of hard-cryptographic security is a challenging prospect. In this paper, we propose a novel method of message security using elevated trust-based routing. Less trusted nodes are given lower number of self-encrypted parts of a message, making it difficult for malicious nodes to gain access to the minimum information required to break through the encryption strategy. Using trust levels, we make multi-path routing flexible enough to be usable in networks with 'vital' nodes and absence of necessary redundancy. In addition, using trust levels, we avoid non-trusted routes that may use brute force attacks and may decrypt messages if enough parts of the message are available to them. Simulation results, coupled with theoretical justification, affirm that the proposed solution is much more efficient than the traditional routing algorithms.

Keywords: Mobile Ad-hoc Networks; Secure Routing Protocol; Elevated Trust-based Routing Protocol; Trust-level model; Judge Rules and Trust-level Evaluation

1 Introduction

MANETs[1] are wireless and do not require any infrastructure to set up. This makes them ideal for military, rescue and relief operations. Now it was gradually extended to the civilian occasions, such as virtual classrooms, sensor networks, home networks and so on. As a wireless mobile network, the major difference between Ad hoc network and traditional mobile networks is that Ad hoc networks tried to interconnect network not rely on any fixed network infrastructure, but through mutual cooperation of mobile nodes. So that each user terminals have both function of routers and host. In addition, it has the following characteristics [2]: self-organizing, dynamic network topology, the limited bandwidth, the limitations of mobile terminals and distributed control networks. But this flexibility and the lack of a central server or access point creates a problem of security. Since the nodes co-operate to route messages, the presence of misbehaving and non-benevolent nodes is nontrivial. There is a number of security issues associated with co-operative routing in multi-hop wireless networks such as MANETs. Various routing protocols have been proposed for MANETs. Based on existing safety issues of the current Mobile Ad hoc network routing the researchers advanced a number of solutions.

2 Security Property of Mobile Ad-hoc Network Routing Protocol

The problem of security is an important matter in research of Mobile Ad hoc network routing

* This work is partially supported by The National Society Science Foundation of P.R. China, Under Grant No. 07BYY033;

protocols .Because of the dynamic network topology structure, making Ad hoc network routing protocol design very complex, which will bring new security threats. Security routing protocol should meet the following conditions [3]:

1) Can prevent the routing messages deceive;

2) Malicious node can not insert spurious routing information in the network;

3) Malicious node can not modify routing information;

4) Malicious nodes can not re-orientation from the shortest path;

5) Unauthorized nodes can not participate in the routing calculations and routing found;

6) The network topology information which contained in routing information can not be exposed to malicious or unauthorized nodes.

It must considering robustness of resist attacks, while design Mobile Ad hoc network routing protocols. Therefore, it is particularly necessary to consider the following aspects [4]:

1) Availability:

This property requires it can get the corresponding routing. The basic approach that attacker used is through Denial of Service attack. In the routing protocol, malicious node interception message but not transmitted to other nodes, it could cause denial of service attacks.

2) Confidentiality

This property requires some important information could not be show to unauthorized nodes. In mobile Ad hoc network, which is difficult to do because the intermediate node achieve the function of router transmit message, so it can wiretap packets easily.

3) Integrity

This property requires deferent message can not be tamper

4) Identity Authentication

This property allows a node to verify identity of its end-node.

5) Undeniable

This property required to ensure that the node can not deny the message which it had sent. The property is very useful to detect mutiny nodes.

3 Current Research Background

At present researcher has been proposed a variety of security routing protocol, but there are a large number of problem still exist in these protocol. As follows there three security routing protocol such as SRP, ARAN, SAODV. But they still have some problems need to solve.

Secure routing protocol (SRP)

Papadimitratos[5] advanced the secure routing protocol (SRP) , the precondition of SRP is the requirements that between the source node and terminal node there must have a secure connection. The design ideas of the protocol are: source node S initializes a routing found, first it creates a routing request packet, which contains one pair of identifier: query serial number (source node S will hold a serial number for each safety communications terminal node. The destination node can use the serial number to detect whether the request routing expired) and random query ID (For each of the query, the source node generates a random query ID, it is the role of let intermediate nodes to identify query request, it is generated by a safety pseudo-random number generator, it is unpredictable); It also includes a MAC Message Authentication Code(The method of calculate MAC is the source node's IP address, the target node's IP address, query ID and public-key which between source and target nodes as one-way hash function input, the output of function is MAC).The IP address of intermediate nodes will be collected in the head of routing query packet. When routing request packets arrive at the terminal node T, node T generates a routing reply packet to calculation and validation MAC, and along the opposite sequence of the IP address sequence which collected by the routing request packet to return routing reply packets, query node upgrade the topology table by validate the effectiveness of reply packet, the protocol can prevent several attacks. The protocol will ensure that the routing request nodes can automatically give up routing reply of wrong topology information. But there are still two security holes exist in SRP: 1) Routing can not avoid the path which contain malicious node; 2) Topology

information will be exposed through the routing information to the enemy and unauthorized nodes, because the packets which contain routing messages transmit along the entire path with clear text, it is unaccepted in the high-risk environment.

Authenticated routing for ad-hoc network (ARAN)

Based on AODV, Dahill [3] advanced an Authenticated Routing for Ad-hoc Networks, ARAN. The protocol divide into two phases, the first phase is very simple, compare with the traditional AODV [6] routing protocol it only need a few additional work, including an initial testify and a compulsory end-to-end verification phases. This is a lightweight phase, and does not need too many resources. The second phase is optional, and only after the completions of the first phase it can be carried out, because this phase need certificate of target node. The second phase is mainly use a safe manner to find the shortest path. The nodes which implement the second phase can greatly enhance the safety of their routing, but would lead to additional and even unacceptable costs (such as battery power shortfall). ARAN uses the encryption certificate to verify identity and ensure non-repudiation. Therefore, it needs a credible third-party certificate server, and must guarantee security of the certificate server absolutely. Server certificates once were captured, and the entire system will not be able to operate. In addition, ARAN protocol call for each nodes maintenance of a routing table like source vs terminal, which may cost more resources than the traditional AODV routing protocols. In particular during the implementation of the second phase, each intermediate node have to give signature to every transmit messages. Like an onion, signatures information layer upon layer to prevent malicious intermediate nodes distort messages, but at the same time increase the load on the network is very serious, even unbearable.

Secure AODV

Zapata [7] advanced Secure AODV (SAODV) routing protocol, He suggested using public-key authentication and Hash chain mechanism to enhance the safety of the AODV protocol routing found process. This makes possible attack node can not claim a nonexistent routing; At the same time, between Route Request(RREQ) and Route Reply(RREP) SAODV has also add the Hash chain field which corresponding to the hop number to check whether the alleged routing Hops are correct. SAODV deficiencies are: It requires a CA system which based on public mechanism exist in the network. This may be a reasonable assumption in the lineate network or the Internet , but for Ad hoc networks, the assumption itself is very difficult to come into existence, so the availability of security mechanism which based on the assumption have a greater doubt. Moreover, public-key authentication need intermediate nodes process a large number of calculate, consequently leading to consume a large amount of CPU, battery energy, and other resources, which in the Ad hoc network is not feasible.

4 Design And Implementation Of ETRP

In Mobile Ad hoc networks, mobile nodes transfer frequently and the distance between each node may be very far, some node can not know trust-level of another node. That is because it did not have sufficient evidence. This uncertainty is a normal phenomenon, so it needs an appropriate model to express this uncertainty. The traditional trust-level model [8] is often used probability model, but subjective logic is more suits to express the trust information. In following section we proposed the elevated trust-based routing protocol (ETRP) using subjective logic model.

Model framework of ETRP

We apply subjective logic [9] to mobile ad hoc network routing protocols, and then establish trust-based [10] routing protocol. As shown in Figure 1, ETRP includes the following modules: Trust-level Recommend (TR), Trust-level Combinations (TC), Trust-level Judgment (TJ), Encrypt Routing (ER), Trust-level Routing (TL), and Trust-level Update (TU). Trust-level Recommend and Trust-level Combinations was used to comprehensive evaluate the trust-level of a node; Trust-level Judgment used to do some corresponding operation according to the trust-level of the node ;when

meet distrustful node or before establishment of relations of trust, it must using encrypt routing module to authenticate identity; Trust-level routing is the routing operation after establishment of relations of trust-level; Trust-level Update module updates the trust-level of the node dynamically according to action of success or failure.

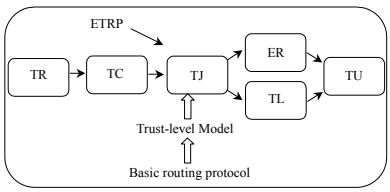


Figure 1 Model Framework of ETRP

Expression of the Trust-level model

Trust evaluation adopt the definition of subjective logic, it is a three-dimensional variables.

Definition of Trust-level evaluation: As show in following Eq. (1), ω_B^A is Trust-level which node A evaluate node B. Element b_B^A, d_B^A and u_B^A respectively expressed the probability of Belief, Disbelief and Uncertainty. They satisfy the following Eq. (2).

$$\omega_B^A = (b_B^A, d_B^A, u_B^A) \tag{1}$$

$$b_B^A, d_B^A, u_B^A = 1 \tag{2}$$

The node will collect and record the evidence of the Trust-level of other nodes in ETRP. According to the evidence, and use the following mapping relation, we can calculate the value of Trust-level evaluation.

Definition of Mapping

Let p and n respectively assume as the number success and unsuccessful of communications between node A and node B. Then, the function of b_B^A, d_B^A and u_B^A is as following:

$$\begin{cases} b_B^A = \dfrac{p}{p+n+2} \\ d_B^A = \dfrac{n}{p+n+2} \\ u_B^A = \dfrac{2}{p+n+2} \end{cases} \tag{3}$$

Node A will collect Trust-level evaluation which other nodes evaluate node B. And then congregate those Trust-level evaluations. In this way, even if there are dishonest node in the network, node A can also evaluate the Trust-level of node B objectively.

Definition of Transfer Combinations: Suppose $\omega_B^A = (b_B^A, d_B^A, u_B^A)$ is Trust-level which node A evaluate node B. $\omega_C^A = (b_C^A, d_C^A, u_C^A)$ is Trust-level which node B evaluate node C. We define transfer combinations as Eq. (4).

$$\omega_C^{AB} = (b_C^{AB}, d_C^{AB}, u_C^{AB}) \tag{4}$$

Thereinto:

$$\begin{cases} b_C^{AB} = b_B^A b_C^B \\ d_C^{AB} = d_B^A d_C^B \\ u_C^{AB} = d_B^A + u_B^A + b_B^A u_C^B \end{cases} \tag{5}$$

So, transfers combinations can calculate along the path of infer Trust-level all the while. For example, ω_B^A is the Trust-level which node A evaluate node B, ω_C^B is the Trust-level which node B evaluate node C, now node A want educe the ω_C^A, it can using transfer combinations to elicit ω_C^A from ω_C^B.

Protocol Implementation

In the original routing table of each node add three new domains: Positive Events (PE), Negative Events (NE) and Trust Evaluation (TE). PE is successful communications number between two nodes, and similarly, NE is unsuccessful communications number between two nodes, the definition of TE in above section. Trust Evaluation can calculate by Eq. (2).

Before describe the Trust-based routing found and process of routing maintenance we define some judge rules of Trust-level as show in Table 1.

In the following section we explain how dose the Trust-level model work? Suppose there are three nodes in mobile ad hoc network, at the beginning Trust-level model as show in Figure 2. Now node A want search for routing to node C. It can using model above to calculate the Trust-level, and then query the rules of Trust-level to found the routing as show in Figure 2.

Table 1 Type Size for Papers

b_B^A	d_B^A	u_B^A	Action
		>0.5	Request and validate digital signature
	>0.5		Distrust the node until routing information expired
>0.5			Trust the node and maintain routing
≤0.5	≤0.5	≤0.5	Request and validate digital signature

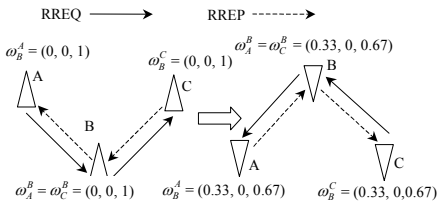


Figure 2 Initialization status of Trust-level mode

We can find that in the initial stage of ETRP for almost all nodes the uncertainty is equal to 1.along with the number of successful and unsuccessful communication increase, Trust-level evaluation gain update constantly. Uncertainty decreased gradually. The trust relationship between nodes was established gradually. In this way each nodes can adopt Trust-level combinations strategy and different Trust-level evaluation to validate the status one another. At this time it can operate the routing found.

5 Simulation And Conclusion

We select matlab as a simulation platform, focusing on the comparison of the initial routing found time. On the condition that the distance inside ten hop of the source and the target node, then record the time and location of the packets and elicit initial routing found time finally as show in Figure 3.

We can find in the Figure 3.The initialization time of routing found is less than other three protocols. Therefore, the ETRP Routing Protocol improves the performance of the security strategy.

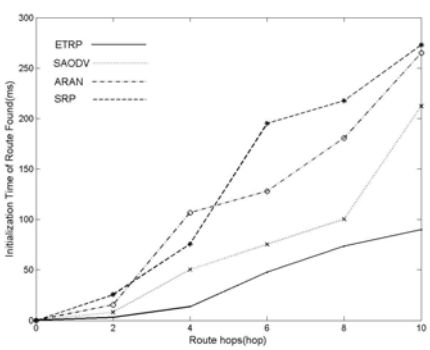


Figure 3 Initialization Routing Found Time of ETRP, SAODV, ARAN and ARP

References

[1] Magnus Frodigh, Per Johansson. Wireless Ad hoc networking - The art of networking without a network. Ericsson Review, 2000, 2(4): 248 ~ 262

[2] Prayag Narulaa, Sanjay Kumar Dhurandher,Security in mobile ad-hoc networks using soft encryption and trust-based multi-path routing Computer Communications, Volume 31, Issue 4, 5 March 2008, Pages 760-769

[3] Kimaya Sanzgiri, Bridget Dahill. A Secure Routing Protocol for Ad hoc Networks [J]. IEEE International Conference on Network Protocols (IC2NP), Paris, France, 2002, (11)

[4] A. Mishra and K.M. Nadkarni, Security in wireless ad-hoc networks. In: M. Ilyas, Editor, The Handbook of Wireless Ad-Hoc Networks, CRC Press (2003) Chapter 30, ISBN 0849313325

[5] Papadimitratos, Z J Haas. Secure Routing for Mobile Ad hoc Networks[C]. Proceedings of SCS Communication Networks and Distributed Systems Modeling and Simulation Conference, San Antonio, USA, 2002

[6] C.E. Perkins, E.M. Royer, Ad-Hoc On demand distance vector routing, in: Proceedings of IEEE WMCSA'99, New Orleans, LA, February 1999, pp. 90-100

[7] Manel Guerrero Zapata. Secure Ad hoc on-demand distance vector (SAODV) routing.ACM SIGMOBILE Mobile Computing and Communications Review, 2002, 8(3): 106~ 107

[8] Jung-Shian Li, Cheng-Ta Lee, Improve routing trust with promiscuous listening routing security algorithm in mobile ad hoc networks Computer Communications, Volume 29, Issue 8, 15 May 2006, Pages 1121-1132

[9] A. Abdul-Rahman, S. Hailes, Supporting trust in virtual communities, in: Proc. 33rd Ann. Hawaii Int'l Conf. Syst. Sci. (HICSS 33), vol. 6, (2000) pp. 6007-6016

[10] A. Boukercha, Xu Lia and K. EL-Khatibb, Trust-based security for wireless ad hoc and sensor networks. Computer Communications, Volume 30, Issues 11-12, 10 September 2007, Pages 2413-2427

SPN-Based Performance Analysis of BGP-S in Satellite networks

Wu Zeng[1] Zhiguo Hong[2]

1 Department of Electric Information Engineering, Wuhan Polytechnic University, Wuhan, Hubei 430023, China
Email: zengwude@yahoo.com.cn

2 Postdoctoral Station, Communication University of China, Beijing 100024, China
Email: hongzhiguo1977@yahoo.com.cn

Abstract

In this paper, a Stochastic Petri Net (SPN) model is constructed to analyze the performance of Border Gateway Protocol – Satellite version (BGP-S) in satellite networks. Then, the effects of satellite coverage angle, data packet size, network data rate on average time delay and average throughput are also analyzed using Stochastic Petri Net Package (SPNP) 6.0. Furthermore, a NS-2-based simulation is implemented to validate the correctness of the SPN model. The numerical results of the SPN model show a good match to simulation results of NS-2. Because of its easiness and accuracy, the proposed approach has great benefit to the design and performance analysis of satellite networks.

Keywords: Satellite Networks, SPN Model, BGP-S, Performance Analysis

1 Introduction

Recent years we have seen the development that satellite networks hold the promise of providing effective and inexpensive global coverage, providing connectivity in areas where existing terrestrial networks are either infeasible or impractical to deploy[1]-[5].

With the assumption of the satellite network being an Autonomous System (AS) with special properties, E. Ekici et al. proposed the Border Gateway Protocol – Satellite version (BGP-S) to coexist with the Border Gateway Protocol version 4 (BGP-4)[6] and support automated discovery of paths that include the satellite hops[7].

Petri nets, which was first developed in 1962 by C.A. Petri in his PhD. dissertation, is powerful in modeling concurrent, distributed, asynchronous behaviors of a system[8]. With algebra theory and the net theory as its mathematical basis, the Petri nets theory has been successfully employed to describe various relations and behaviors of the discrete event system and communication networks[9]–[11].

Satellite's attitude, coverage angle, network data rate are key parameters of describing the node in satellite networks. These parameters have significant impact on the performance of satellite networks. The effect of propagation delay on satellite networks can't be omitted. How to model and analyze the performance of BGP-S in satellite networks is an urgent and important problem. In this paper, a Stochastic Petri Nets (SPN) model is constructed to analyze the impact of some parameters on the performance of BGP-S in satellite networks by taking the geometrical characteristics into account. Also, a NS-2-based simulation is implemented to validate the theoretical results of SPN model.

The paper is organized as follows. Section 2 shows the architecture of BGP-S in satellite networks. Section 3 analyzes the geometrical characteristics of a satellite node, and Section 4 presents the constructed SPN model of BGP-S, and Section 5 calculates the networks performance using Stochastic Petri Net Package (SPNP) 6.0 [12]. Section 6 compares the numerical results of SPN model of BGP-S with NS-2-based simulation results of BGP-S. Finally, Section 7 concludes the paper.

2 The Architecture of BGP-S in Satellite Networks

As Figure 1 shows, BGP-S is proposed under the hybrid terrestrial/satellite network architecture. The terrestrial Internet is organized into Ass. Inside every AS, the routing is accomplished through Interior Gateway Protocols (IGPs). The inter-AS routing is based on an Exterior Gateway Protocol (EGP), specifically, BGP-4.

In Figure 1, two autonomous systems, ASi and ASr are depicted. The autonomous systems are connected to the satellite networks via a gateway. ASi and ASr are also connected with terrestrial links. Note that this figure is only a partial view of a likely network topology. There may be more autonomous systems with possibly different number of gateways and connected in a more complex way [13] . Active Peer Register (APR) in the satellite network is functioned as an agent to interconnect with different Peer GateWays (PGWs).

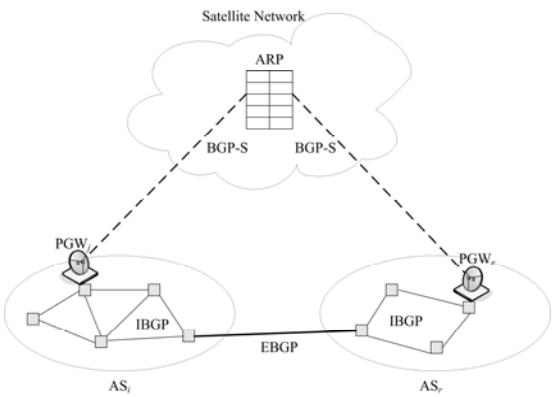


Figure 1 The hybrid terrestrial/satellite network architecture

3 Geometrical Analysis of Satellite Node in Satellite Networks

Figure 2 shows the simplified geometrical model of satellite node in satellite networks[14]. The satellite node is represented by altitude *h*, coverage angle θ etc. Here, *e* is the elevation of satellite and r_e is the radius of the earth as a constant value being 6378.14km. We use d_{AC} to represent the distance between satellite node and the peer gateway and use the "visibility is connection" communication policy, which means that satellite node and the peer gateway can communicate if they can see each other.

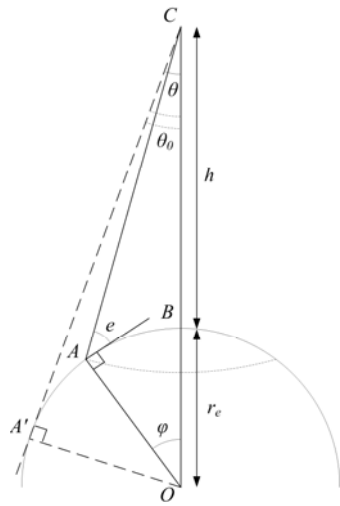


Figure 2 The geometrical model of satellite node in satellite networks

We can derive the formula of calculating d_{AC} as follows:

if θ =0, the position of peer gateway A and the satellite node's track superpose. In this case, $\varphi = 0$ and $e = \frac{\pi}{2}$, we can have $d_{AC} = h$. (1)

if $\theta > 0$, we can get:

$$\varphi = \frac{\pi}{2} - e - \theta \Rightarrow \frac{r_e}{\sin\theta} = \frac{r_e + h}{\sin(\frac{\pi}{2} + e)} = \frac{r_e + h}{\cos e}$$

$$\Rightarrow \cos e = \frac{r_e + h}{r_e}\sin\theta \Rightarrow e = \arccos(\frac{r_e + h}{r_e}\sin\theta)$$

$$\Rightarrow \varphi = \frac{\pi}{2} - \theta - \arccos(\frac{r_e + h}{r_e}\sin\theta)$$

Further, from the quantitative characteristics of ΔOAC, we can derive the following expression by applying law of sines: $\frac{d_{AC}}{\sin\varphi} = \frac{r_e}{\sin\theta}$

From the assumption of the above-mentioned "visibility is connection" communication policy, we can get the upper limit of θ:

As Figure 2 demonstrates, CA' is a tangent of the earth and θ_0 is used to denote θ. Also, from the

quantitative characteristics of $\Delta OA'C$, we can have:

$$\sin\theta_0 = \frac{r_e}{r_e + h} \Rightarrow \theta_0 = \arcsin(\frac{r_e}{r_e + h})$$

Moreover, d_{AC} can be formulated as a function of h and θ:

$$d_{AC} = r_e \cos\left(\theta + \arccos(\frac{r_e + h}{r_e}\sin\theta)\right) \Big/ \sin\theta \quad if\ \theta \in (0, \theta_0] \tag{2}$$

From Eq.(1) and Eq. (2), we can get:

$$d_{AC} = \begin{cases} r_e \cos\left(\theta + \arccos(\frac{r_e + h}{r_e}\sin\theta)\right) \Big/ \sin\theta & if\ \theta \in (0, \theta_0] \\ h & if\ \theta = 0 \end{cases}$$

4 SPN Model for BGP-S

In order to improve the accuracy of modeling and simulating BGP-S, it is necessary to analyze the composition of delay and the communication mode in satellite networks.

4.1 Composition of delay

The end-to-end delay experienced by a data packet traversing the satellite network is the sum of the transmission delay, the uplink and downlink ground segment to satellite propagation delay and the buffering delay etc. Here, we consider the buffering delay to be omitted.

1) Propagation delay

According to the data packet's propagation direction, the propagation delay (t_{prop}) is the sum of uplink User Data Links (UDLs) propagation delay (t_{up}) and download UDLs propagation delay (t_{down}).

In order to simplify the simulation, we further assume that $t_{up} = t_{down} = t_{prop} = t_{AC} = d_{AC}/c$ (c is a constant value being 3×108 m/s)

2) Transmission delay

The transmission delay (t_{trans}) is the time taken to transmit a single data packet at the network data rate. $t_{trans} = packet\ size / data\ rate$ *packet size* denotes the size of data packet, and *data rate* denotes transmission rate of data packet.

4.2 Communicating mode

Because the cost of rare resources in the satellite networks is considerably expensive, half duplex mode was taken in early small satellite systems. However, with the increase of transferring mass data via satellite networks, full duplex mode has been a trend in the future satellite networks. Here we define full duplex mode as follows:

Def. 1. Full duplex mode The mode is referred to as full duplex mode if and only if: node A and node B can pass messages to each other simultaneously. There exists the phenomenon of mutual competition in using resources (buffer, bandwidth etc.)

4.3 SPN model for BGP-S

As Figure 1 demonstrates, PGW_j can communicate PGW_r by traveling UDLs and registering the APR in the satellite network. According to the analysis of satellite node's geometrical characteristics, composition of delay and communicating mode, we make the following assumptions:

1) Different PGWs that communicate each other via APR take full duplex mode;

2) Data packets received from PGWs follow the Poisson process;

3) Data packets are transmitted independently in the satellite networks.

Figure 3 shows our constructed SPN model of BGP-S.

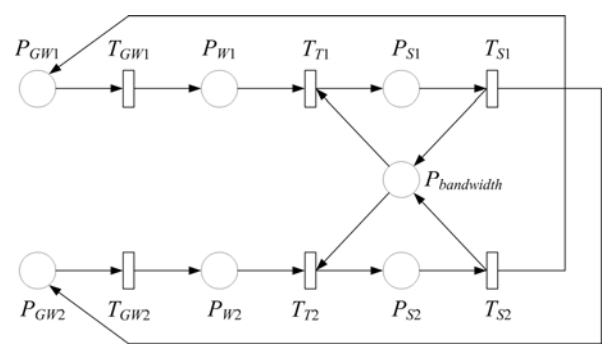


Figure 3 SPN model of BGP-S in a Satellite network

In the SPN model, we associate the Poisson process with the arrival of data packets. An important property of Poisson process is that it has tight relationship with exponential distribution. Let T_a be the interval between data packet's arrival, the interval follows the exponential distribution as follows:

$$P\{T_a < t\} = 1 - e^{-\lambda t}$$
$$E[T_a] = \frac{1}{\lambda} \quad (3)$$

Eq.(3) means that for Poisson arrival process, the interval of average arrival is the reciprocal of arrival rate numerically. Here, we use it as a basis for us to set parameters of the SPN model properly.

From the geometrical analysis of satellite node, we can see that the propagation delay between the ground and satellite node varies with satellite altitude, coverage angle etc. As a result, the impact on the performance of satellite network should not be neglected. In our SPN model, there exist both propagation delay and transmission delay during the data packet's delivering phase via UDLs. Also, the propagation delay of ground segment (including AS1 or AS2) is beyond our consideration, i,e, there only exists transmission delay when sending or receiving data packets.

By taking the full duplex mode of delivering data packets and the symmetry in the SPN model, we set the related parameters as follows:

$$\lambda_t = 1/t_{trans}\ ,\ \lambda_{pt} = 1/(t_{trans} + t_{prop})$$

5 Performance Evaluation

Average time delay and average throughput are two important performance indices of the satellite network. Average throughput means the number of data packets during a unit time. We evaluate networks performance using SPNP 6.0 software to concern average time delay and average throughput in the SPN model.

For the evaluation of performance of BGP-S in satellite networks, four experiments are designed to study different parameters' effects on average time delay and average throughput.

5.1 Case 1

In the case of h=5000, ps=512, dr=100 and k=1, the effect of satellite coverage angle on average time delay and on average throughput are shown in Figure 4 and Figure 5 respectively. For the different parameters of θ=0, π/16 and π/8, three different series are depicted.

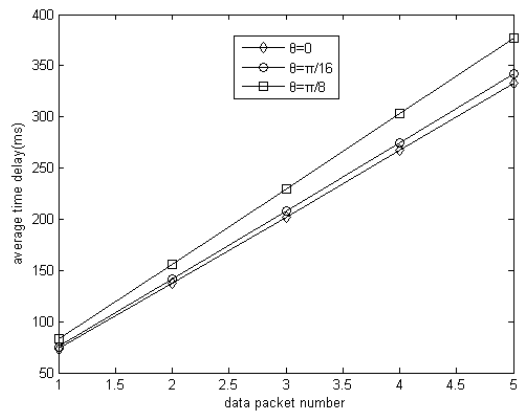


Figure 4 Effect of satellite coverage angle on average time delay

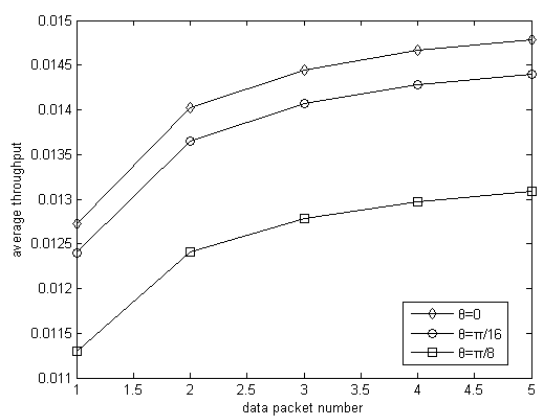


Figure 5 Effect of satellite coverage angle on average throughput

Due to the increase of satellite coverage angle θ, the distance between satellite node and PGW becomes longer which results in the addition of propagation delay via UDLs. Consequently, for one thing, it leads to the increase of average time delay in the SPN model (as Figure 4 demonstrates); for another, it brings about the

decrease of number of delivering data packets in a unit time, i.e. the reduce of average throughput (as Figure 5 demonstrates).

5.2 Case 2

In the case of h=5000, θ=π/12, dr=100 and k=1, the effect of packet size on average time delay and on average throughput are shown in Figure 6 and Figure 7 respectively. For the different parameters of ps=256, 512 and 1024, three different series are depicted.

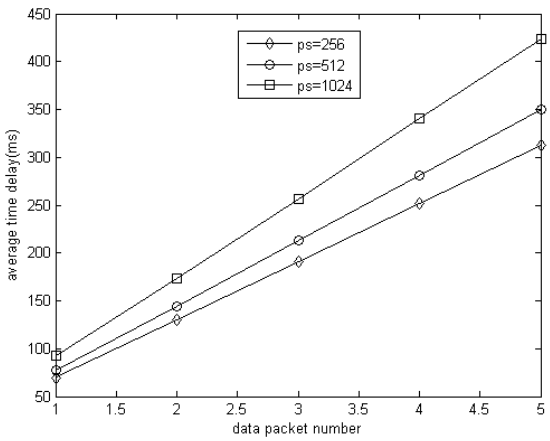


Figure 6 Effect of packet size ps on average time delay

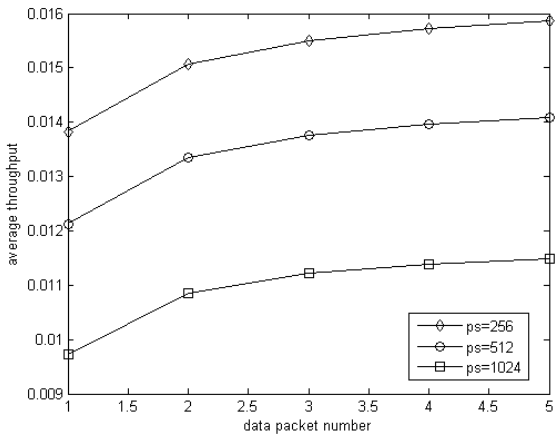


Figure 7 Effect of packet size ps on average throughput

Due to the increase of packet size ps, the propagation delay in total delay becomes bigger. Consequently, for one thing, it leads to the increase of average time delay in the SPN model (as Figure 6 demonstrates); for another, it brings about the decrease of number of delivering data packets in a unit time, i.e. the reduce of average throughput (as Figure 7 demonstrates).

5.3 Case 3

In the case of h=5000, θ=π/12, ps=512 and k=1, the effect of network data rate on average time delay and on average throughput are shown in Figure 8 and Figure 9 respectively. For the different parameters of dr=50, 100 and 1000, three different series are depicted

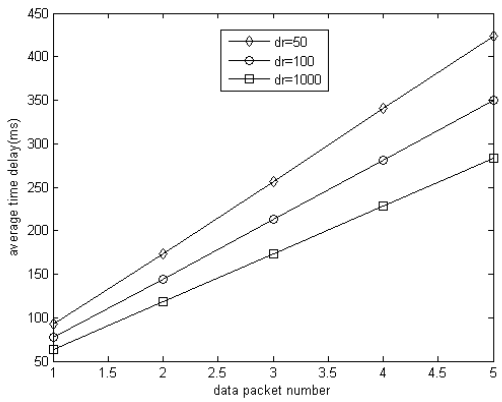


Figure 8 Effect of network data rate dr on average time delay

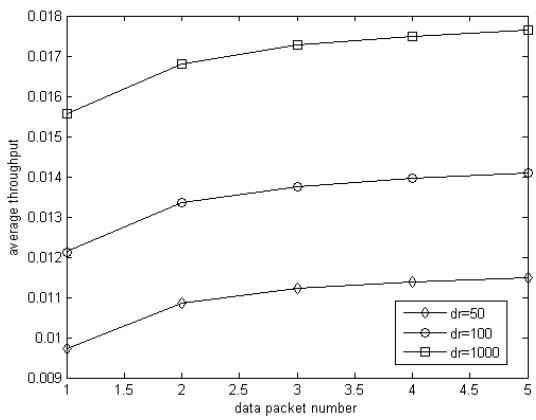


Figure 9 Effect of network data rate dr on average throughput

Due to the increase of packet size dr, the transmission delay in total delay becomes smaller. Consequently, for one thing, it leads to the reduce of average time delay in the SPN model (as Figure 8 demonstrates); for another, it brings about the increase of number of delivering data packets in a unit time, i.e. the addition of average throughput (as Figure 9 demonstrates).

6 Simulations

In order to validate the theoretical results of our SPN model, simulations have been done by the use of NS-BGP[15] which is implemented as an extension to the version of ns-2 network simulator (NS-2.27) [16] by Simon Fraser University, Canada. The parameters for BGP-S, terrestrial and satellite networks are specified in a Tcl file.

Satellite network, in which the terrestrial network has four ASs, two of which contain gateways. As Figure 10 shows, AS1 communicate with AS2 via Satellite S1, which configures the BGP-S routing Agent and the BGP-4 routing Agent. As a result, routing can be implemented either through satellite network or through terrestrial network. However, AS2 communicate AS3 and AS4 via terrestrial links, and there only exists BGP-4 routing Agent in AS4. Here we study the routing from AS1 to AS4.

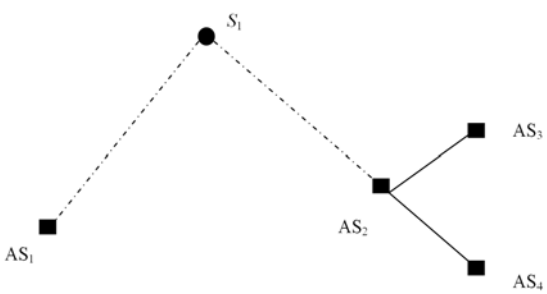


Figure 10 Condensed topology of integrated terrestrial / satellite network

Let *dr* denote satellite uplink/downlink bandwidth with the unit of Mbyte/s; Let *ps* denote data packet size with the unit of byte; Let θ denote satellite coverage angle.

Three experiments are designed to investigate different parameters' effects on average time delay and average throughput in the integrated terrestrial / satellite network.

6.1 Case 1

In the case of *ps*=512 and *dr*=100, the effect of satellite coverage angle on average time delay and on average throughput are shown in Figure 11 and Figure 12 respectively. For the different parameters of θ=0, π/16 and π/8, three different series are depicted.

It can be seen from Figure 11 and Figure 12 that with the increase of satellite coverage angle, average time delay tends to increase and average throughput tends to decline. Compared with Figure 4 and Figure 5, the results are in accordance with those of SPN model of BGP-S in satellite networks.

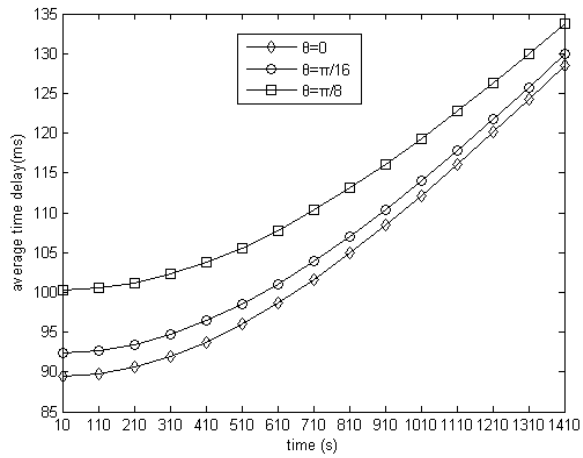


Figure 11 Effect of satellite coverage angle on average time delay

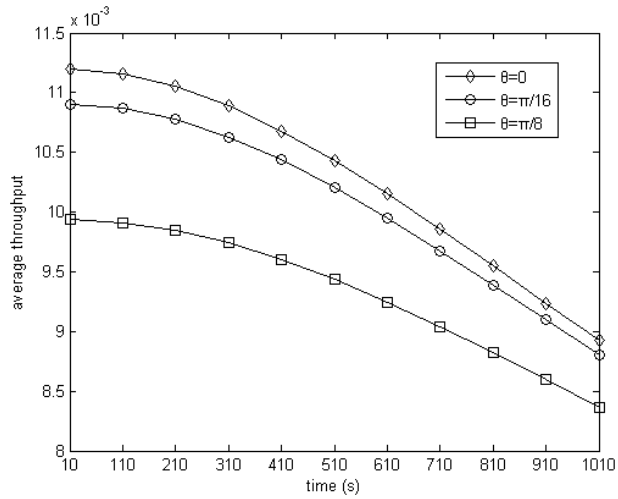


Figure 12 Effect of satellite coverage angle on average throughput

6.2 Case 2

In the case of θ=π/12 and *dr*=100, the effect of data packet size on average time delay and on average throughput are shown in Figure 13 and Figure 14 respectively. For the different parameters of *ps*=256,

512 and 1024, three different series are depicted.

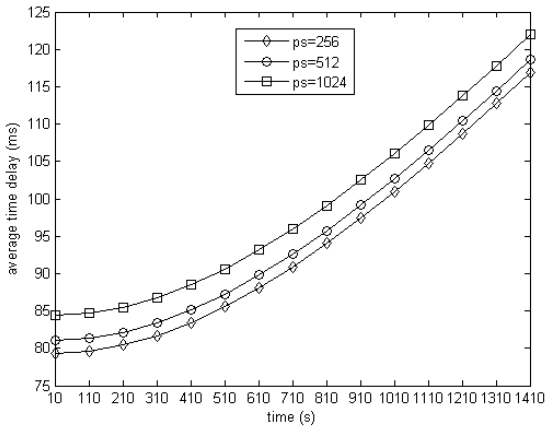


Figure 13 Effect of data packet size ps on average time delay

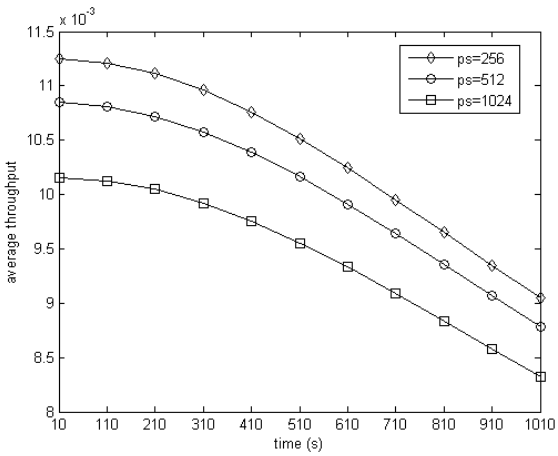


Figure 14 Effect of data packet size ps on average throughput

It can be seen from Figure 13 and Figure 14 that with the addition of data packet size, average time delay tends to increase and average throughput tends to reduce. Compared with Figure 6 and Figure 7, the results are consistent with those of SPN model of BGP-S in satellite networks.

6.3 Case 3

In the case of $\theta=\pi/12$ and ps=512, the effect of network data rate on average time delay and on average throughput are shown in Figure 15 and Figure 16 respectively. For the different parameters of dr=50, 100 and 1000, three different series are plotted.

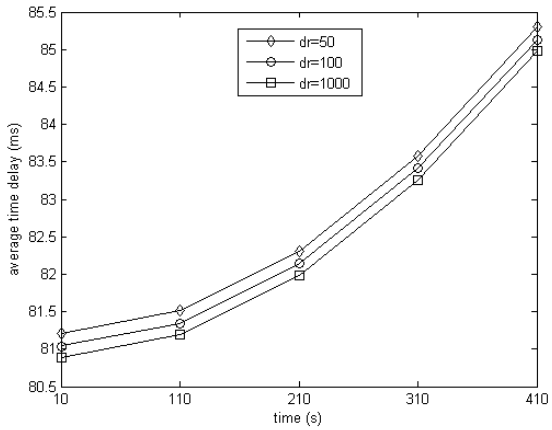


Figure 15 Effect of network data rate dr on average time delay

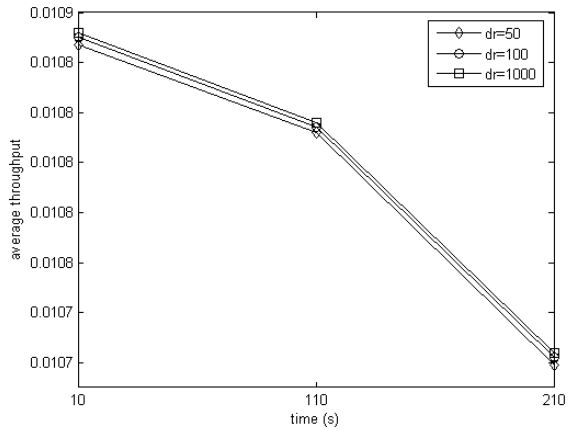


Figure 16 Effect of network data rate dr on average throughput

It can be seen from Figure 15 and Figure 16 that with the increase of network data rate, average time delay tends to become smaller and average throughput tends to become bigger. Compared with Figure 8 and Figure 9, the results accord with those of SPN model of BGP-S in satellite networks.

7 Conclusion

In this paper, a SPN-based model of BGP-S in satellite networks is constructed and the performance of the networks is evaluated by four examples. Moreover, the correctness of the SPN model is validated by NS-2 in three typical cases. It can be concluded that increasing the satellite coverage angle or the data packet size would make average time delay bigger and reduce

average throughput, whereas enhancing network data rate would decrease average time delay and improve average throughput. The results we get can be used for the design and performance optimization of satellite networks.On the basis of current work, further work on the vulnerability of BGP-S can be carried on.

References

[1] G. Compareto and R. Ramirez, "Trends in mobile satellite technology," IEEE Computer, Feb. 1997, pp. 44–52

[2] A. Hung, M.J. Montpetit, and G. Kesidis, "ATM via satellites: A framework and implementation," Wireless Networks, 1998, vol. 4, pp. 141–153

[3] A. Iera, A. Molinaro, S. Marano, and M. Petrone, "QoS for multimedia applications in satellite systems," IEEE Multimedia, Oct. –Dec. 1999, pp.46–53

[4] N.B. Melazzi and G. Reali, "A resource management scheme for satellite networks, " IEEE Multimedia, Oct. –Dec. 1999, pp. 54–63

[5] Y.F. Hu, G. Maral, and E. Ferro, "Service efficient network interconnection via satellite," John Wiley & Sons, 2002

[6] Y. Rekhter and T. Li, "A Border Gateway Protocol (BGP-4)," RFC 1771, March 1995

[7] E. Ekici, I.F. Akyildiz, M.D. Bender, "Network layer integration of terrestrial and satellite IP networks over BGP-S, " Proceedings of the IEEE, 2001, pp. 2698-2702

[8] C.A. Petri, "Communication with automata," Tech. Rep. RADC-TR-65-377, Rome Air Dev. Center, New York, 1966

[9] D. Chen, B. Soong, and K. Trivedi, "Optimal call admission control policy for wireless communication networks," in Proc. Int'l Conf. on Information, Communication and Signal Processing (ICICS), Singapore, Oct. 2001

[10] C. Xiong, T. Murata, and J. Tsai, "Modeling and simulation of routing protocol for mobile ad hoc networks using Colored Petri Nets," Research and Practice in Information Technol., Australian Computer Society, 2002, vol. 12, pp. 145–153

[11] J. Wise, J. Xia, C.K. Chang, and J.C. Huang, "Performance analysis based on requirements traceability," Tech. Rep. 05-04, Dept. Computer Science, Iowa State Univ., 2005

[12] C. Hirel, B. Tuffin, and K.S. Trivedi, "SPNP: Stochastic Petri Nets. version 6.0," B.R. Haverkort et al. (Eds.): TOOLS 2000, LNCS 1786, Berlin Heidelberg :Springer-Verlag, 2000, pp.354–357

[13] E. Ekici, C. Chen, "BGP-S: A protocol for terrestrial and satellite network integration in network layer," Wireless Networks, Vol. 10 No.5, September 2004. pp. 595-605

[14] A. Koutsaftiki, C. Matrakidis, P. Lane, M. Kokkos, and I. Macnamara, "Proposal for spotbeam pattern for GSM-based satellite personal communications," ELECTRONICS LETTERS, 1999, 35(4): 279–280

[15] D.F. Tony, "Implementation of BGP in a Network Simulator" [EB/OL]. http://www.ensc.sfu.ca/~ljilja/cnl/projects/ BGP/Tony_thesis.pdf, 2005-09

[16] ns manual: http://www.isi.edu/nsnam/ns/doc/index.htmlolors in illustrations

A Routing Algorithm Based on the Characteristic of Complex Network for Wireless Sensor Network

Yong Zhang Tingting He

Department of Computer Science, Huazhong Normal University, Wuhan, 430079, China
Email: ychang@mail.ccnu.edu.cn

Abstract

In wireless sensor networks where nodes operate on limited battery energy, some other unique characteristics make these nodes impossible to be replaced or recharged. The sensor nodes are facing the random failure and the selective attack all the time because the nodes are disabled easily. Thus it will cause partial or entire network disintegrating. We have analyzed the common three kind of wireless sensor network topology and overall considered routing thought of flooding algorithm and LEACH algorithm, then proposed a new routing algorithm based on the characteristic of complex network. In the algorithm the threshold function is used to control that numbers of few nodes'load grow excessively quickly. This is helpful in enhancing the network invulnerability. Finally the network performance analysis and the simulation result indicated that it can improve network robustness and invulnerability after applying this algorithm in the wireless sensor network.

Keywords: wireless sensor network, complex network, reliability, robustness, routing

1 Introduction

Recent advances in digital electronics, embedded systems and wireless communication are leading the way to a new class of distributed wireless sensor networks[1,2]. The wireless sensor networks combine the function of sensing, data collection and storage, computation and processing, communication through a wireless medium, and/or actuating. WSNs have been considered and envisioned in a wide spectrum of applications in various military and civil domains. Therefore, the existing and potential applications of WSNs span a wide spectrum in various domains, such as control, communication, computing, intelligence, surveillance, and targeting for military purposes; environmental detection and monitoring; disaster prevention and relief; medical care; home automation; scientific exploration; interactive surrounding, etc.

Now with the development of technique, the cost of sensor node is becoming lower and lower with its strengthening function, which makes the distributed wireless sensor networks that cost too much before can be used in more civil fields and facilitate our lives. Considering the large areas, such as many buildings and complex facilities in campus, it will be very convenient for the college logistics to unification management and environmental monitoring by collecting the data of the noises, temperature, humidity, the light of street lamp and the use state of classrooms.

Sensor networks are composed of numerous sensors by Self-Organization, which relate mutually, process and transmit information through the wireless communication. So it is necessary to consider robustness, reliability and adaptability in the design of sensor networks. But wireless sensor network has some characteristics, for example, nodes are tiny, environment is scariness, node energy is little and so on. Part of nodes may become invalid easily, and it will lead collapse of the network directly or indirectly. In the

* This work was supported by the National Natural Science Foundation of China(60773167) and 973 National Basic Research Program(2007CB310804).

categories of complex network, the phenomenon can be bracketed into two situations[3]: 1) random failure. 2) selective attack

The remainder of the paper is organized as follows: In Section 2, we provide the robustness of complex network. In Section 3, we propose a routing algorithm of wireless sensor network, and analyze it summarily. In Section 4, the simulation work is realized on MATLAB. Finally we conclude our paper in Section 5.

2 The Robustness of Complex Network

From 1998, Watts and Strogatz proposed the microcosm network model, then theory of complex networks has been studied extensively owing to their close relevance to many real networks such as the Internet, and social networks. A large number of real networks are referred to as "scale-free' and 'small-world'.

It is already known that the topological structure and propagation dynamics of complex network are closely dependent on its degree distribution, which, in its turn, is completely determined by its degree distribution exponent. Most of real networks are found, by empirical study, to have the degree distribution exponents located between 2 and 3. But, to our best knowledge, the theoretical basis of this important empirical knowledge is as yet lacking. The following conclusions are obtained: (1) the degree distribution exponents of real networks cannot be less than l; (2) there exist plenty of hub nodes for complex networks whose degree distribution exponents are between 1 and 2, the edges and nodes have nonlinear relations, and the increase of nodes will result in much more increase of edges; (3) for complex networks whose degree distribution exponents are between 2 and 3, the edges are linearly dependent on the nodes, and most networks, whose constructions are heavily controlled by cost, are of this kind; (4) complex networks whose degree distribution exponents are greater than 3 are homogeneous;

Usually a complex network is confronted with two different damages: random failure and selective attack. In other words, if some important points were attacked, it may cause Internet to be paralyzed. So complex system already has robustness, also has vulnerable one side.

Cohen defined a threshold fc as the value of robustness .When the number of invalid sensor nodes is bigger than fc, the network will **collapse. So the bigger this value is,** the better **the** robustness is. When it comes to random failure and intentional attack, this threshold is different.

Define: $fc=fc^{rand}+fc^{target}$ (1)

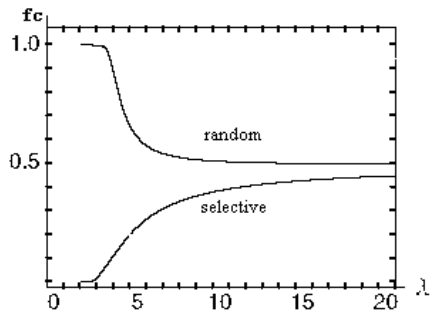


Figure 1. The graph of F_c^{rand} and F_c^{target}

In other words, the network needs the optimum topology structure. The topology structure of self-organization system is decided by the method of self-organization, which is routing protocols, too.

3 The Routing Algorithm of Wireless Sensor Network

In sensor networks all nodes have to communicate with each other because there are no base stations to coordinate the activities of subsets of nodes. Those nodes which serve as AP receive incoming data, process and send them as outgoing data. This process which called as data aggregation can greatly eliminate redundancy, minimize transmission and save energy thereby yield efficient dissemination. However, those AP nodes consume energy greater than others, and partly because advances in battery capacity have developed at a slower rate than advances in processing and communication bandwidth makes it impossible to be recharged, partly because some nodes may deploy in

harsh environments makes it impossible to be replaced[4], so they tend to die early than others and this fatal shortcoming leads to instability of the full sensor networks. By favoring energy efficiency, those AP nodes that perform in-network aggregation can affect the quality (e.g., accuracy and freshness) of the data that ultimately reaches sinks. But because APs endure more loads so that they need more energy; this character makes energy depleted fast, consequentially, one key AP node' failure may bring about the breakdown for sensor networks. Realize the severity of this question there are many papers such as Ref. [5], Ref.[6],Ref.[7] and Ref.[8] submitted trying to solve such problem. A good compromise to conserve battery capacity is to perform appropriate data aggregation[8], and in this paper we use the load-balancing algorithm based on aggregation scheme to try to distribute energy based on the load among these AP nodes to keep loading balanced, through which we can prolong the lifespan of such AP nodes. In fact there are also many papers such as Ref. [9]presents a dynamic data aggregation scheme to ensure the shift of AP nodes [10], so how to distribute the load and to optimize the use of energy among such AP nodes and maximize they lifespan is an important problem.

Routing in sensor networks is very challenging due to several characteristics that distinguish them from contemporary communication and wireless ad hoc networks. First of all, it is not possible to build a global addressing scheme for the deployment of sheer number of sensor nodes. Therefore, classical IP-based protocols cannot be applied to sensor networks. Second, in contrary to typical communication networks almost all applications of sensor networks require the flow of sensed data from multiple regions (sources) to a particular sink. Third, generated data traffic has significant redundancy in it since multiple sensors may generate same data within the vicinity of a phenomenon. Such redundancy needs to be exploited by the routing protocols to improve energy and bandwidth utilization. Fourth, sensor nodes are tightly constrained in terms of transmission power, on-board energy, processing capacity and storage and thus require careful resource management.

Due to such differences, many new algorithms have been proposed for the problem of routing data in sensor networks. These routing mechanisms have considered the characteristics of sensor nodes along with the application and architecture requirements. Almost all of the routing protocols can be classified as data-centric, hierarchical or location-based0 although there are few distinct ones based on network flow or quality of service (QoS) awareness. Data-centric protocols are query- based and depend on the naming of desired data, which helps in eliminating many redundant transmissions. Hierarchical protocols aim at clustering the nodes so that cluster heads can do some aggregation and reduction of data in order to save energy. Location-based protocols utilize the position information to relay the data to the desired regions rather than the whole network. The last category includes routing approaches that are based on general network-flow modeling and protocols that strive for meeting some QoS requirements along with the routing function.

In this paper, we will explore the routing mechanisms for sensor networks developed in recent years. Each routing protocol is discussed under the proper category. Our aim is to help better understanding of the current routing protocols for wireless sensor networks and point out open issues that can be subject to further research.

Flooding is a classical mechanism. In flooding, every sensor receiving a data packet broadcasts it to all of its neighbors and this process continues until the packet arrives at the destination or the maximum number of hops for the packet is reached. Low-energy adaptive clustering hierarchy (LEACH) is one of the first hierarchical routing approaches for sensor networks. The idea is to form clusters of the sensor nodes based on the received signal strength and to use local cluster heads as routers to the sink. Based on two routing protocols, this paper proposes a routing algorithm directed mainly at wireless sensor network in campus. In accordance with Scale-Free model and preferential attachment principle, this algorithm introduces the thought of greedy algorithm, and adjusts α, β and λ. Then self- organization reliability of wireless sensor network in campus was increased.

Considering two kind of routing method thought, this paper proposes a new routing algorithm in view of

the network reliability and robustness. First we define a threshold function F.

$$F = W_1 H \times W_2 L \times W_3 E \quad (2)$$

In Eq.(2), H is the number of hops from the sensor node to base station. L is the value of sensor 'load. E denotes the residual energy of sensor node. W1, W2 and W3 are the weights. Because wireless radio frequency transmission energy and the distance square is proportional, therefore in the routing more jumps need less energy consumption in the same distance. So we can balance the energy consumption during the sensor nodes through adjust the weights W1, W3, which will enhance the energy validity as far as possible. If the value of L is too large, on the one hand it may cause the node's energy consumption to be rapid and the energy exhausts finally, which will affect the lifetime of the entire sensor network. On the other hand also easily when suffers the selective attack expires, thus causes the multi-strip route link to receive affects, direct either indirect causes partial or the entire network disintegrates, reduces the entire network robustness. But if the load count L is low, it also be able to affect the entire network correspondence.

According to the Scale-Free model, the network growth obeys the principle that new nodes lean to connect with the node whose load is bigger.

$$P(k) = \frac{the\ amount\ of\ nodes\ whose\ L\ equal\ k}{nodes\ total}$$

$$\text{and } p(k) \sim k^{-\lambda} \quad (2)$$

According as Figure 2, when $\lambda \in [2.5, 3]$ the entire network robustness will be ideal. If the value of λ is too big resistance random ability may drop suddenly. So the detailed thought of routing algorithm is described as follows.

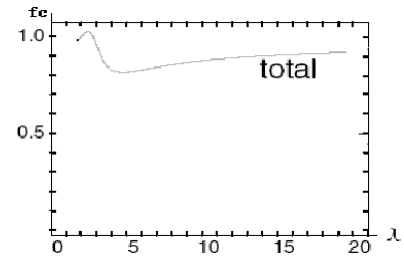


Figure 2 The graph of λ in Scale-Free network

First the sink base station activates one batch of nodes whose signals are strongest and surplus energy is biggest. After that the node record base station jumps counts H, the node load counts L and the node surplus energy measures E automatically in own routing list.

Then the nodes that were activated broadcast information immediately that includes its ID, the load number as well as it jumps from the base station and so on.

In the network the surplus nodes calculate the values by threshold function F. According to the values they choose ID of next nodes and establish the data link. Then base station activates this node. At the same time one node N2 will be chose at random and its ID will be recorded. Then it carries out the step (2).

When first choice node N1 expires suddenly and is unable to retransmits the data correctly, node N2 can be used immediately and established the data link. Then base station carries out the last step. Each activated node enters the ready stage and starts to gather and transmit the data.

Because at the very start the activation of nodes is stochastic in the network, randomicity of the network is enhanced when the energy is optimized. At the same time the robustness of the network is enhanced In addition, this algorithm prepared a redundant node for next jump. When the goal node is disabled because of energy exhaustion or attack, the prepared node will establish the data link and guarantee the data to transmit reliably. This may avoid that with the entire network disintegrates as a result of some node invalidation. Simultaneously the algorithm enhanced the network robustness and reduced the vulnerability.

4 The Analysis of Network performance

Taking characteristics of sensor network into account, we suppose that the detection and monitoring area is uniform distribution. We didn't consider the time of constructing routing and uploading data. So network efficiency can be defined easily as E =M /N[12]. It is the ratio of nodes that were uploaded successfully in a turn of the network. Directed at two invalid situations, we analyze the routing algorithm which is proposed in our paper.

The routing process of wireless sensor network is also the process which a network grows. According to the description in the Scale-Free model, the network growth obeys the principle that new nodes lean to connect with the node whose load is bigger. So in the wireless sensor network the new sensor nodes also lean to connect with the node whose load is bigger. In this way it is easy to form few main center node whose load number are extremely high. When the center nodes were attacked prepensely, it is very possible to bring up the entire network paralysis. Therefore it may cause the robustness of entire network is very low.

In view of above situation, the paper proposed the new routing algorithm which uses a threshold function to control that the sensor nodes choose the new nodes for next jump every time. Except the number of hops and the energy question which were usually must be considered in the wireless sensor network, it was joined the consideration about the node load number. Through to the adjustment of parameter W1, W2, W3, it forms some nodes whose load number are moderate and controls that few nodes load number growth excessively quickly. Then it may guarantee the network can grow reposefully. When the network was even, it satisfies the need that search different information and enhances the network stability. Simultaneously it also enhances the survivable ability of the sensor network. Directed at two invalid situations, we analyze the routing algorithm which is proposed in our paper.

We suppose that the distribution of nodes in the monitor area is uniform. The routing establishment time and the data upload time were not considered provisionally. Therefore the network efficiency may define simply as $I = \frac{N'}{N}$. N is the total of the sensor node in the network. N' is the number of nodes that each turn can succeed in uploading data in the network. From Figure 3, we can know that the more sensor nodes were invalidated, the more sensor nodes were affected and didn't succeed in uploading data. When the expired nodes achieve 5% in the network, we may see that the entire network still maintained the good network efficiency.

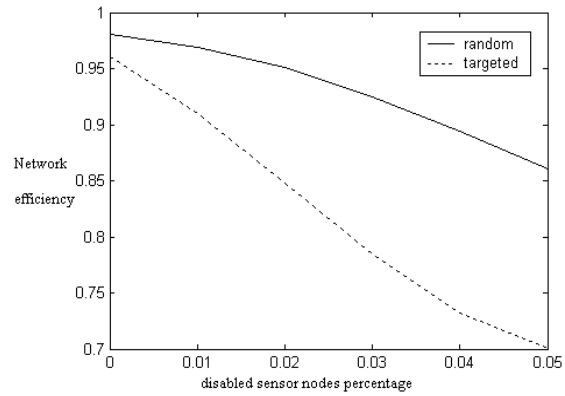


Figure 3 The graph of network efficiency

5 Conclusions & Future Works

The distributed wireless sensor networks was integrated with many advanced techniques in order to supply the way of getting information, processing information for people. Now with the development of technique, the cost of sensor node is becoming lower and lower with its strengthening function, which makes the distributed wireless sensor networks that cost too much before can be used in more civil fields and facilitate our lives, such as street lamp illumination, traffic control and intelligent home. Meanwhile it also advances a great deal of new challenges. Routing protocol is needed

to improve the performance of the wireless sensor networks.

Based on the characteristic of complex network, this paper proposes a routing algorithm directed mainly at wireless sensor network. In accordance with Scale-Free model and preferential attachment principle, this algorithm introduces the thought of greedy algorithm, and adjusts α, β and λ. Then self-organization reliability of wireless sensor network in campus was increased. But this algorithm also has certain limitation in the practical application. For example it is only suitable for the small wireless sensor network. Therefore some aspects also need further improvement. We would also like to extend our proposed algorithm in our future work.

References

[1] I.F. Akyildiz, W. Su, et al, “Wireless sensor networks: a survey”, Computer Networks, Vol.38, 2002, pp.393-422

[2] Deepak Ganesan, Alberto Cerpa, et al, “Networking issues in wireless sensor networks”, J. Parallel Distrib. Comput., Vol.64, 2004, pp.799-814

[3] Paul Gerry, Tanizawa T, Havlin H, et al. Op timization of robustness of complex networks [J]. Eur Phys J B, Vol.38, 2004, pp.187 -191

[4] Utz Roedig, Andre Barroso and Cormac J.Sreenan, “Determination of Aggregation Points in Wireless Sensor Networks,” In Proceedings of the 30th Euromicro Conference (EUROMICRO2004), Rennes, France, pp. 503–510, IEEE Computer Society Press, August 2004

[5] H. Dai, R. Han, "A Node-Centric Load Balancing Algorithm For Wireless Sensor Networks," IEEE GLOBECOM – Wireless Communications' 2003, Volume: 1, 1-5, pp. 548 - 552

[6] J.Gao and L.Zhang, “Load Balanced Short Path Routing in Wireless Networks,” The 23rd Conference of the IEEE Communications Society (INFOCOM), March 2004.

[7] I.Howitt and J.Wang, “Energy balanced chain in wireless sensor networks,” in Proceedings of the IEEE wireless Communications and Networking Conference (WCNC), 2004

[8] P. H. Hsiao, A. Hwang, H. T. Kung and D. Vlah, "Load balancing routing for wireless access networks", in Proc. IEEE INFOCOM, Apr. 2001, pp. 986-995

[9] Tarek Abdelzaher, Tian He, John Stankovic, “Feedback Control of Data Aggregation in Sensor Networks,” 43rd IEEE Conference on Decision and Control, Paradise Island, Bahamas, December 2004

[10] S. Chatterjea and P. Havinga “A Dynamic Data Aggregation Scheme for Wireless Sensor Networks,” ProRISC 2003, Veldhoven, the Netherlands, 26-27 November 2003

[11] Deepak Ganesan, Alberto Cerpa, et al, “Networking issues in wireless sensor networks”, J. Parallel Distrib. Comput., Vol.64, 2004, pp.799-814

[12] Shijue Zheng, Ying Su, Design of Distributed Wireless Sensor Networks in Campus, The fifth International Conference on Distributed Computing and Applications for Business, Engineering and Sciences, (DCABES-2006) , Hangzhou, P.R.China, 2006.10.12-15, pp:553-556

A Wireless Security Protocol Based on Ecc

Yuehua Zhao Fangkui Nong

College of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, Jiangsu, China
Email: nfnfforever@yahoo.com.cn

Abstract

With the rapid development of computer technology and mobile communication, at the same time, mobile data value-added service is widely used for it's efficiency and convenience. However, security is still an overwhelming problem relating to it. After studying elliptic curve cryptography and analyzing security features of mobile network, an effective wireless security protocol is presented, then security analysis of this improved protocol is given. At the end of this article, we can safely draw a conclusion that the improved protocol, which we put forward in the article can preferably meets the demands of mutual authentication, anonymity, confidentiality, non-repudiation requirements.

Keywords: ECDSA, Wireless Protocol, Elliptic Curve Cryptography, Confidentiality

1 Introduction

With the rapid development of mobile communication technology, mobile communication has more and more range of applications, various value-added data services, such as mobile shopping, mobile bank account transfer, mobile micro payment, advanced order-tiket, mobile phone changing, over the air technology and so on, are widely used for their efficiency and convenience in wireless environment [1~6]. At the same time, however, security is also an overwhelming problem relating to it. The opening and security vulnerability of mobile communication system, such as one way certification, vulnerability of algorithm, eavesdropping, easy to be suffered interception, camouflage, etc [7][8]. So there are more safety hidden troubles in wireless environment than that of wired environment. Therefore, people are afraid of safety problem when use value-added mobile services, especially involving transaction services. In the environment of mobile, how to make user be sure about the trade is the core and key to promote an effective development in mobile.

As a result of analysis, safety services, mainly indicate authentication service, confidentiality service, privacy service and audit service in the environment of mobile. Authentication service is mainly used for preventing camouflage, confidentiality service is used to ensure information not to be eavesdropped and modified by illegal user, privacy service guarantee the information not to be traced by third party when user use the mobile service, audit service is used for furnishing evidence after the mobile transaction. The security features we talk above could be implemented by using security authentication protocol and security session key.

Public key encryption system is an effective method to implement user authentication. However, thanks to the own restriction of moving equipment, such as the small storage space, narrow computing ability, small capacity of battery and so on, have decided that it is not suitable for mobile equipment to calculate the encryption algorithm such as RSA public-key algorithm, which demands complicated hardware equipment, and this is also the main reason that GSM system implements identity authentication by using symmetric key algorithm rather than public key system. In 1998, Aydosetal presented a wireless authentication protocol based on elliptic curve cryptosystems, this cryptographic algorithm can provide the safety guarantee same as RSA algorithm, but the key length is much longer in RSA algorithm, therefore, the algorithm baseed on ECC that

Aydosetal presents has low requiement to computing capability and storage space. At the same time, it has the feature of short time of producting key, little calculating amount and low power, it is particularly well adapted to the application that used smart card as carrier. However, in 2004, Hwung-Min Sun pointed out that serious security vulnerability problem existing in Aydos protocol, and it is not able to really realize mutual authentication of both sides. In order to overcomes the shortcomings of Aydosetal protocol, this paper presents an improved wireless security protocol, which solves the problem of identification deceit, key security, not able to really realize bidirectional identity authentication, furthermore, it effectively prevent retransmission by using counter and random number.

The main contend of this article are: At first, the paper gives a brief introduction to the digital signature algorithm ECDSA based on elliptic curve cryptography, then it gives a introduction to the mutual authentication protocol based on elliptic curve cryptography, binding the advantages of public key algorithm and private key algorithm, and producing session key dynamically in the procession of mutual authentication implementation [9][10]. Finally, analysis the safety of the protocol which put forward in the paper.

2 Introduction of ECDSA Signature Algorithm

Assumption that there exist a elliptic curve E, which be defined in the domain of GF(p) or GF(2^k), P is a point in elliptic curve E, it's order is prime number n, $H(\cdot)$ is one-way hash function, signature information is m, concrete algorithm are as follows.

The generation of the key of ECDSA algorithm:

1) User of A randomly select a integer d in the domain between 2 and n-2.

2) Calculating the formula $Q = d \times p$.

3) Regarding (E, p, n, Q) and d respectively as A's public key and private key.

4) Sending the signature information (r, s, m) to user of B.

Verification of signature of ECDSA: User of B executes following steps after receiving the information (r, s, m).

1)Calculating $c = s^{-1} \bmod n$ and $H(m)$.

2)Calculating $u_1 = H(m) \cdot c \bmod n$ and $u_2 = r \cdot c \bmod n$.

3)Calculating $u_1 \times p + u_2 \times Q = (x_0, y_0)$ and $v = x_0 \bmod n$.

4)Judging whether the formula $v = r$ is true or false? If it is true, then the verification is success.

3 A New Security Protocol that Presents in this Paper

Assumption that there exist a trusted center in the system, we just call the center CA. It is responsible for producing and distributing the certificate for user and server, the certification includes the public key and unique identity of user or server and, the CA uses its private key d_{ca} to sign on the certification, its corresponding public key is $Q_{ca} = d_{ca} \times p$. The CA could be constructed and operated by mobile operation business. This protocol includes two stages, one is registration, another is mutual authentication and key generation.

Symbol definition: $Q_k(x)$ and $R_u(x)$ respectively represent x coordinate of $Q_k(x)$ and $R_u(x)$, $H(\cdot)$ is one-way hash function, $H(d;m)$ is the hash function that controlled by the key of d, such as SHA and MD5, $sign(d;m)$ represent the ECC signature that used the key of d to sign on the information m. $E(d,m)$ and $D(d,m)$ are symmetric key algorithm that controlled by key of d, such as DES and 3DES.

3.1 Registration phase

Users are required to register on the CA before used any value-added services. Users commit their public key Q_u through safe channels or directly go to the business station where treat the CA. Then it calculates the user's certificate, and the CA writes the user's certificate $cert_u = (r_u, s_u, I_u)$ and parameters e_u,

Q_u down the user's SIM card, among them the I_u is the only marking for user. The registration phase of server is the same as that of the user, just the expression is different, which regards the subscripts to replace u.

Not as the ECDSA scheme that talk above, we modify the formula $s = k^{-1}(m + dr) \bmod n$ as $s = (k - d(r + m)) \bmod n$, correspondingly, verification equation is expressed as follow: $Q_k = sp + rQ$, $r \equiv (Q_k(x) + m) \bmod n$, the purpose to introduce this form of signature equation is to avoid inverse operation in the process of verification and to reduce operations.

3.2 Mutual authentication and the stage of producing the key

Mutual authentication and the process of producing the key is detailed description as Figure 1, the whole protocol are explained as follows.

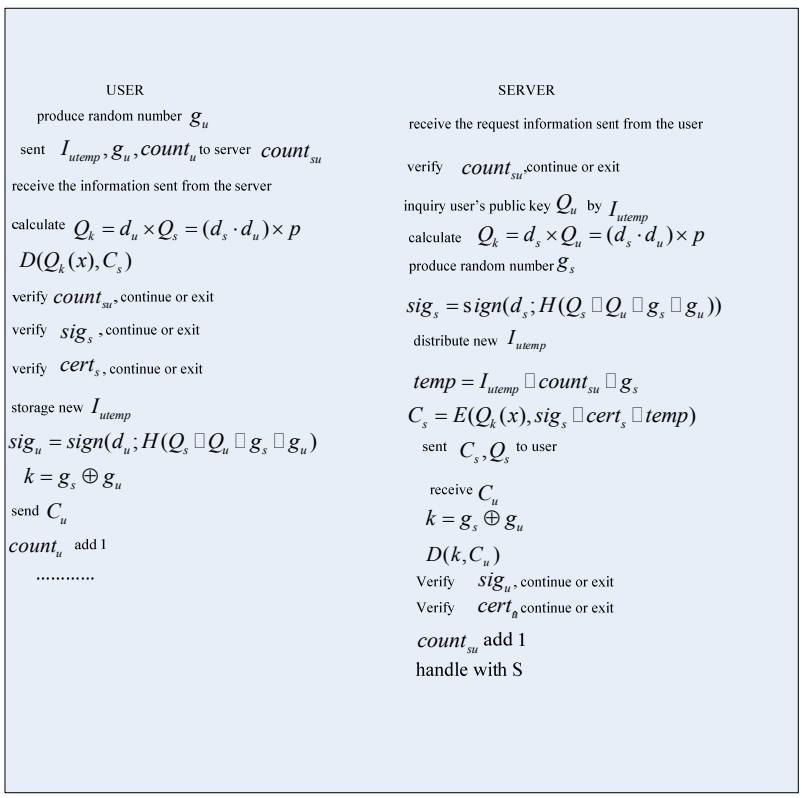


Figure 1 The process of mutual authentication and production of the key

1) It is a protocol that based on the form of challenge-response, and it is a process of real-time interactive authentication. The user put forward a challenge to the server when he want to use one of the value-added service, meanwhile, providing his temporary identity and a counter, we call it $count_u$, which is used to prevent replay attack. Under normal condition, the user and server respectively maintain a counter $count_u$ and $count_{su}$, furthermore, they are synchronization. When the server receives the request sent through the user, firstly, the server judges the two counters whether they are equal, if they are not, it shows that the two counters are out of step or had been suffered serious retransmission attack, this time server may send a synchronization information to the user to make them to get the state of synchronization again, meanwhile, the server will preserve n random number that receive recently. When the server receives the request information, it will judge the random whether it has exist at all, if it exists, then the server holds that the request is not legal, and refuses the request. On the other hand, because the identity is updated randomly, therefore, it is impossible for the attacker to modify the request parameters to make illegal request.

2) For each application, the server will rearrange a temporary identity for the user, moreover, build the

corresponding relation for the only identity in the server meanwhile, the session key, $k = g_s \oplus g_u$, changed dynamically each time, using this simple form of key generation. On one hand, it is to reduce the calculation, on the other hand, because the protocol uses the mode that one transmission requires one password, its safety can satisfy the demand of practical. Although the g_u is open for the third part, the g_s is sent to the user through the form of cipher text. So the attacker can not obtain the g_s, also the attacker can not know the key k.

3) The user uses the key $Q_k(x)$ to decrypt the C_s after he receives the C_s, then he compared the two counters whether they are equal, and whether the information is effective. These two steps are in order to avoid the calculation that is not necessary. In general case, the server is responsible to make the counter synchronous, meanwhile, temporary storage the random numbers that sent from the server, the server does not delete the random number until complete the transaction, by doing this can judge whether the replay attack has happen and enhance the safety of the system.

4) For the general user's authentication, the certificate is sent in the form of plaintext. In this protocol, both the certificate and the signature are encrypt in order to avoid spoofing attack, furthermore, it firstly authenticate the server in this protocol, if the authentication is success, then the user do corresponding calculation and send request S to server, reduce user's calculation and load as much as possible.

5) When it comes to Q_k, it just demand to calculate at the first authentication, then it storage in the SIM card, afterward, the transaction can be directly call for, so it can further reduce the calculation of the client.

4 Security Analysis of the System

4.1 Mutual authentication

The two parts of the transaction both generate a response after receiving their challenges. And they use their respective private key to sign to the response, at the same time, use the key $Q_k(x)$ to encrypt the signature and the certificate. The $Q_k(x)$ can only be produced by legal people, after receiving the response, respectively use the other party's public key and CA's public key and CA's certificate to verify the identity, it is equal to make three verification to the identity of the other part, so it strictly guarantee the legitimacy of the counterpart and effectively avoid the fraud behavior in the transaction.

4.2 The confidentiality of user's identity

From the security protocol that we talk above, we know that the user's open key and the other sensitive information are never transport in the air interface. On the other hand, the server distributes a temporary identity I_{utemp} for the user after receiving the user's request, and uses the key $Q_k(x)$, which only know by each other, to encrypt. Meanwhile, it builds a corresponding relation for temporary identity and unique identity in the server. The user does not update temporary identity marking until pass through the verification of the server. Therefore, the third part is impossible to know the real identity of the user in the transaction.

4.3 Non-Repudiation of the transaction

Usually, the non-repudiation of the transaction can be obtained by using on-line digital signature, the signature information includes user's private key information, and this private key is only kept by the user himself. Meanwhile, in this protocol, the signature information is the open key of two parts, also the random number is produced by two parts. Moreover, subsequently, the session key is produced by these random numbers, so any part of the transaction can not repudiate that fact that had taken part in the transaction.

4.4 Security of the information

The transaction information has been encrypted by

DES or 3DES algorithm, furthermore, the session key is different in the transaction every time, that is say that one transaction needs one key, by doing this can make it impossible for the attacker to obtain the data in the real transaction, so guarantee the safety of the information of the transaction.

In the protocol, we introduce counter and random number to prevent replay attack instead of the method of timestamp, if we use the later, it will be much difficult for us to make the whole system in synchronization in time.

5 Conclusions

After studying the security features of value-added service of mobile data, a protocol is presented based on the elliptic curve cryptography, which meet the demand of mutual authentication, anonymity, confidentiality, non-repudiation requirements.

References

[1] KIPA, "A Study of Cooperative Overseas Expansion Model of Mobile Solution and Contend", January 2006

[2] Baskerville.Global, "Mobile Forecasts to2010", 2005

[3] Riverst R L, Shanmir A, and Adleman L M, "A method for obtaining digital signatures and public key crytosystems", Comm of the Acm, Vol. 21, No. 2, 2001, pp. 120-126

[4] Zhenzhou Lei, "3G Market Positioning and Service Analysis Mobile Communication", January 2007

[5] Armstrong, "Competition in Telecommunication", Oxford Review of Economic Policy, No. 13, pp. 64-82

[6] Yu Xu, "Global Telecom Increment Business Development and Present Situation Information Network", No. 9, pp. 11-13

[7] Moe Rahnema, "Overview of the CSM system and protocol architecture", IEEE communications Magazine, 1993, pp. 92-100

[8] C. Lee, S. Hwang and P. Yang, "Extension of authentication protocol for GSM", IEEE Proc Commun, Vol. 150, No. 2, 2003, pp. 91-95

[9] A.K.LenStra and E.R.Verheul, "Selecting cryptographic key sizes", In the 3rd workshop on ECC, 1999

[10] Yixian Yang, Wei Sun, and Xinyi Qiu, "Modern New Theory Password", Science Press, 2002

Improvement and Research of Node Location Algorithm Based on Robust Position in Wireless Sensor Network

Wei Zhao Xiumei Wen Hui Pang

Department of Computer Science and Technology , Hebei Institute of Architecture Civil Engineering Zhangjiakou, Hebei, 075024,China
Email: shui_yun_zhaowei@163.com

Abstract

Node location is one of studying hotspot in Wireless sensor network. Now, it doesn't exist the best optimized node location algorithms because different application system has different request on node location. Two typical distributed node location algorithms were analyzed in this paper firstly and gave the analysis result of location precision and energy consumption. A modified node location algorithm(named BB_RP algorithm) was draw out. By simulated test, we compared with the primary algorithms from node location precision and energy consumption. Comparing result showed that the node location precision of BB_RP algorithm is a little lower than primary algorithms but it makes great improvement in energy consumption and coverage speed and delays the lifetime of Wireless sensor network.

Keywords: wireless sensor network(WSN),node location, location precision, energy consumption

1 Introduction

Wireless Sensor Network (WSN)contains sensor technology, embed technology, wireless communication technology and distributed information processing technology. It is a network which composed by a great lot inexpensive cost, having sensation ability, the computation power, and the wireless communication ability's sensor node. WSN will change the interactive mode that between us and object world by deploying mass sensor nodes to target area. WSN future applications will exceed our imagination[1].

It makes the nodes to have uncontrollability when the nodes are deployed because they are usually deployed complex and inclement environment. To the most applications, the sensation data has not significance if we don't know sensors' location[2]. Sensor nodes must be clear about their own position to be able to gather the message to send to the viewer, implement to locate the external object and to trace it[3]. On the side, if we know the position of sensor nodes ,it can increase router efficiency, provide naming space to the network, report the network coverage quality to the person of deployment, and implement the network load balancing and topology configuration itself.

In recent years, the domestic and foreign experts and the research institution had already done many research work in this aspect, and they had brought forward many location algorithms that using the sensor network specially. But different algorithm has its suitable range, also has the respective virtues and defects. For example centralization algorithm MDS-MAP was mentioned in REFERENCES[4], this algorithm has not limited nearly to the computation load and the reserves, and it can obtain the precise position relatively. But its defects include that the nodes which near by the center node can consume all the electrical energy prematurely because exceeding communication costs, as a result it causes the communication to be broken between the whole network and center node, and can not locate real time[5]; distributed location algorithm DV-Hop was mentioned in REFERENCES[6,7], though the computation and communication that it needs are appropriate, and nodes need not have measure distance ability, this algorithm has higher request for the network topological structure . For the anomalistic network topology structure, their location precision can descend rapidly. So there is not

optimized location algorithm, most algorithms still exist great improvement space.

The stand or fall of the performance of WSN location algorithm can effect its usability straight[1], we can estimate a location algorithm performance from location precision, anchor nodes density, energy consumption and costs.

(1) Location precision: it is the first important estimation criterion of location technology, usually we use proportion between error value and node wireless range to denotation. For example, location precision is 20% to denote location error equal to 20% node wireless range.

(2) Anchor nodes density: The anchor node refers to the node that its location is confirmed by special location system or manpower deployment. There are a small quantity of anchor nodes in the WSN. Because sensor nodes are in the complex environment, it makes very difficult to manpower deployment, however , the expense is considerable if we use location system such as GPS to deploy nodes, so anchor nodes density will effect the whole network expense and location precision straight.

(3) Energy consumption: energy finiteness is one of important characteristic for the sensor network. Because sensor network use battery as its energy source, we must consider computation costs and communication costs when we design a location algorithm. Reducing energy consumption can delay the nodes lifetime, consequently, it can assure the whole WSN lifetime also.

(4) cost: it includes two aspects about time costs and spacial costs. Time costs includes a system installing time, a system configure time, location time. Spacial costs includes the infrastructure that a location system or location algorithm needed, the number of network nodes, and hardware size, etc.

Then we will give quantitative analysis about two typical distributed location algorithms according to estimate criterion above.

2 Distributed Location Algorithms Analysis

Before giving analysis concretely, we have some assumptions and important parameter marks as follows:

Assumption: ①the nodes are deployed stochastic; ②environment is limited to in two dimensions.

Parameter mark: ①N: the number of nodes in WSN; ②A: the number of anchor nodes in WSN; ③G: WSN average connectivity; ④K: the number of the location anchor nodes which participates in once multilateral measure.

2.1 Bounding box algorithm[8]

Bounding Box Algorithm was brought forward by Mr. S.N.Semic in California University. The communication of this algorithm is discrete. The main idea of this algorithm is that assuming all the nodes are in the area D, and the D is divided into n2 cells. Let ρ be cells of communication radius, and communication area is a square which its length of side equals to 2ρ. As shown in Figure 1.

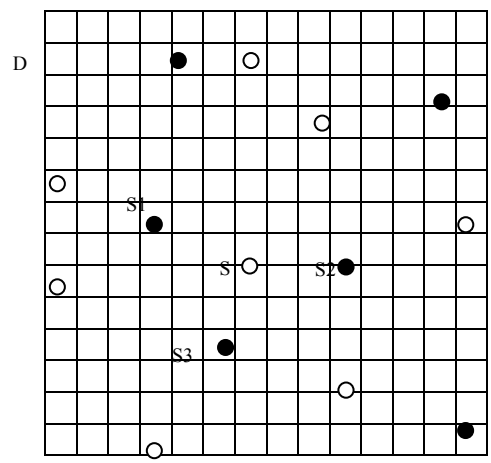


Figure 1 discrete model(let solid dot be anchor node, let hollow dot be unknown nod, ρ=3)

For the unknown node S in Fig1, we can estimate its position by using formula(1) with its neighbor anchor nodes S1,S2 and S3. Thus, if we assume that unknown node has m neighbor anchor nodes(m is the number of its neighbor anchor nodes), unknown node position is obtained by intersection of rectangle area.

$$[\max(xi-\rho),\max(yi-\rho)]\times[\min(xi+\rho),\min(yi+\rho)] \quad i=1,2,\ldots,m \qquad (1)$$

Location precision: experiment[8] shows that anchor nodes location error of Bounding Box algorithm is 10%, the proportion of anchor nodes is 20%, location

precision is 53% under the condition of measure distance error is 25%.

Computation costs: computation of this algorithm includes two pair of maximum values and minimum value when we compute each intersection. For each intersection, it needs 2G times compare operation and 4G times addition operation, so total computation costs are 6G times flop.

Communication costs: when anchor node sends broadcast message, the message is transmitted its one skip neighbor nodes only, and each node needs to communicate with its neighbor nodes one times, so the number of messages is NG in the whole network.

Energy consumption: nodes energy consumption includes computing and communication costs mainly. Synthesizing above analysis result, and using the number of sending data packages as communication energy consumption. Assuming that single flop energy consumption is F, network communication costs of this algorithm is NG, node computation costs is 6GF.

Bounding Box Algorithm's computation and communication costs is very little, and coverage speed is very fast, precision of position estimation is enhanced with increasing the number of anchor nodes. Because Bounding Box Algorithm is a distributed algorithm, it is also expanded, and every node computation complexity has nothing to do with the network scale. The algorithm's defect is that it needs higher anchor nodes density, otherwise location precision and coverage speed will be very low. So, Bounding Box Algorithm suits the node that its computation power is very finite.

2.2 Robust position algorithm[9]

Robust Position algorithm is composed of two stages: preliminary stage and refining stage. We use Hop-TERRA IN algorithm to offer each node original position estimate in the preliminary stage. The node tries to improve position estimation precision in the refining stage. The node obtains all the distance of its one skip neighbor, at the same time it is updating its own position in turn. This algorithm uses Least-squares estimation to estimate triangle measure location to compute all nodes position. It introduces trust degree to improve performance of the refining stage. Trust degree is used as weighted in triangle location. Set all unknown nodes' trust degree original value as 0.1, and set all anchor nodes' trust degree original value as 1.0. When all unknown nodes update their position estimation, they also update their trust degree to equal to average value of their neighbor nodes' trust degree. Thus, the average trust degree of network will be increased with iterative times, coverage range and precision will be enhanced also.

When unknown node obtains three or more than three anchor nodes' distance, it performs triangle measure location. We assume that the unknown node coordinate is A (x, y),the anchor node coordinate is A1 (x1 , y1), …, Ak (xk , yk) ,and the distance from unknown node to anchor node is d1, d2, …,dk, then we can establish linear equation and denote it as formula(2) as follows:

$$Cx=B \tag{2}$$

$$C=-2\times\begin{Bmatrix}(x_1-x_k)(y_1-y_k)\\(x_2-x_k)(y_2-y_k)\\ \cdot \quad \cdot\\ \cdot \quad \cdot\end{Bmatrix} X=\begin{Bmatrix}x\\y\end{Bmatrix}$$

According to distance formula between two points:

$$d_2=(x_1-x_2)^2+(y_1-y_2)^2 \tag{3}$$

$$B=\begin{Bmatrix}d_1^2-d_1^2-x_1^2+x_k^2-y_1^2+y_k^2\\d_1^2-d_1^2-x_1^2+x_k^2-y_1^2+y_k^2\\ \cdot\\ \cdot\\ \cdot\\ d_1^2-d_1^2-x_1^2+x_k^2-y_1^2+y_k^2\end{Bmatrix}$$

We can get the matrix as follows:

After establishment system of linear equations, unknown node's position estimation will be obtain by using minimal double multiplication.

$$L=(C^TC)^{-1}C^TB \tag{4}$$

Location precision: experiment shows[9] that when we deploy 400 nodes randomly in the 100m×100m area, anchor node location error of Robust Position algorithm is 10%, the proportion of anchor node is 5%, location precision is 40% under the average connectivity of the network is 12.

Table 1 Total energy consumption

	computation costs	communication costs
preliminary stage	[17 (k - 1) + 2^3 /3]F	2AN
refining stage	s[18G - 16 + 2^3 /3]F	sN
total	[17(k-1)+(18G-16)s+2^3/3 (1+s)]F	(2A+s)N

Computation costs: In the preliminary stage, each node will perform 17 (k - 1) + 23 /3 times flop because we use Hop-TERRAIN algorithm. Computation of the refining stage is similar as Hop-TERRAIN. Each iterative among the refining, the node will do once Least-squares estimation and produce new trust matrix. If iterative degree of the algorithm is s, each Least-squares estimation will include one skip neighbor of node. So, each minimal double multiplication operation will consume 17 (G-1) + 23 /3 flop. Every trust degree computation will need G times addition and one time division. Each node will need s[17 (G- 1) + 23 /3+ G + 1] = s[18G - 16 + 23 /3] flop among the refining process.

Communication costs: In the preliminary stage, each anchor node sends broadcast data package, middle node only sends data package that has not be sent, so each node sends A data package average. In the refining stage, node sends its position information to its one skip neighbor node only, so each node sends s(iterative times) data package.

Robust Position algorithm is obtained precision better, it can contain distance error when network connectivity is better. Because node communicates with its one skip neighbor node mainly. But this algorithm is forced computing because iterative process. This algorithm may be not obtain accurate estimation if original position is estimated very inaccurate or error is interrelated. The algorithm may be need very long coverage time because it depends on network topology structure.

2.3 Compare and analysis

As shown above analysis result, Both Bounding Box and Robust Position are distributed, and both are extended also, but they have defect respectively still. Bounding Box algorithm's location precision and coverage speed relate to the anchor nodes density closely, it can obtain position estimation better only if it has higher anchor nodes density. However, because Robust Position algorithm uses Least-squares method to estimation the node position both in the preliminary stage and refining stage, its energy consumption is very large and its coverage time is very long.

Based on the above characteristic, we consider to improve Robust Position algorithm. We use the idea of Robust Position algorithm as its position estimation in the preliminary stage. What we consideration is that because Bounding Box algorithm's computation and communication costs is very little, and its coverage speed is very fast[10], this can make up the limitation of Robust Position algorithm at energy costs and coverage speed. At the same time, the refining stage of Robust Position algorithm can make up the effect that anchor node brings forward. As a result it can obtain better location precision ultimately. Improved algorithm is named as BB_RP algorithm.

3 Implementation

3.1 Idea of BB_RP algorithm

BB_RP algorithm is an improved algorithm based on Robust Position algorithm. It will still divide the node location into two stage: preliminary stage and refining stage. To explain the idea of BB_RP in detail, firstly, the important parameters and functions are shown as follows:

For simple and convenient discussion, we do some assumption for the algorithm as follows: ① the position of anchor node is known in advance, it means accurate coordinate position of anchor node is known, denote it as A(x,y); ② all the nodes know their one skip neighbor node.

Preliminary stage: we use Bounding Box algorithm to compute unknown node position, this position value is original estimation. Initialize all nodes: set "location" value of anchor node as "true" and value of trust degree as 1.0; set "location" value of unknown node as "false" and value of trust degree as 0.1. Estimate the position

for the node that "location"=false, use formula (1) to compute intersection of rectangle area, then compute average value according to maximal and minimal value , set average value as position estimation eventually in the preliminary stage. At the same time we use average trust degree of its neighbor node to update its trust degree, set the value of "location" as "true". Iterative termination condition is "location" of all the nodes to become "true".

Refining stage: we use idea of Robust Position algorithm to locate node. But we use function center() to compute center of mass according to estimate values which obtained by repetitious triangle measure and make it as ultimately position estimation result. Simultaneity, we upgrade the refining node to anchor node, make it as reference node for other unknown node to locate. The consideration of above can reduce not only the number of anchor nodes in the network but also reduce the network costs, but it can bring forward some defects of increasing location error. Firstly, initialize two Boolean variable:"Last location" and" orphan" of all the nodes in the refining stage. For the node that its neighbor node number is less than 3, set its "orphan" as "true". When iterative starts, check and register the value of "Last_location" and "orphan" of each node. If "orphan" equal to "false", select three nodes randomly from its neighbor node to locate, and register each location value, compute its center of mass. Enter the next iterative. Because at least one node becomes to orphan node or upgrades to anchor node in each iterative, iterative can be terminated finally.

Table 2 Important parameters and functions in improved algorithm

Parameter and Function	Meaning
location	mark whether unknown node position has to be estimated in the preliminary stage
Last_location	mark whether unknown node has to be obtained final position in the refining stage
orphan	mark unknown node that its neighbor node <2
trust	trust degree of node
neighber	The number of neighbor nodes
Location_node	The number of neighbor nodes that has known node's position
Aver ( )	Average function
Center( )	Center of mass function

3.2 Algorithm step

Preliminary stage:

(1) initialize "location" and "trust" of all nodes:

Anchor node: *location=true, trust=1.0;*

unknown node: *location=false; trust=0.1;*

(2) *FOR i=1 to m* // m is the number of all nodes in the network

{ IF (NOT (location))

{ search this node's neighbor node;

IF(neighbor>=2 AND location)

use formula(1)to compute maximal value and minimal value;

Get_location =aver ( );

// compute average value as original position estimation

location=true;

use average value of its neighbor node's trust degree to update its trust degree;}}

Refining stage :

(3)*IF (Get_GPS)*

Get_location=GPS location; Last_Location=true; orphan=false;

// If this node is anchor node, obtain its position and update last_Location=true,orphan=false

ELSE (neighbor>=3)

Last_location=false；orphan=false;

// the neighbor node of this node is more than 3 or equal to 3

ELSE

Last_location=false；orphan=true;

// the neighbor node of this node is less than 3

(4)*WHILE(NOT (Last_location OR orphan))*

{register value of "Last_location" and " orphan" of node;

IF(neighbor<3 AND NOT(Last_location)) orphan=true;

IF(Location_node<3 OR Last_location OR orphan)

GOTO(4);

//If the number of neighbor nodes that known node position is less than 3, namely it is not satisfied with condition of triangle location and return start point of

iterative}

(5)*FOR i= 1 to s* // iterative times

select 3 neighbor nodes randomly to locate;

register each measure value;

Get_location =center(each measurement value);

// compute center of mass as last position estimation

Last_location=true;

END

4 Experiment results

Location precision: we deploy fifty nodes in the square area of 40m×40m, average transmission distance between the nodes is 10m, the number of anchor nodes is 10 and its proportion is 20%, location error of anchor node is 10%, measure distance error is 25%. Under these conditions, we compare Bounding Box algorithm with Robust Position algorithm, BB_RP algorithm from location precision. The experiment result as shown: location precision of Bounding Box algorithm is 53%, Robust Position algorithm is 35%, and BB_RP algorithm is 37%.When the number of anchor nodes are reduced to 5% and other conditions are maintained, the experiment result as shown: location of three algorithms is 80%, 40% and 41% respectively.

From the experiment result, we can get such conclusion as follows: location precision of Robust Position algorithm is the highest among the three algorithms, location precision of BB_RP is better than Bounding Box, while location precision of Bounding Box reduces rapidly. Because Robust Position algorithm uses triangle location measure to compute node position both in two stage, and one skip neighbor nodes of unknown node are all anchor node, this algorithm can obtain the better location precision. Improved algorithm BB_RP algorithm is lower than Robust Position at location precision , it has two reasons as follows: ①in the preliminary stage, we use idea of Bounding Box algorithm to locate unknown nodes, this location precision is lower than using triangle location method; ②we use triangle measure to locate in the refining stage, but BB_RP algorithm can update known node which has been located to anchor node. Thus, it can not assure one skip neighbor nodes of unknown node are all anchor nodes that has been located before measurement, however, it may be located node just. It makes location precision of improved BB_RP algorithm is lower than Robust Position algorithm because there are two factors above. Location precision of Bounding Box algorithm reduces rapidly because it is effected greatly by the anchor nodes density; location precision of Robust Position reduces also.

Because different system has different request for location precision. For example, when we locate a object inner the building, we may be locate the room that it is in but not a certain spot. Thus, we can consider improve other network performance of WSN. So BB_RP algorithm immolates a little location precision, but it can obtain great elevation from energy costs, it can make the whole network lifetime to delay finally.

Energy consumption: energy costs of BB_RP algorithm is showed as Table.3.

Table 3 Total energy consumption of improved algorithm

	computation costs	communication costs
Preliminary stage	(6G+3)F	GN
Refining stage	s[18G - 16 + 23 /3]F	SN
total	[6G+3 +S(18G-16+2^3/3)]F	(G+S)N

As shown in Table.3. we can get such conclusion as follows: it makes energy costs of the whole algorithm to reduce obviously because we use idea of Bounding Box algorithm in the preliminary stage. Computation costs can add one time division operation and twice addition operation because we compute average value to maximal and minimal value both in the preliminary stage. We still use Least-squares method to estimate node position in the refining stage, so its computation and communication costs are maintained.

It makes the total coverage speed of BB_RP algorithm to increase because its coverage speed is very fast in the preliminary stage. But coverage speed in the refining stage is still slow, so the space of improvement coverage speed is not large.

5 Conclusion

Two typical distributed node location algorithms are analyzed deeply in the thesis, and improved node location algorithm is brought forward based on it, name BB_RP algorithm. Though the location precision of BB_RP is lower than Robust Position, the parameter of energy costs is obtained improvement, and nodes lifetime is delayed, so the whole network lifetime is delayed. Indeed, so far researching for node location algorithm of WSN has not optimized, each node location algorithm has different suiTablele range. BB_RP algorithm that brought bring forward in the thesis suit for application system that request lower for node location precision but request higher for node lifetime. BB_RP algorithm can provide a little illumination for node location algorithm researching in the a future.

References

[1] Yuan-feng Ren, Hai-ning Huang, Chuang Lin, Wireless Sensor Network, Journal of Software, 2003, pp.14(2): 1148~1157

[2] Rabacy. JJ, Ammer. MJ, da Silva Jr. JL, Patel. D, Roundy. S. Picorodio supports ad hoc ultra-low power wireless networking. Computer, 2000,pp.33(7):42~48

[3] Savarese. C, Rabaey. JM, Beutel .J. Location. in distributed ad-hoc wireless sensor network. In: Proc. of the 2001 IEEE Int'l Conf. on Acoustics, Speech, and Signal. Vol.4, Salt Lake: IEEE Signal Processing Society, 2001,pp.2037~2040

[4] Shang Y, Ruml. W, Zhang Y, Fromherz. MPJ. Localization from mere connectivity. In: Proc. of the 4th ACM Int'l Symp. on Mobile Ad Hoc Networking & Computing. Annapolis: ACM Press, 2003,pp. 201~212

[5] Savvides. A, Han C-C, Srivastava. MB. Dynamic fine-grained localization in ad-hoc networks of sensors. In: Proc. of the 7th Annual Int'l Conf. on Mobile Computing and Networking. Rome: ACM Press, 2001,pp. 1661~1679

[6] N iculescu. D , Nath. B . Ad-Hoc Positioning Systems (APS) [A]. Proceedings of 2001 IEEE Global Telecommunications Conference (IEEE GLOBECOMp01) [C]. San Antonio, TX, USA: IEEE Communications Society, May 2001, pp. 2926~2931

[7] N iculescu. D, Nath. B. DV based positioning in ad hoc networks[J]. Journal of Telecommunication Systems, 2003,pp.22 (1/4) : 267~280

[8] Simic. SN, Sastry. S. UCB /ERL M02 /26, Distributed localization in wireless ad hoc networks[R]. UC Berkeley, 2002

[9] Savarese. C, Rabay. J, Langendoen. K. Robust Positioning Algorithms for Distributed Ad-Hoc Wireless Sensor Networks [A].ELL IS CS, ed. Proceedings of the USEN IX Technical Annual Conference [C]. Monterey, CA, USA: USEN IX Press, 2002,pp.317~327

[10] Hightower J, Boriello G. Location systems for ubiquitous computing.Computer,2001, pp.57~66

A Stable QoS Routing Protocol for Mobile Ad hoc Network*

Xiaoyan Zhu

School of Mathematics & Computer Science, Jianghan University, Wuhan, 430056, China
Email: zxy_jhun@163.com

Abstract

Due to the dynamic nature of the network topology and restricted resources, quality of service (QoS) and stability routing in mobile ad hoc network (MANET) is a challenging task. The paper proposes a stable QoS routing protocol (SQRP) for Mobile Ad Hoc Networks. SQRP is to construct the new metric-entropy and select the stable path with the help of entropy metric to provide stability guarantee and reduce the number of route reconstruction in mobile ad hoc networks (MANET). The SQRP is based on AODV. The simulation results shows that the SQRP approach provide an accurate and efficient method with a stable path and an enough bandwidth in MANET.

Keywords: QoS, QoS Routing, AODV, Routing protocol, MANET

1 Introduction

Mobile ad hoc networks (MANET) are comprised of mobile nodes that are communicating via either directed wireless links or multi-hop wireless links through a sequence of intermediate nodes. They are autonomously formed without any pre-configured infrastructure or centralized control. There is no static infrastructure such as base station. The hosts are free to move around randomly, thus changing the network topology dynamically. These types of networks have many advantages such as self reconfiguration and adaptability to highly variable mobile characteristics like the transmission conditions, propagation channel distribution characteristics and power level. For such networks, an effective routing algorithm is critical for adapting to node mobility as well as possible channel error to provide a feasible path for data transmission [1-4].

The traditional routing protocol used wired networks are not suited for MANET. At present, many applications need to provide quality of services (QoS). In MANET, QoS routing protocol has been a research hotspot and presented by many scholars. These protocols can be broadly classified into table-driven and on-demand routing protocols [5, 6]. AODV (Ad hoc On Demand Distance Vector) is a typical on-demand routing protocols. AODV is capable of both unicast and multicast routing. It is an on demand algorithm, meaning that it builds routes between nodes only as desired by source nodes. AODV builds routes using a route request / route reply query cycle. When a source node desires a route to a destination for which it does not already have a route, it broadcasts a route request (RREQ) packet across the network. Nodes receiving this packet update their information for the source node and set up backwards pointers to the source node in the route tables. In addition to the source node's IP address, current sequence number, and broadcast ID, the RREQ also contains the most recent sequence number for the destination of which the source node is aware. A node receiving the RREQ may send a route reply (RREP) if it is either the destination or if it has a route to the destination with corresponding sequence number greater than or equal to that contained in the RREQ. If this is the case, it unicasts a RREP back to the source. Otherwise, it rebroadcasts the RREQ. Nodes keep track of the RREQ's source IP address and broadcast ID. If

* This work is supported by a grant from Science Foundation of Ministry of Education of Wuhan of China (No.20070750).

they receive a RREQ which they have already processed, they discard the RREQ and do not forward it[5-8]. But the topology of MANET often changes because the nodes are frequently mobile. AODV has to rebuild the routing path from the source to the destination. Thus, it increases the overload of the network.

This paper proposes a stable QoS routing protocol (SQRP) for mobile ad hoc networks based on entropy and bandwidth constraint in AODV. The SQRP introduces a QoS routing model. The algorithm can find a stable route based on entropy in ad hoc networks that satisfies bandwidth constraint. It can be extended in other QoS requirements. Simulation results show that the proposed algorithm can obviously improve packet delivery ratio and reduce end to end delay.

The rest of the paper is organized as follows: Section 2 and Section 3 depict the QoS routing model and entropy metric. Section 4 presents a stable routing protocol based on entropy and bandwidth constraint. Section 5, provides simulation results. Section 6 describes the conclusion.

2 QoS Model of Network

A network can be represented as a weighted digraph G =(N, E) where N denotes the set of nodes and E denotes the set of symmetric communication links. Without loss of universality, we only consider digraphs in which there is at most one edge between two ordered nodes. For simplicity, we do not consider QoS constraints of the nodes.

Definition1.For $\forall n_i \in N, \forall n_j \in N$ and $n_i \neq n_j$ in G(N,E), (i , j) denotes the link between n_i and n_j.

Definition2. The model only takes into account the QoS constraints of the links, since the node and the link is equivalence. Assume $p(s,d)$ denotes a path form the source to the destination, where $s \in N$ and $d \in (N - \{s\})$. Assume R is the collection of positive real number and R+ is the collection of non- negative real number. For $e \in E$, the metrics of QoS is defined by functions as followed:

$$delay(e): E \to R$$
$$\cos t(e): E \to R$$
$$bandwidth(e): E \to R$$
$$delay - jitter(e): E \to R^+$$

Then the QoS of the whole path is defined:

$$delay(p(s,d)) = \sum_{e \in p(s,d)} delay(e) \quad (1)$$

$$bandwidth(p(s,d)) = \min\{bandwidth(e), e \in p(s,d)\} \quad (2)$$

$$delay - jitter(p(s,d)) = \sum_{e \in p(s,d)} delay - jitter(e) \quad (3)$$

Definition3. The QoS that has selected the routing path must satisfy promissory QoS constraints, namely:

$$\begin{aligned} &delay(p(s,d)) \le D \\ &bandwidth(p(s,d)) \ge B \\ &delay - jitter(p(s,d)) \le DJ \end{aligned} \quad (4)$$

where D, B and DJ denote the delay constraint, bandwidth constraint and delay jitter constraint, respectively.

In this paper, we only consider the bandwidth constraint. So the QoS routing model based on bandwidth is following:

The bandwidth that has selected the routing path must satisfy promissory QoS constraints, namely $bandwidth(p(s,d)) \ge B$. where $p(s,d)$ and B denote a path form the source to the destination, the bandwidth constraint, respectively.

3 Entropy Metric

We also associate each node m with a set of variable features denoted by $a_{m,n}$ where node n is a neighbor of node m. Any change of the system can be described as a change of variable values am,n in the course of time t such as $a_{m,n}(t) \to a_{m,n}(t+\Delta t)$. Let us denote by v(m, t) the velocity vector of node m and by v(n, t) the velocity vector of node n at time t. The relative velocity v(m, n, t) between nodes m and n at time t is defined as: v(m, n, t)= v(m, t) - v(n, t), Let us also denote by p(m, t) the position vector of node m and by p(n, t) the position

vector of node n at time t. The relative position p(m, n, t) between nodes m and n at time t is defined as: p(m, n, t) = p(m, t) - p(n, t), Then, the relative mobility between any pair (m, n) of nodes during some time interval is defined as their absolute relative speed and position averaged over time[9]. Therefore, we have:

$$a_{m,n} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\, p(m,n,t_i) + v(m,n,t_i)\times\Delta_{t_i}\,| - |\, p(m,n,t_{i+1})\,|}{R} \quad (5)$$

where N is the number of discrete times ti that velocity information can be calculated and disseminated to other neighboring nodes within time interval Δt. R is radio range of nodes. In general the entropy Hm(t, Δt) at mobile is calculated as follows:

$$H_m(t,\Delta_t) = \frac{-\sum_{k\in F_m} P_k(t,\Delta_t)\log P_k(t,\Delta_t)}{\log C(F_m)}, \quad (6)$$

$$F'(s,d) = \prod_{i=1}^{N_r} H_i(t,\Delta_t) \quad (7)$$

where $P_k(t,\Delta t) = (a_{m,k} \Big/ \sum_{i\in F_m} a_{m,i})$, Nr denotes the number of intermediate mobile nodes over a route between the two end nodes (s, d).

$$F(s,d) = -\ln F'(s,d) = -\sum_{i=1}^{N_r} \ln H_i(t,\Delta_t) \quad (8)$$

We are computing $F(s, u)$, and queuing it from the smallest to the biggest, namely, $F(s,u_1) \le F(s, u_2) \le \ldots \le F(s, um)$, then, the min value is the best stability path.

4 Description and Correctness of SQRP

4.1 Description of SQRP

In SQRP, the design is supposed as follows: Two neighboring nodes use GPS to communicate and obtain the coordinate of nodes. The goals of SQRP are: If the link is disconnection, the link can be renewed promptly. When the network topology is changed, the success rate of data transmission should be assured. The SQRP can decrease the overload ultimately in order to transmit packets. SQRP is based on the AODV routing protocol and uses entropy metric. The whole of SQRP includes two phases—routing discovery, routing maintenance. The routing discovery searches a path from the source to the destination with bandwidth constraint. When the path is disconnected, the routing maintenance is to renew the path as soon as possible.

The Route Discovery of SQRP Algorithm is designed as follows:

Step1: if the routing table of source node exists routing to the destination, the source directly sends message. Otherwise, go to step2.

Step2: The source broadcasts for the destination by routing request packet

Step3: If the received packet is not a duplicate and $B_{ij} \le B_c$, go to step4. Otherwise, the node should discard the packet.

Step4: If the node is not the destination, the node forwards the routing request packet. Otherwise, a route path is found.

Step5: When the destination receive RREQ packet, it do not reply the source and wait for period of time in order to the RREQ packet from the same source. Thus SQRP can be found the collection of routing path form source node to the destination.

Step6: If the collection is not null, the destination calculates F(s,d) by Eq.(8) that the path satisfies QoS , selects the most stable routing path form the collection according to the value of F(s,d). The min value of F(s,d) is the best stability path Otherwise, the destination sends routing- reconstruct information to source node. the destination sends routing reply packet to the source.

Due to the dynamic nature of the network topology and restricted resources, the established route often becomes invalid. When the source does not receive the replay packets, the upriver-node sends routing-reconstruct packet to the source. The source starts to discovery the route over again. If the source receives routing- reconstruct packet and routing reply packet at the same time, the source discards the routing reply packet and deals with routing- reconstruct packet.

4.2 Correctness of SQRP

The route found by SQRP is no loop because the

SQRP is based on AODV and AODV uses sequence numbers to ensure the freshness of routes. At the same time, the correctness of SQRP is not influenced even though the QOS constraint is used in SQRP. When a certain routing path is invalid in mobile ad hoc network, the source can receive the routing disconnection information in the process of routing maintenance because the network graph is a strong connected graph.

5 Simulation

In our simulation environment, we randomly generated 40 mobile nodes in a 900×900 meters square areas. The radio transmission rate is set to 250 meters and the data transmission rate is 2Mb/s. We use CBR as data source. The mobile speed is 0~20m/s in random direction. The QoS bandwidth requirement is 2 slots, and each slot is 5ms.The source nodes and the destination nodes are generated randomly. Each simulation time is 600 second. All simulation results are average values of multiple experiments. To evaluate the SQRP, it is compared with the AODV_QoS[10] routing protocols. In this performance evaluation the following performance metrics were evaluated: percentile of data transmission rate and route overload. The network environment for the ns2[11] simulator is given in Table 1.

Table 1 Simulation Setting

Parameters	Value
MAC Layer	IEEE802.11
Simulation Area(m)	900x 900
Simulation Time(s)	600
Mobile Nodes	40
Node Mobility Speed	0-20m/s
Node Moving Pattern	Random Way Point
Traffic Type	CBR
Packet Size	512byte
Channel bandwidth	2Mbps
transmission range	250

Figure 1 depicts a comparison of data transmission rate AODV_QoS and SQRP scheme. The data transmission rate is still higher than that of AODV_QoS, which means it is more suitable for the routing choosing under timely data transmission application and dynamic network structure. The route overload of both protocols increase with the increasing of the node mobile speed as shown in the Figure 2, because SQRP takes path stability into account based on entropy.

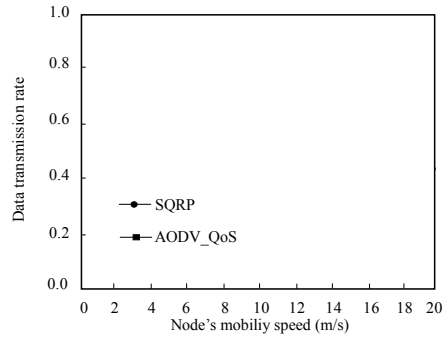


Figure 1 Data transmission rate vs. Node's mobility speed

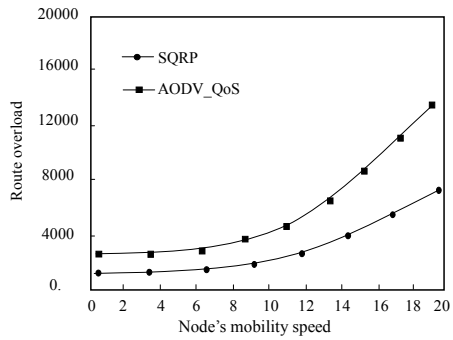


Figure 2 Route overload vs. Node's mobility speed

Figure 3 depicts a comparison of cost to control information between SQRP and AOVD_QoS. The cost of proposed scheme is less than the AOVD_QoS with increasing the scale of the network. Though the cost of SQRP increases too, the growth of cost is lesser. So SQRP will not incur the flooding storm because it can find a stable routing by using entropy. SQRP has apparent advantages in mobile ad hoc network with limited resources.

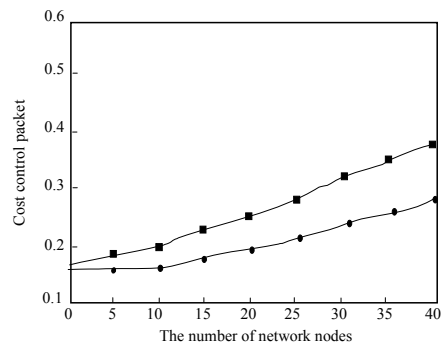


Figure 3 Cost control packet vs. Number of network nodes

6 Conclusion

In this paper, we applied entropy mechanism to ad hoc network routing protocol and present a stable qos routing protocol (SQRP). SQRP can obtain the more stable route path by entropy metric, provides a quick response to changes in the network and minimizes the waste of network resources. SQRP algorithm has produced significant improvements in data transmission rate, and route overload.

References

[1] Sun, B. L., and Li, L. Y.,"A QoS Multicast Routing Optimization Algorithms Based on Genetic Algorithm". Journal of Communications and Networks, Vol. 8, No.1, (2006) 116-122

[2] Sun Baolin,Gui Chao and Lian Jin.QoS Multicast Routing Algorithm for MANET Based on Fuzzy Controllers. The 2nd International Conference on Computer Science & Education(ICCSE 2007), July 25-28, Wuhan, China,pp:255-259

[3] Lian Jin, Li Layuan and Zhu Xiaoyan. A Multiple QoS Constraints Routing Protocol based on mobile predicting in Ad Hoc Network. The IEEE International Conference on Wireless Communications, Networking and Mobile Computing (WICOM07), Shanghai, China, 21-23 Sept, 2007,pp:1608-1611

[4] Lian Jin, Li Layuan ,Zhu Xiaoyan. A Routing Protocol with Link Status Predicting in Mobile Ad hoc Network.2007 International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES 2007), August 14-17, Yichang, China,pp:278-280

[5] B. L. Sun, and L. Y. Li, "A QoS Based Multicast Routing Protocol in Ad Hoc Networks," Chinese Journal of Computers, vol. 27, no. 10, pp. 1402-1407, 2004 (in Chinese)

[6] Zheng Feng, Li Layuan and Lian Jin. A MAODV_Based QoS Routing Protocol for Mobile Ad Hoc Networks. International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES 2007), August 14-17, Yichang, China, pp:289-291

[7] S. Chen, and K. Nahrstedt, "Distributed Quality of Service routing in Ad hoc networks," IEEE journal on selected Areas in Commn., vol. 17, no. 8, 1488-1504, 1999

[8] Sun, Q., Li, L. Y.: An Efficient Distributed Broadcasting Algorithm for Ad Hoc Networks. Advanced Parallel Processing Technologies (APPT 2005), Lecture Notes in Computer Science, Vol. 3756, Springer Verlag Berlin Heidelberg, (2005) 363-372

[9] An, B., and Papavassiliou, S.: An Entropy-Based Model for Supporting and Evaluating Route Stability in Mobile Ad hoc Wireless Networks. IEEE Communications Letters, Vol. 6, No. 8, (2002) 328-330

[10] Perkins, C., Royer, E. B., and Das, S.: Ad hoc On Demand Distance Vector (AODV) Routing. RFC 3561, July (2003)

[11] The NS Manual, A Collaboration between researchers at UC Berkeley,LBL,USC/ISI and Xerox PARC. Available at http://www.isi.edu/nsnam/ns

A Multiple Constrained Long-life QoS Multicast Routing Algorithm in MANET *

Chunhua Xia[1] Rui Yang[1] Chao Gui[2] Baolin Sun[2]

1 Department of Electrical Engineering, Hubei University of Economics, Wuhan, Hubei 430205, P. R. China
Email: goodxch2004@126.com

2 School of Computing, Hubei University of Economics, Wuhan, Hubei 430205, P. R. China
Email: blsun@163.com

Abstract

A mobile ad hoc network (MANET) is a collection mobile network nodes dynamically forming of wireless temporary without the use of any existing network infrastructure or centralized administration. This paper introduces a multiple constrained long-life QoS multicast routing algorithm in MANETs (MCQMR). The key idea of MCQMR algorithm is to construct the new metric-entropy and select the stability path with the help of entropy metric to reduce the number of route reconstruction so as to provide QoS guarantee in MANET. The simulation results show that the proposed approach and parameters provide an accurate and efficient method of estimating and evaluating the route stability in dynamic MANETs.

Keywords: Mobile Ad Hoc Networks (MANET), Entropy, Multicast Routing

1 Introduction

A mobile ad hoc network (MANET) is a collection mobile network nodes dynamically forming of wireless temporary without the use of any existing network infrastructure or centralized administration. They are autonomously formed without any pre-configured infrastructure or centralized control. Ad hoc networks also have the feature of self-creating, self-organization and elf-management as well as deploy and remove network easily. Ad hoc network has above advantages. However, the ad hoc network environment has the following restricts [1-9], including of network topology instable, limited energy constrained and limited network bandwidth-constrained. Since nodes are mobile, the network topology changes at any time whenever a wireless link is broken or reestablished due to a pair of nodes moving toward or away from each other. Moreover, they are usually deployed in an unattended environment, such as battlefields or disaster areas, and have to rely on battery power. These characteristics demand a new way of designing and operating this type of networks. For such networks, an effective routing protocol is critical for adapting to node mobility as well as possible channel error to provide a feasible path for data transmission [1-9]. As shown in Figure 1, the topology of a MANET can be very dynamic due to the mobility of mobile nodes.

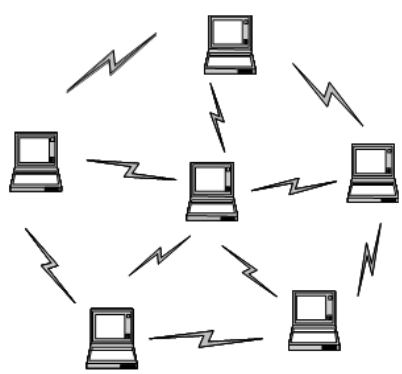


Figure 1 A typical mobile ad hoc network (MANET)

* This work is supported by National Natural Science Foundation of China (No. 60672137), NSF of Hubei Province of China (No. 2007ABA062), and Key Scientific Research Project of Hubei Education Department (No. D20081904, q200717003).

In [5], Wang et al. discussed the dynamic quality of service guarantees scheme and the intermediate adapter scheme for the networks with QoS adapter. In [6], Shen et al. proposed a complete distributed routing algorithms with low cost and applicable to high dynamic environment to choose route, through "local multicasting" scheme and location information in ad hoc networks. Chen and Nahrstedt [7] proposed another complete distributed QoS routing algorithms to choose route through local status information, solved the unique-casting in ad hoc networks better. Jetcheva and Johnson [8] proposed the design and initial evaluation of the Adaptive Demand driven Multicast Routing protocol (ADMR), the ADMR protocol that attempts to reduce as much as possible any non-on-demand components within the protocol. Multicast routing state is dynamically established and maintained only for active groups and only in nodes located between multicast senders and receivers. Each multicast data packet is forwarded along the shortest-delay path with multicast forwarding state, from the sender to the receivers, and receivers dynamically adapt to the sending pattern of senders in order to efficiently balance overhead and maintenance of the multicast routing state as nodes in the network move or as wireless transmission conditions in the network change.

There are two main issues in MANETs: the quality of service (QoS) and mobility. QoS is very important since multimedia services have become popular. Over the past few years, there have been a considerable number of studies on QoS [2, 6, 7, 8]. In a mobile environment, because of the mobility of mobile nodes in MANETs, the shortest path is not necessarily the best path. If the stability of a routing path is not considered, then wireless links may be easily broken. Many efforts have been made to design reliable routing protocols that enhance network stability [5, 8, 9].

Entropy [10, 11] presents the uncertainty and a measure of the disorder in a system. There are some common characteristics among self-organization, entropy, and the location uncertainty in mobile ad hoc wireless networks. The corresponding methodology, results and observations can be used by the routing protocols to select the most stable route between a source and a destination, in an environment where multiple paths are available, as well as to create a convenient performance measure to be used for the evaluation of the stability and connectivity in mobile ad hoc networks.

In this paper, we designed a Multiple Constrained long-life QoS Multicast Routing protocol in MANET (MCQMR). The key idea of MCQMR protocol is to construct the new metric-entropy and select the stability path with the help of entropy metric to reduce the number of route reconstruction so as to provide QoS guarantee in the ad hoc network. The goal of this paper is to develop a protocol to find out QoS-based multicast routing provisioning for guaranteed QoS, and to reduce the protocol's complexity through the local broadcasting feature in the ad hoc networks.

The rest of the paper is organized as follows: In section 2, we present the QoS models, Section 3 we present entropy metric in MANET. Section 4 introduces the MANET model and routing issues. Some simulating results are provided in section 5. Finally, the paper concludes in section 6.

2 QoS Models

It is necessary for MANETs to have an efficient routing and quality of service (QoS) mechanism to support diverse applications. RFC2386 [12] characterizes Quality of Service (QoS) as a set of service requirements to be met by the network while transporting a packet stream from the source to the destination. For the current Internet there are two different models to obtain a QoS guarantee: the Integrated Services (IntServ) [13] and Differentiated Services (DiffServ) [14].

The Integrated Services (IntServ) approach [13] aims to provide applications with a guaranteed share of bandwidth. IntServ operates on a per-flow basis, and the requested QoS for a flow is either fully granted or denied. The IntServ approach assumes that an explicit setup mechanism is used to convey resource requests to routers so that they can provide the requested services to

flows that require them.

The basic goal of the Differentiated Services [14] architecture is to meet the performance requirements of the users. Different traffic classes have different priority levels and scheduling algorithms have to ensure high priority packets are forwarded before low priority ones. DiffServ provides differential forwarding treatment to traffic, thus enforcing QoS for different traffic flows. It is a scalable solution that does not require per-flow signalling and state maintenance. DiffServ is a fully distributed and stateless model.

AQOR [15] uses a reservation-oriented method to decide admission control and allocate bandwidth for each flow. FQMM [16] is designed to provide QoS in ad hoc networks by mixing the IntServ and DiffServ mechanisms. High priority applications are provided by IntServ per-flow QoS guarantee, while lower priority applications are provided with per-class differentiation based on DiffServ. INSIGNIA [17] employs an in-band signaling protocol rather than out-of-band signaling protocol like RSVP to decrease reservation overhead. SWAN [18] is based on a reservation-less approach. By avoiding signaling, it simplifies the whole architecture and provides a differentiation between real-time and best effort in spite of not being able to guarantee the QoS needs of each flow for the whole session due to frequently changing topology and limited wireless bandwidth restriction.

3 Entropy Metric

We also associate each node m with a set of variable features denoted by $a_{m,n}$ where node n is a neighbor of node m. In this paper, two nodes are considered neighbors if they can reach each other in one hop. These variable features $a_{m,n}$ represent a measure of the relative speed among two nodes and are defined rigorously later in this section. Any change of the system can be described as a change of variable values $a_{m,n}$ in the course of time t such as $a_{m,n}(t) \rightarrow a_{m,n}(t+\Delta t)$. Let us also denote by $v(m, t)$ the velocity vector of node m and by $v(n, t)$ the velocity vector of node n at time t. Please note that velocity vectors $v(m, t)$ and $v(n, t)$ have two parameters, namely speed and direction. The relative velocity $v(m, n, t)$ between nodes m and n at time t is defined as: $v(m, n, t) = v(m, t) - v(n, t)$. Let us also denote by $p(m, t)$ the position vector of node m and by $p(n, t)$ the position vector of node n at time t. Please note that position vectors $p(m, t)$ and $p(n, t)$ have two parameters, namely position. The relative position $p(m, n, t)$ between nodes m and n at time t is defined as: $p(m, n, t) = p(m, t) - p(n, t)$.

Then, the relative mobility between any pair (m, n) of nodes during some time interval is defined as their absolute relative speed and position averaged over time. Therefore, we have:

$$a_{m,n} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\, p(m,n,t_i) + v(m,n,t_i)\times\Delta_{t_i}\,| - |\, p(m,n,t_{i+1})\,|}{R} \quad (1)$$

where N is the number of discrete times t_i that velocity information can be calculated and disseminated to other neighboring nodes within time interval Δt. R is radio range of nodes. In general the entropy $Hm(t, \Delta t)$ at mobile is calculated as follows:

$$H_m(t,\Delta_t) = \frac{-\sum_{k\in F_m} P_k(t,\Delta_t)\log P_k(t,\Delta_t)}{\log C(F_m)} \quad (2)$$

where $P_k(t, \Delta_t) = (a_{m,k} \big/ \sum_{i\in F_m} a_{m,i})$.

Let us present the route stability (RS) between two nodes s and $u \in U$ during some interval Δt as RS. We also define and evaluate two different measures to estimate and quantify end to end route stability, denoted by $F'(s, u)$ and $F(s, u)$ and defined as follows respectively:

$$F'(s,u) = \prod_{i=1}^{N_r} H_i(t,\Delta_t) \quad (3)$$

where N_r denotes the number of intermediate mobile nodes over a route between the two end nodes (s, u).

$$F(s,u) = -\ln F'(s,u) = -\sum_{i=1}^{N_r}\ln H_i(t,\Delta_t) \quad (4)$$

4 Preparation of Manuscripts

In this paper, we model the MANET system as a weighted digraph $G(N, E)$. The N is a set of mobile

nodes in MANET. The E is all set of arbitrary couple of nodes i and j, and they can communicate by the radio wave. $|N|$ and $|E|$ denote the number of nodes and links in the network respectively, without loss of generality, only digraphs are considered in which there exists at most one link between a pair of ordered nodes. Since the nodes have mobility, N and E will change dynamically

In $G(N, E)$, considering a QoS constrained multicast routing problem from a source node to multi-destination nodes, namely given a non-empty set $M=\{s, u_1, u_2, \ldots, u_m\}$, $M \subseteq N$, s is source node, $U=\{u_1, u_2, \ldots, u_m\}$ be a set of destination nodes, c_{ij} is cost of link (i,j). In multicast tree $T=(N_T, E_T)$, where $N_T \subseteq N$, $E_T \subseteq E$. if $C(T)$ is the cost of T, $P(s,u)$ is the path from source node s to destination $u \in U$, $C_{p(s,u)}$ are the cost of path $P(s,u)$, i.e.:

$$C_{p(s,u)} = \sum_{(i,j)\in p(s,u)} c_{ij} \tag{5}$$

Definition 1. The cost of multicast tree T is:

$$C(T_{ij}) = \sum_{(i,j)\in E_T} C(i,j), \quad (i,j) \in E_T \tag{6}$$

Definition 2. The bandwidth of multicast tree T is the minimum value of link bandwidth in the path from source node s to each destination node $u \in U$. i.e.

$$B_T(s,u) = \underset{u\in U}{Min}\{B_{p(s,u)}\} \tag{7}$$

Definition 3. Assume the minimum bandwidth constraint of multicast tree is B, the minimum entropy metric constraint of multicast tree is F, given a multicast demand R, then, the problem of bandwidth, and entropy metric constrained multicast routing is to find a multicast tree T, satisfying:

(1) Bandwidth constraint: $B_T(s, u) \geq B, u \in U$.

(2) Entropy metric constraint: $F_T(s, u) \leq F, u \in U$.

Suppose $S(R)$ is the set, $S(R)$ satisfies the conditions above, then, the multicast tree T which we find is:

$$C(T) = \min(C(T_s), T_s \in S(R)) \tag{8}$$

5 Simulation Experiments

Simulation Model

To conduct the simulation studies, we have used randomly generated networks on which the algorithms were executed. This ensures that the simulation results are independent of the characteristics of any particular network topology [19].

To effectively evaluate MCQMR's performance, we compare it with other famous multicast routing protocols ADMR [8] for cost to control information, average link-connect time, the success rate to find the path and the feature of data transmission. Our simulation modeled a network of mobile nodes placed randomly within 1000 m × 1000 m area with 100 mobile nodes. The radio transmission range was assumed to be 250 m. the speed of each mobile node was assumed to be 0 to 10 m/s. the data packet size was assumed to be 3 Mbps and the data sending rate was assumed to be 2 Kbps. The random waypoint mobility model was employed. Each node randomly selects a position and moves toward that location with a speed between the minimum and the maximum speed. Each simulation is executed for 600 seconds of simulation time. Multiple runs with different seed values were conducted for each scenario and collected data was averaged over those runs. A free space propagation model was used in our experiments. A traffic generator was developed to simulate CBR sources. Data sessions with randomly selected sources and destinations were simulated.

Simulation Results

The results of the simulation are positive with respect to performance. We use the ns-2 simulator [20] to evaluate the MCQMR protocol

Figure 2 depicts a comparison of cost to control information two protocols. In Figure 2, the control information increased as the number of mobile nodes increases. We can see that MCQMR's cost is smaller than that of ADMR with the increase of the scale of the network, the extend QoS constraints into ADMR, the cost to control information also increases; but for MCQMR, with its feasible path and QoS restrictive diffuse scheme, the growth of cost to control information is lower, so MCQMR will not incur the flooding storm. Due to the scarcity of MANET resource, to MANET multicast routing problems, MCQMR has apparent advantages.

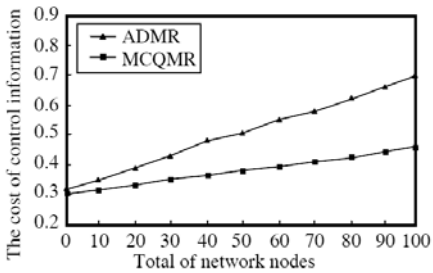


Figure 2 Cost-Comparison with control information

Figure 3 depicts the comparison of data transmission rate under nodes' changing movement speed for these three protocols: the faster the node's movement speed, the smaller the protocol's data transmission rate, due to the fact that when the movement speed increase for the nodes, the network's topology structure changes faster. From the Figure 3 we can see that when the node's movement speed increases, MCQMR data transmission rate is higher than that of ADMR. When the node movement speed is control with a range, the network's topology structure will not change fast, the link's break rate of the multicast tree is low, make MCQMR QoS constraints assured within most of user's movement speed range, so MCQMR has a good performance within the network node's constrained movement speed scope.

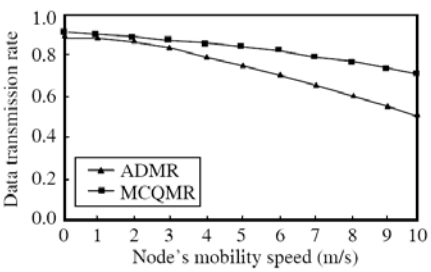


Figure 3 Comparison of data transmission rate

Figure 4 depicts a comparison of number of route reconstructions against mobility through ADMR, MCQMR protocols. Whenever path error occurs, it needs to reconstruct, and route number of reconstructions characterize the route's stability to some extent. From Figure 4 we can see that the time of route reconstructions for MCQMR is superior and more stable.

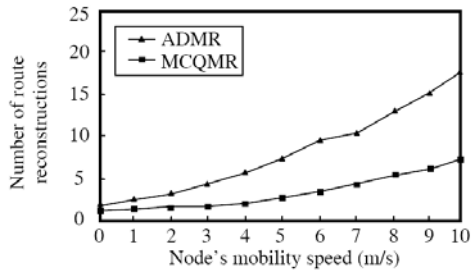


Figure 4 Number of route reconstructions against mobility

6 Conclusion

This paper discusses the multicast routing problem with multiple QoS constraints, which may deal with the bandwidth, entropy and cost metrics, and describes a network model for researching the MANET QoS multicast routing problem. It presents a Multiple Constrained long-life QoS Multicast Routing protocol in MANET (MCQMR). The key idea of MCQMR protocol is to construct the new metric-entropy and select the stability path with the help of entropy metric to reduce the number of route reconstruction so as to provide QoS guarantee in MANET. The proposed protocol can be easily extended to other mobile networks QoS routing problems with NP complexity. The simulation results show that the proposed approach and parameters provide an accurate and efficient method of estimating and evaluating the route stability in dynamic MANET.

References

[1] B. L. Sun, L. Y. Li, "A QoS Multicast Routing Optimization Algorithms Based on Genetic Algorithm", Journal of Communications and Networks, Vol. 8, No. 1, 2006, pp. 116-122

[2] B. L. Sun, L. Y. Li, "A QoS Based Multicast Routing Protocol in Ad Hoc Networks", Chinese Journal of Computers, Vol. 27, No.10, 2004, pp. 1402-1407. (in Chinese)

[3] B. L. Sun, "Long-Life Multicast Routing Protocol in MAODV Based on Entropy", *Journal of Computational Information Systems, Vol. 1, No. 2, 2005, pp.* 263-268

[4] B. L. Sun, C. Gui, H. Chen, "On the Reliability of ODMRP in Ad Hoc Networks", Journal of Computational Information Systems, Vol. 3, No. 1, 2007, pp. 133-137

[5] H. T. Wang, S. R. Zheng, L. H. Song, "The Researches on Guarantee Mechanisms of QoS in Ad Hoc network", Journal of China Institute of Communications, Vol. 23, No. 10, 2002, pp. 114-120. (in Chinese)

[6] H. Shen, B. X. Shi, L. Zou, et al, "The Location-Based QoS Routing Algorithm in Ad Hoc Network", Journal of China Institute of Communications, Vol. 24, No. 9, 2003, pp. 27-34. (in Chinese)

[7] S. Chen, K. Nahrstedt, "Distributed Quality-of-Service Routing in Ad Hoc Networks", IEEE Journal on Selected Areas in Communications, Vol. 17, No. 8, 1999, pp. 1488-1505

[8] G. J. Jetcheva, B. D. Johnson, "Adaptive Demand-Driven Multicast Routing in Multi-hop Wireless Ad Hoc Networks", Proceedings of the 2nd ACM International Symposium on Mobile Ad Hoc Networking & Computing. New York: ACM Press, 2001, pp. 33-44

[9] N. S. Chen, L. Y. Li, W. S. Dong, "Multicast routing algorithm of multiple QoS Based on widest-bandwidth", Journal of Systems Engineering and Electronics, Vol. 17, No. 3, 2006, pp. 642-647

[10] B. An, and S. Papavassiliou, "An Entropy-Based Model for Supporting and Evaluating Route Stability in Mobile Ad hoc Wireless Networks", IEEE Communications Letters, Vol. 6, No. 8, 2002, pp. 328-330

[11] A. Shiozaki, "Edge extraction using entropy operator, Computer", Vision, Graphics, and Image Processing, Vol. 36, No. 1, 1986, pp. 1-9

[12] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick. "A framework for QoS-based routing in the Internet". Request for Comments 2386, Internet Engineering Task Force, August 1998

[13] R. Braden, D. Clark, and S. Shenker. "Integrated Services in Internet Architecture – an Overview". IETF RFC1663, June1994

[14] S. Blake. "An Architecture for Differentiated Services". IETF RFC2475, December 1998

[15] Q. Xue and A. Ganz, "Ad hoc QoS on-demand routing (AQOR) in mobile ad hoc networks", Jounal of Parallel and Distributed Computing, (63), pp.154-165, 2003

[16] H. Xiao, W. K. G.. Seah, A. Lo, and K. C. Chua. "A Flexible Quality of Service Model for Mobile Ad hoc networks", In Proceedings of IEEE Vehicular Technology Conference, Tokyo, Japan, May 2000, pp. 445-449

[17] X. Zhang S.B. Lee, A. Gating-Seop and A.T. Campbell, "INSIGNIA: An IP-Based Quality of Service Framework for Mobile Ad hoc Networks". Journal of Parallel and distributed Computing, 60(4), 374-406, April 2000

[18] Gahng-Seop Ahn, Andrew T. Campbell, A. Veres, and L. Sun, "Supporting Service Differentiation for Real Time and Best Effort Traffic in Stateless Wireless Ad hoc Networks (SWAN)", In IEEE Transactions on Mobile Computing, vol. 1, no. 3, September 2002, pp. 192-207

[19] B. Waxman, "Routing of Multipoint Connections", IEEE Journal on Selected Areas in Communications, Vol. 6, No. 9, 1988, pp. 1617-1622

[20] The Network Simulator - ns-2, http://www.isi.edu/nsnam/ns

Direct Kinematics Analysis of a Special Class of the Stewart Manipulators

Xiaogang Ji Yi Cao Hui Zhou Jinghu Yu

School of Mechanical Engineering, Jiangnan UniversityWuxi, Jiangsu 214122, China

Email：bhearts@jiangnan.edu.cn, caoyi@jiangnan.edu.cn, jhyu@jiangnan.edu.cn

Abstract

This paper addresses the direct kinematics of a special class of the 6-6 SPS Stewart manipulators in which the mobile and base platforms are two similar semisymmetrical hexagons. After proposing a mathematical model of the direct kinematic problem of this class of the 6-6 SPS Stewart manipulators, a multivariate polynomial equations set with respect to the moving platform position parameters and orientation parameters is constructed in which input-parameters are design parameters and six given link-lengths of this special class of the Stewart manipulators. Based on this multivariate polynomial equations set, the complete analytical expression of the solutions which have in total at most 28 in the complex domain, to the direct kinematics of this special class of the Stewart manipulators can be easily and directly solved by utilizing a sophisticated commercial symbolic computation software, MATHEMATICA. Computation results have shown that not only this approach presented in this paper is simple and efficient than existing approaches, but also an arduous and complicated derivation and solving task can be avoided and a lot of computation time can be saved. Examples of the direct kinematics analysis of a 6-6 SPS Stewart manipulator under investigation are given to demonstrate the aforementioned theoretical results. Direct kinematics analysis of this special class of the 6-6 SPS Stewart manipulators paves underlying theoretical grounds for the workspace analysis, path planning and control of this class of the Stewart manipulators.

Keywords：Stewart Manipulator; Direct Kinematics; Symbolic Computation; MATHEMATICA Software

1 Introduction

During the past two decades, parallel manipulator systems have become one of the research attentions in robotics. This popularity has been motivated by the fact that parallel manipulators possess some specific advantages in accuracy, rigidity, stiffness and load-carrying capacity, better dynamic performance, etc. over serial manipulators. Among them, the best-known parallel manipulator is the Stewart platform that was introduced as an aircraft simulator by Stewart in 1965. The general form of the Stewart platform consists of a moving platform (mobile platform) and a base one connected via six extensible links, often referred to as legs, through appropriate kinematic pairs either of the type SPS (where S stands for Spherical pair and P Prismatic pair) or equivalent to the SPS from a kinematic point of view.

One of the most important and quite challenging problems in robot kinematics is the direct kinematics analysis of a parallel manipulator that has initiated much research attentions during the past decade. A complete solution to the direct kinematic problem of a parallel manipulator is expected to provide a better insight into the kinematics performance of the mechanism and thus enhance its applications. As to Stewart manipulator, the direct kinematics analysis of the manipulator is to determine the position and orientation of the moving

platform of the manipulator for six given leg-lengths or link-lengths of the manipulator. It involves the solution of a system of highly nonlinear coupled algebraic equations with respect to the variables describing the moving platform pose (position and orientation) of the manipulator. As the complete solution to these algebraic equations mentioned above is quite difficult and challenging, the direct kinematic problem enjoys the central status in the research on the Stewart manipulators during the past decade, numerous approaches have been presented by various researchers in the direct kinematics analysis of many simplified types of 6-6 SPS Stewart manipulators, such as 3-3, 3-6, 4-4, 4-5, 4-6, and 5-5 SPS Stewart manipulators, see for instance, Griffis and Duffy [2]; Merlet [3]; Wen and Liang [4]; Sreenivasan and Waldron [5]; Wampler [6]; Husty [7]; Sika and Kocandrle et al [8]; Ku [9]; Wang and Xu [10]; Bonev and Ryu [11]; Lee and Shim [12]; Prakash Bande and Martin Seibt et al. [13]; and so on. However, to the best of the authors' knowledge, much less work on direct kinematic problem has been reported for the direct kinematics analysis of a special class of the 6-6 SPS Stewart manipulators, whose moving and base platforms are two similar semi- symmetrical hexagons, the most relevant investigations have been made in [7] and [10], without mentioning the direct kinematics analysis of the general form of the 6-6 SPS Stewart manipulators.

There are mainly two most common approaches to solve the direct kinematic problem of parallel manipulators:（1）to use numerical iteration method or（2）to use algebraic elimination method. For the numerical iteration method, as the motion of a parallel manipulator is continuous, at each iteration step, the initial estimate is the pose of the mobile platform at previous iteration step. The convergence problem of iteration method can be settled mainly by choosing a very small iteration step size, which entails efficient iteration algorithms and expensive computer hardware. Hence, numerical iteration method not only is not suitable for real-time applications but also cannot essentially determine the total (maximum) number of solutions of the direct kinematic problem of a parallel manipulator, which is of great importance to the understanding of kinematic performance of a parallel manipulator. For the latter, a feature of the algebraic elimination method is that by means of various efficient algebraic elimination methods, a univariate polynomial equation in one unknown can be derived theoretically, from which positions and orientations of the mobile platform of a parallel manipulator can be solved. However, one point to note is that the derivation of this univariate polynomial equation is a quite arduous and time-consuming task, because it would involve complicated variable transformations and require one to employ all kinds of elimination methods, and to perform very rigorous and complicated mathematical manipulation. In addition, even such a univariate polynomial equation can be successfully derived, to search all of its possible solutions once again becomes another challenging and time-consuming task.

In spite of these challenges, direct kinematics analysis of the Stewart manipulator can be very instructive due to its great importance. This paper addresses the direct kinematics of a special class of the 6-6 SPS Stewart manipulators in which the mobile and base platforms are two similar semisymmetrical hexagons, as shown in Figure 1. After proposing a mathematical model of the direct kinematic problem of this class of the 6-6 SPS Stewart manipulators, a multivariate polynomial equations set with respect to the moving platform position parameters and orientation parameters is constructed in which input- parameters are design parameters and six given link-lengths of this special class of the Stewart manipulators. Based on this multivariate polynomial equations set, the complete analytical expression of the solutions which have in total at most 28 in the complex domain, to the direct kinematics of this special class of the Stewart manipulators can be easily and directly solved by utilizing a sophisticated symbolic computation software, MATHEMATICA. Computation results have shown that not only this approach presented in this paper is simple and efficient than existing approaches mentioned above, but also an arduous and complicated derivation and solving task can be avoided and a lot of computation time can be saved. Examples of the direct kinematics analysis of a 6-6 SPS Stewart manipulator under

investigation are given to demonstrate the aforementioned theoretical results. Direct kinematics analysis of this class of the 6-6 SPS Stewart manipulators paves underlying theoretical grounds for the workspace analysis, path planning and control of this class of the Stewart manipulators.

The organization of this paper is as follows. Firstly, in Section 1, an overview of the research on direct kinematics analysis of parallel manipulators is concisely introduced. Then, in Section 2, the geometric modeling of a special class of 6-6 SPS Stewart manipulators is briefly described, and inverse kinematics problem of this special class of 6-6 SPS Stewart manipulators is discussed. In Section 3, direct kinematics problem of this special class of the Stewart manipulators is presented in detail. After proposing a mathematical model of the direct kinematic problem of this class of 6-6 SPS Stewart manipulators, a multivariate polynomial equations set with respect to the moving platform position parameters and orientation parameters is constructed in which input- parameters are design parameters and six given link-lengths of this special class of the Stewart manipulators. Based on this multivariate polynomial equations set, the complete analytical expression of the solutions which have in total at most 28 in the complex domain, to the direct kinematics of this special class of the Stewart manipulators can be easily and directly solved by utilizing a sophisticated symbolic computation software, MATHEMATICA. Examples of the direct kinematics analysis of a 6-6 SPS Stewart manipulator under investigation are provided to demonstrate the aforementioned theoretical results in Section 4. Finally, conclusions are presented in Section 5.

2 Inverse Kinematics Analysis of the Manipulator

A moving reference frame P-$X'Y'Z'$ and a fixed one O-XYZ are respectively attached to the moving platform and the base one of the manipulator, as shown in Figure 1, where origins P and O are corresponding geometric center of the moving platform and the base one. The Cartesian coordinates of the moving platform are given by the position of origin P with respect to the fixed frame, designated by (X, Y, Z), and the orientation of the moving platform is usually represented by three standard Z-Y-Z Euler angles (ϕ, θ, ψ).

Furthermore, design parameters of this special class of 6-6 SPS Stewart manipulators can be described as follows. The circumcircle radius of the base hexagon, $C_1C_2...C_5C_6$, is R_a, and the one of the moving hexagon, $B_1B_2...B_5B_6$, is R_b. While β_0 denotes the central angle of circumcircles of the hexagons corresponding to sides C_1C_2 and B_4B_5, as shown in Figure 1. The coordinates of six vertices, B_i (i=1, 2, …, 5, 6), of the moving platform are denoted by $\boldsymbol{B}_i'$: $(B_{iX}', B_{iY}', B_{iZ}')$ with respect to the moving frame, and $\boldsymbol{B}_i$: (B_{iX}, B_{iY}, B_{iZ}) in the fixed frame. Similarly, $\boldsymbol{C}_i$: (C_{iX}, C_{iY}, C_{iZ}) represent the coordinates of six vertices, C_i (i=1, 2, …, 5, 6), of the base hexagon with respect to the fixed frame.

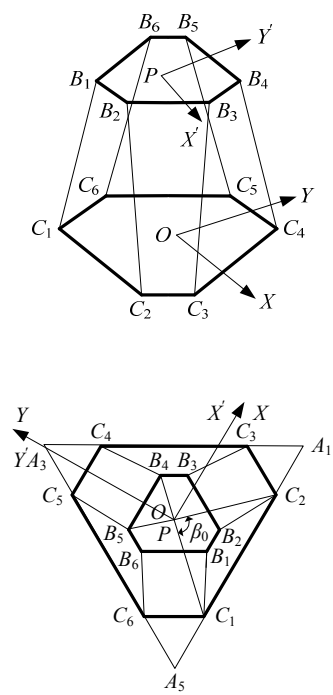


Figure 1 A schematic of a Stewart manipulator

The inverse kinematic problem of this special class of the Stewart manipulators is to determine six link-lengths of the manipulator for particular position (X, Y, Z) and orientation (ϕ, θ, ψ) of the moving platform of

the manipulator. Given a pose (position and orientation) of the moving platform, the necessary link-vectors, denoted by li (i=1, 2, …, 5, 6), can be straightforwardly computed by using the following formula

$$\boldsymbol{l}_i = \boldsymbol{B}_i - \boldsymbol{C}_i \quad (i=1, 2, \dots, 5, 6) \qquad (1)$$

where Bi, Ci (i=1, 2, …, 5, 6) are vertex vectors defined above. The norm of li, i.e., ‖li‖, representing six link-lengths of the manipulator for any given pose, can be obtained

$$\rho_i = \|\boldsymbol{l}_i\| = \|\boldsymbol{B}_i - \boldsymbol{C}_i\| \quad (i=1, 2, \dots, 5, 6) \qquad (2)$$

It can be seen clearly that for a particular pose of the moving platform, there will be a unique solution ρi (i=1, 2, …, 6), namely, six link-lengths, to the inverse kinematic problem of this special class of the Stewart manipulators and the inverse kinematics analysis of this special class of the 6-6 SPS Stewart manipulators is nearly straightforward.

3 Direct Kinematics Analysis of the Manipulator

Though the inverse kinematics analysis of this special class of the Stewart manipulators is nearly straightforward, the direct kinematic problem is a very difficult and challenging one that has initiated much research attentions during the past decade. This problem is to determine the position (X, Y, Z) and orientation (ϕ, θ, ψ) of the moving platform for six given link-lengths ρi (i=1, 2, …, 5, 6) of the manipulator, it is required to solve the following kinematic constraint equations

$$\|\boldsymbol{B}_i - \boldsymbol{C}_i\|^2 = \rho_i^2 \quad (i=1, 2, \dots, 5, 6) \qquad (3)$$

Geometrically, it is equivalent to the problem of placing a rigid body such that six of its given points Bi (i=1, 2, …, 5, 6) lie on six given spheres centered at Ci (i=1, 2, …, 5, 6) with the given radius ρi (i=1, 2, …, 5, 6), respectively. From a theoretical point of view, the pose of the moving platform of the manipulator can be solved by Eq.（3）.

Further investigation shows that Eq.（3）is a multivariate polynomial equations set with respect to the moving platform position parameters (X, Y, Z) and trigonometric functions of orientation parameters (ϕ, θ, ψ), but not a multivariate polynomial equations set directly in (ϕ, θ, ψ), of this class of the Stewart manipulators. Therefore, the solution to the direct kinematic problem, i.e., Eq.（3）, involves the solution of a system of highly nonlinear coupled algebraic equations in the variables describing the moving platform pose of this class of the Stewart manipulators. A complete analytical solution to the direct kinematic problem of this special class of the Stewart manipulators has not been presented yet owing to the fact of the complexity of this multivariate polynomial equations set and the periodicity of trigonometric functions.

3.1 Mathematical Model of the Direct Kinematic Problem of the Manipulator

Based on the conclusion presented in [13] that the orientation of a rigid body in three-dimensional (3-D) rotation space can always be represented by a 3 by 3 proper orthogonal identity matrix, a proper orthogonal identity matrix of order three, [T], will be introduced into the representation of the orientation of the moving platform of this special class of the 6-6 SPS Stewart manipulators with respect to the fixed frame, which can be written as follows

$$[\boldsymbol{T}] = \begin{bmatrix} \boldsymbol{u} & \boldsymbol{v} & \boldsymbol{u}\times\boldsymbol{v} \end{bmatrix} = \begin{bmatrix} u_1 & v_1 & u_2v_3 - u_3v_2 \\ u_2 & v_2 & u_3v_1 - u_1v_3 \\ u_3 & v_3 & u_1v_2 - u_2v_1 \end{bmatrix} \qquad (4)$$

where three unit column vectors, u=[u1, u2, u3]T, v=[v1, v2, v3]T and u×v =[u2v3–u3v2, u3v1–u1v3, u1v2–u2v1]T, of the matrix [T] describe direction cosine of the moving frame P-X'Y'Z', i.e., the orientation of the moving platform of this class of the Stewart manipulators which actually means that elements (u1, u2, u3, v1, v2, v3) is orientation parameters of the moving platform of the manipulator, with respect to the fixed frame O-XYZ, symbol × denotes cross-product of two vectors and [.]T denotes the transpose of [.]. Moreover, the following conditions should be satisfied which can guarantee the rotation matrix [T] is a proper orthogonal identity matrix

$$\begin{cases} \sum_{i=1}^{3} u_i^{\,2} = 1 \\ \sum_{i=1}^{3} v_i^{\,2} = 1 \\ \sum_{i=1}^{3} u_i v_i = 0 \end{cases} \quad (5)$$

Similarly, the position vector P=[X, Y, Z]T indicates the position of the origin P of the moving frame P-X'Y'Z' with respect to the fixed frame O-XYZ, then the relationship between the Cartesian coordinates of six vertices, Bi (i=1, 2, …, 5, 6), of the moving platform with respect to the moving frame denoted by $\boldsymbol{B}'_i : (B'_{iX}, B'_{iY}, B'_{iZ})$ and $\boldsymbol{B_i} : (B_{iX}, B_{iY}, B_{iZ})$ designated with respect to the fixed frame can be expressed as below, that is

$$\begin{Bmatrix} B_{iX} \\ B_{iY} \\ B_{iZ} \end{Bmatrix} = [\boldsymbol{T}] \begin{Bmatrix} B'_{iX} \\ B'_{iY} \\ B'_{iZ} \end{Bmatrix} + \boldsymbol{P} \quad (\text{i=1, 2, …, 5, 6}) \quad (6)$$

It is worth noting that the proposing of this mathematical mode of the direct kinematic problem of this special class of the 6-6 SPS Stewart manipulators is of great importance for the complete analytical solution to this problem owing to the fact that it provides the underlying theoretical ground which makes the complete analytical solution to the direct kinematic problem of the manipulator considered in the paper feasible.

3.2 Direct Kinematics Analysis of the Manipulator

Back substitution of expressions (4) and (6) into Eq. (3) and in combination with Eq. (5) will yield the following kinematic constraint equations after some rearrangements, such that

$$\begin{cases} f_{0i} = \left\| \boldsymbol{T}\boldsymbol{B}_i^{\cdot} + \boldsymbol{P} - \boldsymbol{C}_i \right\|^2 - \rho_i^2 = 0 \ (i = 1, 2, \ldots 5, 6) \\ f_{07} = \sum_{i=1}^{3} u_i^2 - 1 = 0 \\ f_{08} = \sum_{i=1}^{3} v_i^2 - 1 = 0 \\ f_{09} = \sum_{i=1}^{3} u_i v_i = 0 \end{cases} \quad (7)$$

Eq. (7), a system of nonlinear coupled algebraic equations in the nine variables (x, y, z; u1, u2, u3, v1, v2, v3), is a multivariate polynomial equations set with respect to the moving platform position parameters (X, Y, Z) and orientation parameters (u1, u2, u3, v1, v2, v3) of this class of the 6-6 SPS Stewart manipulators for any given set of design parameters, Ra, Rb and β0 , and six given link-lengths ρi (i=1, 2, …, 5, 6). Hence, the direct kinematic problem of this class of the Stewart manipulators is transformed into a complete analytical solution to these nonlinear coupled algebraic equations.

Though, the complexity of this multivariate polynomial equations set is, analytical expressions of the complete solution to these nonlinear coupled algebraic equations can be easily and directly accomplished by utilizing a sophisticated symbolic computation software, MATHEMATICA, consequently this approach presented in this paper is very simple and efficient than existing approaches mentioned above due to the fact an arduous and complicated derivation and solving task can be avoided and a lot of computation time can be saved. The solution to this multivariate polynomial equations set was successfully carried out in MATHEMATICA, on account of space limitations, we will not present here the detailed algorithm and computation process, for the same reason, more and complete descriptions of the software MATHEMATICA will not be described here as well, we refer the reader to the well-detailed one given in [14]. Most of all, Computation results first show that the maximum number of the complete analytical solution to the direct kinematic problem of this special class of the 6-6 SPS Stewart manipulators is up to 28 in the complex domain for any given set of design parameters, Ra, Rb and β0, and six given link-lengths ρi (i=1, 2, …, 5, 6) of the manipulator considered in the present paper.

As mentioned above, for any given set of design parameters, Ra, Rb and β0, and six given link-lengths ρi (i=1, 2, …, 5, 6) of this special class of the 6-6 SPS Stewart manipulators, the direct kinematic problem of the manipulator have at most 28 solutions in the complex domain indicating that there are at mot 28 groups of (x, y, z; u1, u2, u3, v1, v2, v3) in the complex domain corresponding to a given set of Ra, Rb , β0 and

ρi (i=1, 2, …, 5, 6) of this class of the Stewart manipulators. For each group of (u1, u2, u3, v1, v2, v3), Euler angles (ϕ, θ, ψ), which also represents the orientation of the moving platform in 3-D rotation space with respect to the fixed frame O-XYZ , can be obtained by the following formula

$$\begin{cases} \phi = A\tan 2(u_3 v_1 - u_1 v_3,\ u_2 v_3 - u_3 v_2) \\ \theta = A\tan 2(\sqrt{u_3^2 + v_3^2}, u_1 v_2 - u_2 v_1) \\ \psi = A\tan 2(v_3, -u_3) \end{cases} \quad (8)$$

where function c=Atan 2(a, b), four-quadrant inverse tangent function, returns an array c the same size as b and a containing the element-by-element, four-quadrant inverse tangent of the real parts of a and b. From Eq.（8）, it can be seen clearly that a one-to-one mapping exists between (ϕ, θ, ψ) and (u1, u2, u3, v1, v2, v3). That is to say, for every value of (u1, u2, u3, v1, v2, v3), there is a unique corresponding (ϕ, θ, ψ); and for every value of (ϕ, θ, ψ), there is a unique image (u1, u2, u3, v1, v2, v3) in the image space. This mapping is very important for the following orientation computation.

So it can be safely concluded that for any given set of design parameters, Ra, Rb and β0, and six given link-lengths ρi (i=1, 2, …, 5, 6) of this special class of the 6-6 SPS Stewart manipulators, the direct kinematic problem of this class of the Stewart manipulators have at mot 28 groups of (x, y, z; u1, u2, u3, v1, v2, v3) or (x, y, z; ϕ, θ, ψ) in the complex domain representing positions (x, y, z) and orientations (ϕ, θ, ψ), respectively, of the moving platform of the manipulator with respect to the base platform.

4 Numerical Examples

As stated in sections 2 and 3, for any given set of design parameters, Ra, Rb and β0 of this special class of the 6-6 SPS Stewart manipulators considered in this paper, there is a unique solution ρi (i=1, 2, …, 6), namely, six link-lengths of the manipulator, to the inverse kinematic problem of this class of the Stewart manipulators corresponding to a particular position (X, Y, Z) and orientation (ϕ, θ, ψ) of the moving platform of the manipulator; whereas, there are in total at most 28 solutions, i.e., (x, y, z; u1, u2, u3, v1, v2, v3) or (x, y, z; ϕ, θ, ψ), in the complex domain which represent positions (x, y, z) and orientations (ϕ, θ, ψ) of the moving platform, respectively, with respect to the base platform corresponding to six given link- lengths ρi (i=1, 2, …, 5, 6) of the manipulator. In order to demonstrate the aforementioned theoretical results, the inverse kinematic problem and direct kinematic problem of a special class of the 6/6-SPS Stewart manipulators will be numerically studied sequentially. The design parameters of the manipulator are given as follows Ra=8 cm, Rb=6 cm and β0=90° where meanings of Ra, Rb and β0 have been defined in Section 2.

Only two examples are presented here, the position and orientation of the moving platform of the manipulator described above is (x, y, z)=(0 cm, 0 cm, 16 cm), (ϕ, θ, ψ)=(0°, 0°, 0°) in the first example and will be (x, y, z)=(6cm, 0 cm, 16 cm), (ϕ, θ, ψ)=(3°, 3°, 45°) in the second example. Then, solutions ρi (i=1, 2, …, 6), namely, six link-lengths of the manipulator, to the inverse kinematic problem of the manipulator described above can be calculated by Eq.（2）, as shown in Table 1, corresponding to two above-mentioned poses of the moving platform of the manipulator described above.

For the current design parameters, Ra, Rb and β0, and two groups of six link-lengths ρi (i=1, 2, …, 5, 6) shown in Table 1 of the manipulator described above, solutions, as listed in Table 2 (from the second column 2 to the tenth column), to the direct kinematic problem of the manipulator described above were successfully accomplished in MATHEMATICA by solving Eq.（7）. From Table（2）, it can be seen clearly that for the current design parameters, Ra, Rb and β0, and two groups of six given link-lengths ρi (i=1, 2, …, 5, 6) of the manipulator described above, the direct kinematic problem of the manipulator actually have 28 groups of (x, y, z; u1, u2, u3, v1, v2, v3) in the complex domain representing positions (x, y, z) and orientations (u1, u2, u3, v1, v2, v3), respectively, of the moving platform of the manipulator with respect to the base platform. It suggests that this approach introduce in this paper works well and may be a good starting point for future on this problem. Furthermore, out of 28 solutions only 16 are

real roots in the first example and 4 in the second example, for these real roots, then, Euler angles (ϕ, θ, ψ), as shown in Table 2 as well (from the eleventh column to the thirteenth column), representing the orientation of the moving platform in 3-D rotation space with respect to the fixed frame O-XYZ , can be obtained by Eq.（8）.

Table 1　Solutions to the inverse kinematic problem

Sol.	ρ_1 (cm)	ρ_2 (cm)	ρ_3 (cm)	ρ_4 (cm)	ρ_5 (cm)	ρ_6 (cm)
1	16.519	16.519	16.519	16.519	16.519	16.519
2	22.106	15.516	16.628	16.226	21.917	21.954

Table 2　Solutions to the direct kinematic problem

Sol.	x(cm)	y(cm)	z(cm)	u_1	u_2	u_3	v_1	v_2	v_3	ϕ(°)	θ(°)	ψ(°)
1	0	0	16	1.00	0	0	0	1.00	0	0	0	0
	0	0	9.47223	-1.0000	0	0	0	-1.0000	0	180	0	0
	0	7.34957	12.2175	1.0000	0	0	0	-0.26814	0.96338	-90	105.554	90
	-6.3649	-3.6748	12.2175	0.04890	-0.5491	-0.8343	-0.54912	0.68297	-0.4817	30.00	105.554	-30.0
	6.36491	-3.6748	12.2175	0.04890	0.54912	0.83431	0.54912	0.68297	-0.4817	150.00	105.554	-150.
	0	-9.1469	8.46548	-1.0000	0	0	0	0.57826	-0.8159	-90	125.329	-90
	-7.9215	4.57346	8.46548	0.18370	-0.6834	-0.7066	-0.68341	-0.60543	0.40792	150.00	125.329	30.0
	7.92146	4.57346	8.46548	0.18370	0.68341	0.70655	0.68341	-0.60543	0.40792	30.00	125.329	150
	0	0	-16	1.0000	0	0	0	1	0	0	0	0
	0	0	-9.4722	-1.0000	0	0	0	-1.0000	0	180	0	0
	0	7.34957	-12.218	1.0000	0	0	0	-0.26814	-0.9634	90	105.554	-90
	-6.3649	-3.6748	-12.218	0.04890	-0.5491	0.83431	-0.54912	0.68297	0.48169	-150.00	105.554	150.
	6.3649	-3.675	-12.218	0.04890	0.54912	-0.8343	0.54912	0.68297	0.48169	-30.0	105.554	30.0
	0	-9.147	-8.4655	-1.0000	0	0	0	0.57826	0.81585	90	125.329	90
	-7.9215	4.57346	-8.4655	0.18370	-0.6834	0.70655	-0.68341	-0.60543	-0.4079	-30.00	125.329	-150
	7.92146	4.57346	-8.4655	0.18370	0.68341	-0.7066	0.68341	-0.60543	-0.4079	-150.00	125.329	-30.0
	-42.217	24.374	42.080**i**	7.3085	-3.6422	-8.104**i**	-3.6422	3.1028	4.6790**i**			
	-42.217	24.374	-42.08**i**	7.3085	-3.6422	8.1043**i**	-3.6422	3.1028	-4.679**i**			
	0	-48.748	42.080**i**	1.0000	0	0	0	9.4113	-9.358**i**			
	0	-48.748	-42.08**i**	1.0000	0	0	0	9.4113	9.3580**i**			
	42.217	24.374	42.080**i**	7.3085	3.6422	8.1043**i**	3.6422	3.1028	4.6790**i**			
	42.217	24.374	-42.08**i**	7.3085	3.6422	-8.104**i**	3.6422	3.1028	-4.679**i**			
	-11.889	6.8640	0.5198**i**	0.77654	-1.0257	-0.809**i**	-1.0257	-0.40782	0.4673**i**			
	-11.889	6.8640	-0.519**i**	0.77654	-1.0257	0.8094**i**	-1.0257	-0.40782	-0.467**i**			
	0	-13.728	-0.519**i**	-1.0000	0	0	0	1.3687	0.9346**i**			
	0	-13.728	0.5198**i**	-1.0000	0	0	0	1.3687	-0.934**i**			
	11.889	6.8640	0.5198**i**	0.77654	1.0257	0.8094**i**	1.0257	-0.40782	0.4673**i**			
	11.889	6.8640	-0.519**i**	0.77654	1.0257	-0.809**i**	1.0257	-0.40782	-0.467**i**			
2	6	0	16	0.17678	0.91856	-0.3535	-0.883883	0.306186	0.35355	30	30	45
	6.05383	0.18656	14.5311	-0.3383	0.92320	-0.1826	-0.879239	-0.241129	0.41085	41.7446	26.7089	66.0777
	6	0	-16	0.17678	0.91856	0.35355	-0.883883	0.306186	-0.3535	-150.0	30.000	-135
	6.05383	0.18656	-14.531	-0.3383	0.92320	0.18225	-0.879239	-0.241129	-0.4108	-138.26	26.7089	-113.92
	378.391	-173.55	419.39**i**	-49.513	33.0458	-59.52**i**	31.2434	-19.4391	36.784**i**			
	378.391	-173.55	-419.4**i**	-49.513	33.0458	59.52**i**	31.2434	-19.4391	-36.78**i**			
	-321.87	-229.75	398.63**i**	-53.437	-27.368	60.03**i**	-29.1701	-13.6645	32.19**i**			
	-321.87	-229.75	-398.6**i**	-53.437	-27.368	-60.03**i**	-29.1701	-13.6645	-32.19**i**			
	6.79951	-61.288	54.843**i**	0.93645	0.98754	-0.923**i**	-0.814907	11.6409	-11.63**i**			
	6.79951	-61.288	-54.84**i**	0.93645	0.98754	0.923**i**	-0.814907	11.6409	11.63**i**			
	18.3583	-17.040	11.32**i**	1.19923	1.98475	-2.092**i**	0.182308	4.26878	-4.154**i**			
	18.3583	-17.040	-11.32**i**	1.19923	1.98475	2.092**i**	0.182308	4.26878	4.154**i**			
	4.7243 +0.401**i**	-20.1471-6. 16**i**	-7.80566+1 1.7**i**	-1.29492+0. 30**i**	0.8085 +0.03**i**	-0.319601-1 .2**i**	-0.993942+ 0.0346**i**	2.31079 +1.3714**i**	1.36162 -2.302**i**			

Continued

Sol.	x(cm)	y(cm)	z(cm)	u_1	u_2	u_3	v_1	v_2	v_3	ϕ(°)	θ(°)	ψ(°)
	4.7243 -0.401**i**	-20.1471+6.16**i**	-7.80566-11.7**i**	-1.29492-0.30**i**	0.8085 -0.03**i**	-0.319601+1.2**i**	-0.993942-0.0346**i**	2.31079 -1.3714**i**	1.36162 +2.302**i**			
	-3.6946	22.347	-0.379**i**	4.76167	0.08218	-4.656**i**	-1.72026	1.03519	1.7409**i**			
	-3.6946	22.347	0.379**i**	4.76167	0.08218	4.6562**i**	-1.72026	1.03519	-1.740**i**			
	-9.6610	17.9967	13.453**i**	1.05868	-0.4326	0.5549**i**	-2.23501	-1.91716	-2.769**i**			
	-9.6610	17.9967	-13.45**i**	1.05868	-0.4326	-0.554**i**	-2.23501	-1.91716	2.7696**i**			
	-5.74497 -4.44**i**	7.23426 -0.617**i**	14.5464 +1.26**i**	0.599133 +0.89**i**	-0.0947 -0.38**i**	-1.32839 +0.43**i**	-1.89716 -0.38331**i**	-0.5197 +1.00103**i**	0.15668 -1.32**i**			
	-5.74497 +4.44**i**	7.23426 +0.617**i**	14.5464 -1.26**i**	0.599133 -0.89**i**	-0.0947 +0.38**i**	-1.32839 -0.43**i**	-1.89716 +0.38331**i**	-0.5197 -1.00103**i**	0.15668 +1.32**i**			
	5.0987 +8.057**i**	5.17455 +3.894**i**	-12.2 +7.5906**i**	0.30898 +1.519**i**	1.70353 +0.695**i**	-1.44864 +1.14**i**	-0.0989099 +0.695**i**	-0.454458 +0.8477**i**	-1.44375 -0.31**i**			
	5.0987 -8.057**i**	5.17455 -3.894**i**	-12.2 -7.5906**i**	0.30898 -1.519**i**	1.70353 -0.695**i**	-1.44864 -1.14**i**	-0.0989099 -0.695**i**	-0.454458 -0.8477**i**	-1.44375 +0.31**i**			
	4.7243 -0.401**i**	-20.1471 +6.16**i**	7.80566 +11.69**i**	-1.29492 -0.30**i**	0.8085 -0.034**i**	0.319601 -1.15**i**	-0.993942 -0.03460**i**	2.31079 -1.3714**i**	-1.36162 -2.30**i**			
	4.7243 +0.401**i**	-20.1471 -6.16**i**	7.80566 -11.69**i**	-1.29492 +0.30**i**	0.8085 +0.034**i**	0.319601 +1.15**i**	-0.993942 +0.03460**i**	2.31079 +1.3714**i**	-1.36162 +2.30**i**			
	15.0987 -8.057**i**	5.17455 -3.894**i**	12.2+7.5906**i**	0.30898 -1.519**i**	1.7035 -0.695**i**	1.44864 +1.141**i**	-0.0989099 -0.6951**i**	-0.454458 -0.84770**i**	1.44375 -0.314**i**			
	15.0987 +8.057**i**	5.17455 +3.894**i**	12.2-7.5906**i**	0.30898 +1.519**i**	1.7035 +0.695**i**	1.44864 -1.141**i**	-0.0989099 +0.6951**i**	-0.454458 +0.84770**i**	1.44375 +0.314**i**			
	-5.74497 -4.44**i**	7.23426 -0.617**i**	-14.5464 -1.26**i**	0.5991 +0.8944**i**	-0.0947 -0.383**i**	1.32839 -0.430**i**	-1.89716 -0.38331**i**	-0.5197 +1.00103**i**	-0.1566 +1.320**i**			
	-5.74497 +4.44**i**	7.23426 +0.617**i**	-14.5464 +1.26**i**	0.5991 -0.8944**i**	-0.0947 +0.383**i**	1.32839 +0.430**i**	-1.89716 +0.38331**i**	-0.5197 -1.00103**i**	-0.1566 -1.320**i**			

5 Conclusions and Future Work

This paper addresses the direct kinematics of a special class of the 6-6 SPS Stewart manipulators in which the mobile and base platforms are two similar semisymmetrical hexagons. After proposing a mathematical model of the direct kinematic problem of this class of the 6-6 SPS Stewart manipulators, a multivariate polynomial equations set with respect to the moving platform position parameters and orientation parameters is constructed in which input-parameters are design parameters and six given link-lengths of this special class of the Stewart manipulators. Based on this multivariate polynomial equations set, the complete analytical expression of the solutions which have in total at most 28 in the complex domain, to the direct kinematics of this special class of the Stewart manipulators can be easily and directly solved by utilizing a sophisticated commercial symbolic computation software, MATHEMATICA. Computation results have shown that not only this approach presented in this paper is simple and efficient than existing approaches, but also an arduous and complicated derivation and solving task can be avoided and a lot of computation time can be saved. Examples of the direct kinematics analysis of a 6-6 SPS Stewart manipulator under investigation are given to demonstrate the aforementioned theoretical results. Direct kinematics analysis of this special class of the 6-6 SPS Stewart manipulators paves underlying theoretical grounds for the workspace analysis, path planning and control of this class of the Stewart manipulators.

It should be noted that the investigation of the direct kinematic problem of the Stewart manipulators presented in this paper only deals with a special yet very useful class of the Stewart manipulators whose moving and base platforms are two similar semisymmetrical hexagons, this work, just a good starting point for our future work on this problem, can be extended to the general Stewart manipulator, which is the current aim of the authors. This work is under investigation at our laboratory.

References

[1] D. Stewart, "A Platform with Six Degrees of Freedom", Proc. of the Institution of Mechanical Engineers, London, UK, 1965, pp.371~378

[2] M. Griffis, J. Duffy, "A Forward Displacement Analysis of a Class of Stewart Platforms", Journal of Robotic Systems, Vol.10, No.3, 1989, pp.703~720

[3] J.-P. Merlet, "Direct Kinematics and Assembly Modes of Parallel Manipulators", The International Journal of Robotics Research, Vol.11, No.2, 1992, pp.150~162

[4] F. A. Wen, C. G. Liang, "Displacement Analysis of the 6-6 Stewart Platform Mechanisms", Mechanism and Machine Theory, Vol.29, No.4, 1994, pp.547~557

[5] S. V. Sreenivasan, K. J. Waldron, "Closed-Form Direct Displacement Analysis of a 6-6 Stewart Platform", Mechanism and Machine Theory, Vol.29, No.6, 1994, pp. 855~864

[6] C. W. Wampler, "Forward Displacement Analysis of General Six-in-Parallel Stewart Platform Manipulators Using Soma Coordinates", Mechanism and Machine Theory, Vol.31, No.3, 1996, pp.331~337

[7] Z. Sika, P. Kocandrle et al "An Investigation of Properties of the Forward Displacement Analysis of the Generalized Stewart Platform by Means of General Optimization Methods", Mechanism and Machine Theory, Vol.33, No.3, 1998, pp.245~253

[8] Huang Zhen, Kong Linfu et al, Theory of Parallel Robotic Mechanisms and Control, Beijing: China Mechanical Press, 1997

[9] D. M. Ku, "Direct Displacement Analysis of a Stewart Platform Mechanism", Mechanism and Machine Theory, Vol. 34, No.3, 1999, pp.453~465

[10] I. A. Bonev, J. Ryu, "A New Method for Solving the Direct Kinematics of General 6-6 Stewart Platforms Using Three Linear Extra Sensors", Mechanism and Machine Theory, Vol.35, No.3, 2000, pp.423~436

[11] T. Y. Lee, J. K Shim, "Forward Kinematics of the General 6-6 Stewart Platform Using Algebraic Elimination", Mechanism and Machine Theory, Vol.36, No.9, 2001, pp.1073~1085

[12] Y. X. Wang, Y. M. Wang et al, "Direct Displacement and Configuration Analyses of Symmetrical Stewart Platform Mechanisms", China Mechanical Engineering, Vol.13, No.9, 2002, pp.734~736

[13] Xiong Youlun, Ding Han et al, Robotics, Beijing: China Mechanical Press, 1993

[14] S. Wolfram, The Mathematica Book, Cambridge：Cambridge University Press, 1996

Improve and Secure a Mediated Certificateless Signature Scheme*

Xuezhong Qian Xu Wang

School of Information Technology, Jiangnan University, Wuxi 214122, China

Email: qxzvb@163.com, pieces1229@hotmail.com

Abstract

Certificateless public key cryptography can be viewed as a model that is intermediate between traditional public key cryptography and identity-based public key cryptography. Mediated certificateless cryptography equips certificateless cryptography with instantaneous revocation function. Yang et al. recently proposed an efficient mediated certificateless signature scheme and claimed that their scheme is secure. This paper shows that their scheme suffers from the key replacement attack. An improved scheme is subsequently proposed and formal security proof presented in the paper demonstrates that the improved scheme is existentially unforgeable against fully- adaptive chosen message attack in the random oracle model. With our complementary efforts, the improved mediated scheme is provably secure.

Keywords : Mediated Certificateless Signature; Certificateless Public Key Cryptography; Cryptanalysis; Security Model; Pairing

1 Introduction

In 2001, Boneh et al. [1] first introduced a method for obtaining instantaneous revocation in public key cryptosystem. The basic idea is to use an online mediator called security mediator (SEM) to control users' cryptographic capabilities. Once the SEM being told that a user's ability is to be revoked, he refuses to cooperate with the user so that the user can not perform any signing or decryption operations. In 2003, Al-Riyami and Paterson [2] introduced an intermediate paradigm, known as certificateless public key cryptography (CLPKC), between the traditional public key cryptography (TPKC) and the identity-based cryptography (IBC). CLPKC inherits the advantages and eliminates the disadvantages from TPKC and IBC. In 2005, Ju et al. [3] presented an immediate revocation strategy in the CLPKC and proposed the corresponding mediated encryption and signature schemes. Their mediated certificateless signature (MCLS) scheme was based on the certificateless signature (CLS) scheme of [2], which was proved insecure in [4]. Yang et al. [5] and Yap et al. [6] both showed that Ju et al.'s MCLS scheme was insecure and proposed their improved MCLS schemes respectively. In [5], Yang et al. claimed that their improved scheme was secure and presented sketch proof. In this paper, we show that their scheme is unfortunately insecure and that there are some flaws in their proof. We improve their scheme and formally prove the improved MCLS scheme secure in a strong security model.

2 Security Model of MCLS

The first formalized security model for mediated certificateless cryptography was introduced in [7] and was mainly about encryption. Then the idea was

* This study is supported by Natural Science Foundation of Jiangsu Province of China (BK2003017).

extended to MCLS in [6]. There are three kinds of participant – key generation center (KGC), SEM and a set of users in a MCLS scheme. While neither the master key of KGC nor the secret key of a user is available to a SEM, a malicious SEM is explicitly weaker than the KGC as a Type II adversary or than a user as a Type I adversary. Hence, the security model of a MCLS scheme is similar to that of a CLS scheme. Nevertheless, we have to consider whether a SEM will accidentally leak his secret to a user. Therefore, some extra oracles related to the SEM have been presented in the security model. If these oracles are really helpful to the adversary, he might break MCLS schemes.

The two types of adversary in the MCLS security model can be briefly described as follows:

1) Type I adversary A_I represents a dishonest user or an attacker from the outside, who has already accessed to a challenged user's secret/private key or has the ability of replacing users' public keys, but has no access to the KGC's master key. Furthermore, he can access to either the partial secret key of the SEM or that of the challenged user.

2) Type II adversary A_{II} acts as an honest-but-curious KGC who owns the master key, but has no access to a challenged user's secret key and can not replace the challenged user's public key either.

It is noteworthy that another adversary in CLPKC called malicious-but-passive adversary [8] was proposed recently. However, we are not going to consider this kind of adversary because it is hard to construct a secure scheme against the adversary for his tremendous ability of arbitrarily generating parameters.

Now we summarize four kinds of existential forgery attack according to different signing oracles. They are

1) existential forgery under no-message attack, provided without signing oracles,

2) existential forgery under partially-adaptive chosen message attack, equipped with signing oracles to get valid signatures only if the query is on an unchallenged user,

3) existential forgery under adaptive chosen message attack, which accesses to valid signatures from signing oracles as long as the public key of the queried identity has not been replaced,

4) existential forgery under fully-adaptive chosen message attack, which can achieve valid signatures from signing oracles almost under any circumstances.

Obviously, a security model against the existential forgery under fully-adaptive chosen message attack is the securest one.

Details of challenge games and oracles will be described in section 6.

3 Definitions

Throughout the paper, G_1 denotes an additive group of prime order q and G_2 a multiplicative group of the same order. We let P denote a generator of G_1.

Definition 1 (Pairing). A pairing is a map e: $G_1 \times G_1 \rightarrow G_2$ with bilinear, non-degenerate and computable properties.

Definition 2 (Computational DH (CDH) Assumption). Given two elements aP, $bP \in G_1$, where a, b are selected uniformly at random from $\mathbf{Z}_q^*$, the CDH problem (CDHP) in G_1 is to output abP. An adversary A has at least an ε advantage if $\Pr[A(P, aP, bP)=abP] \geq \varepsilon$. We say that the ($\varepsilon$, t)-CDH assumption holds in G_1 if no algorithm running in polynomial time at most t can solve the CDHP in G_1 with non-negligible advantage at least ε.

Definition 3 (Paring-based Bilinear DH (BDH) Assumption [9]). Given a pairing (e, G_1, G_2) and three elements aP, bP, $cP \in G_1$ where a, b, c are selected uniformly at random from $\mathbf{Z}_q^*$, the BDH problem (BDHP) in G_1 is to output $e(P, P)^{abc}$. An adversary A has at least an ε advantage if $\Pr[A(P, aP, bP, cP)=e(P, P)^{abc}] \geq \varepsilon$. We say that the ($\varepsilon$, t)-BDH assumption holds in G_1 if no algorithm running in polynomial time at most t can solve the BDHP in G_1 with non-negligible advantage at least ε.

Definition 4 (Paring-based Generalized Bilinear DH (GBDH) Assumption [2]). Given a pairing (e, G_1, G_2) and three elements aP, bP, $cP \in G_1$ where a, b, c are selected uniformly at random from $\mathbf{Z}_q^*$, the GBDH

problem (GBDHP) in G_1 is to output $e(P, Q)^{abc}$ where Q is (not necessarily randomly) selected from G_1^*. An adversary A has at least an ε advantage if $\Pr[A(P, aP, bP, cP)=<Q, e(P, Q)^{abc}>] \geq \varepsilon$. We say that the ($\varepsilon$, t)-GBDH assumption holds in a group G_1 if no algorithm running in polynomial time at most t can solve the NGBDHP in G_1 with non-negligible advantage at least ε.

Definition 5 (Non-pairing-based Generalized Bilinear DH (NGBDH) Assumption [10]). Given a group G_1 and two elements aP, $bP \in G_1$ where a, b are selected uniformly at random from $\mathbf{Z}_q^*$, the NGBDH problem (NGBDHP) in G_1 is to output ($abcP$, cP) where $Q=cP$ is selected from G_1^*. An adversary A has at least an ε advantage if $\Pr[A(P, aP, bP)=(abcP, cP)] \geq \varepsilon$. We say that the ($\varepsilon$, t)-NGBDH assumption holds in a group G_1 if no algorithm running in polynomial time at most t can solve the NGBDHP in G_1 with non-negligible advantage at least ε.

4 Review of Yang et al.'s Scheme

Their mediated certificateless signature scheme consists of four algorithms, which can be presented as follows.

Setup: Given a security parameter k, the KGC generates (e, G_1, G_2), chooses a generator $P \in G_1$, randomly picks $s \in \mathbf{Z}_q^*$, computes $P_0=sP$ and chooses two cryptographic hash functions $H:\{0,1\}^* \times G_1 \rightarrow \mathbf{Z}_q^*$ and $H_1:\{0,1\}^* \rightarrow G_1$. KGC finally publishes the parameters as (e, G_1, G_2, P, P_0, H, H_1) and keeps s as his secret master key.

Key-Generation: For a user with the identity ID, KGC computes $Q_{ID}=H_1(ID)$ and $D_{ID}=sQ_{ID}$. He chooses a random element $D_{ID,\text{user}} \in G_1$ as the user's partial secret key and computes $D_{ID,\text{SEM}}= D_{ID} - D_{ID,\text{user}}$ as the SEM's one. KGC sends them to the user and the SEM respectively through a confidential and authentic channel. The user selects $x_{ID} \in \mathbf{Z}_q^*$ as his secret key and constructs the corresponding public key $P_{\text{pub}}=(X_{ID}, Y_{ID})=(x_{ID}P, x_{ID}P_0)$.

Sign: To sign a message M, the user with the identity ID interacts with the SEM via a secure channel as follows:

1) The user chooses a random number $r_1 \in \mathbf{Z}_q^*$, computes $R_1= r_1X_{ID}$ and sends (M, R_1, ID, X_{ID}, Y_{ID}) to the SEM.

2) The SEM checks whether the ID is revoked. If not, he checks whether the equation $e(X_{ID}, P_0)= e(Y_{ID}, P)$ holds. If the equation holds, the SEM randomly picks $r_2 \in \mathbf{Z}_q^*$ and computes

$$R_2=r_2X_{ID},\ R=R_1+R_2,\ h=H(M, R) \text{ and}$$
$$U_{\text{SEM}}= r_2Y_{ID}+ hD_{ID,\text{SEM}}. \quad (1)$$

Then he sends (R, U_{SEM}) back to the user.

3) Receiving (R, U_{SEM}), the user computes

$$h=H(M, R),\ U_{\text{user}}=r_1X_{ID}+ hD_{ID,\text{user}},$$
$$U=x_{ID}(U_{\text{SEM}} +U_{\text{user}}), \quad (2)$$

and checks whether $e(P, U)=e(Y_{ID}, R+hQ_{ID})$ holds. If it does, the signature will be $\sigma=(R, U)$.

Verify: Given $\sigma=(R, U)$ on the message M under the identity ID, the verifier computes $h=H(M,R)$ and $Q_{ID}=H_1(ID)$. He accepts the signature if and only if $e(P,U)=e(Y_{ID}, R+hQ_{ID})$ holds.

5 Cryptoanalysis of Yang et al.'s Scheme

5.1 Key replacement attack against the scheme

Now we show that Yang et al.'s scheme is vulnerable to the public key replacement attack launched by a Type I adversary since there is no authenticity of the public key.

The attack against the scheme is really a strong one, which is a no-message attack, i.e. no signing oracle is needed. That means the forger does not need any help from the SEM's or the users' signing messages before he outputs a forgery.

Forgery: To forge a signature for a message M_A with an identity ID_A, the forger randomly selects $R_A \in G_1^*$ and $y_A \in \mathbf{Z}_q^*$, computes $Q_A=H_1(ID_A)$ and $h_A=H(M_A, R_A)$, and sets $U_A= y_A(R_A+h_AQ_A)$. Then, he

computes $Y_A = y_A P$ and replaces Y_{ID} with Y_A. The forgery is $\sigma = (R_A, U_A)$.

Plainly, the forged signature is valid since it can be verified by the Verify algorithm.

5.2 Analysis of Yang et al.'s proof of their scheme

The flaws in their proof are as follows.

First of all, the security model of a related CLS scheme is in general supposed to be against the Type I adversary and the Type II adversary, which was not considered in [5].

On the proof of theorem 3 in [5], the attacker B, as a simulator, was supposed to generate indistinguishable parameters for the forger C, and otherwise C might recognize that it was not a real attack and subsequently halt. Besides, B still can not solve BDHP although B had set $P_1 = x_A P$ since C might have changed the value of P_1 during the game.

Furthermore, the scheme should not be called secure against existential forgery under adaptive chosen message attack for no signing oracles presented.

6 Improvement and Security Proof

6.1 Improvement of Yang et al.'s scheme

The improvement is quite straightforward because the original scheme actually is an articulate and efficient one except for neglecting the adversary's ability of key replacement. When a verifier checks the validity of a signature with the public key submitted by a signer, firstly he should make sure that the public key has an important authentic ingredient – the master key. Hence, the solution is to add a check to the beginning of the Verify algorithm as its first step and the check is whether $e(X_{ID}, P_0) = e(Y_{ID}, P)$ holds.

6.2 Security proof of the improved scheme

Theorem 1. The scheme is secure against Type I existential forgery under fully-adaptive chosen message attack in the random oracle model under NGBDH assumption.

Proof: Let A_I be a forger that breaks the improved scheme. We show that how B make use of A_I to solve a NGBDHP instance (P, aP, bP).

B sets $P_0 = aP$, generates other parameters and then sends parameters $(e, G_1, G_2, P, P_0, H, H_1)$ to A_I. Let H and H_1 be random oracles controlled by B. Note that B actually has no idea about his master key a. B also maintains a list of tuples $(ID_i, Q_i, d_i, D_i, D_{i,\mathrm{SEM}}, D_{i,\mathrm{user}}, x_i, P_i)$ denoted by H_1^{list} and a list of tuples (M_i, U_i, h_i) denoted by H^{list}. B randomly selects an index j such that $1 \le j \le q_H$, where q_H represents the maximum number of different queries to the random oracle H_1. B simulates oracles as follows:

H_1 Oracle: When A_I queries H_1 with an input ID_i, B checks H_1^{list} and outputs Q_i if such a value exists. Otherwise B sets $Q_i = bP$ if $i = j$. If $i \neq j$, B randomly chooses $d_i \in \mathbf{Z}_q^*$ and computes $Q_i = d_i P$. B then updates H_1^{list} and outputs Q_i.

H Oracle: When A_I queries H with inputs (M_i, U_i), B checks H^{list} and outputs h_i if such a value exists. Otherwise B randomly selects $h_i \in \mathbf{Z}_q^*$, updates the list and outputs h_i.

SEM-Extract Oracle: When A_I queries the oracle with ID_i, B outputs the value if $D_{i,\mathrm{SEM}}$ exists. Otherwise, if $i \neq j$, B computes $D_i = d_i(aP)$, chooses a random element $D_{i,\mathrm{user}} \in G_1$ and computes $D_{i,\mathrm{SEM}} = D_i - D_{i,\mathrm{user}}$. If $i = j$ and $D_{i,\mathrm{user}} = \perp$, then B selects a random element $D_{i,\mathrm{SEM}} \in G_1$. Otherwise, B aborts. B finally updates H_1^{list} and outputs $D_{i,\mathrm{SEM}}$.

It is worth noting that D_j in fact equals abP.

User-Extract Oracle: It is similar to the oracle above. B outputs $D_{i,\mathrm{user}}$.

Secret-Extract Oracle: When A_I queries the oracle with ID_i, B outputs x_i if x_i exists. Otherwise, if P_i doesn't exist, B selects $x_i \in \mathbf{Z}_q^*$ and constructs the corresponding $P_i = (X_i, Y_i) = (x_i P, x_i P_0)$. B then updates H_1^{list} and outputs x_i.

Public-Require Oracle: When A_I queries the oracle with ID_i, if P_i exists, B outputs it. Otherwise B selects $x_i \in \mathbf{Z}_q^*$ and constructs $P_i = (X_i, Y_i) = (x_i P, x_i P_0)$. B then

updates H_1^{list} and outputs P_i.

Public-Replacement Oracle: Suppose the query is on ID_i and P_i'. B replaces the corresponding P_i with P_i' and sets $x_i = \perp$. B then updates H_1^{list}.

SEM-Sign Oracle: When A_I queries the oracle with (ID_i, M_i, $R_{i,1}$), if $i \neq j$, B can generate a valid partial signature $\sigma_{i,\text{SEM}}$ since he gets access to P_i and $D_{i,\text{SEM}}$ by making use of the Sign algorithm. Otherwise, if $D_{j,\text{SEM}}$ has already been on the H_1^{list}, then B can generate $\sigma_{i,\text{SEM}}$ with it. Otherwise B selects $D_{j,\text{SEM}} \in G_1$ at random and then generates $\sigma_{i,\text{SEM}}$.

User-Sign Oracle: When A_I queries the oracle with (ID_i, M_i), if P_i has not been replaced, B can generate a valid signature σ_i since (P_i, D_i, x_i, $\sigma_{i,\text{SEM}}$) are available to him. Otherwise, B simulates a valid signature as follows. B selects $h,r \in \mathbf{Z}_q^*$ at random, computes $R_i=rP-hQ_i$ and $U_i=rY_i$, sets $H(M_i, R_i)=h$, updates H^{list} and finally the signature is $\sigma_i=(R_i,U_i)$, which is valid since it can be justified via the Verify algorithm.

After having queried these oracles polynomial times, A_I outputs a valid forged signature σ^* on message M^* under ID^* with non-negligible advantages. B aborts if $ID^* \neq ID_j$. In addition, if (ID^*, M^*) have been submitted to the User-Sign Oracle, B aborts too. Otherwise according to forking lemma [11], B can get two valid forgeries $\sigma^*=(R,U)$ and $\sigma'=(R,U')$, where $U=xR+(xhD_j)$ and $U'=xR+(x\,h'\,D_j)$. B computes

$$U-U'=(h-h')abxP \qquad (3)$$

and then

$$abxP=(U-U')(h-h')^{-1}. \qquad (4)$$

B successfully solves the NGBDHP by presenting ($abxP$, X_j), where $X_j=xP$. □

Theorem 2. The scheme is secure against Type II existential forgery under fully-adaptive chosen message attack in the random oracle model under CDH assumption.

Proof: Let A_{II} be a forger who breaks the improved scheme. B can make use of A_{II} to solve the CDHP instance (P,aP,bP). We omit the details and subtle differences for the sake of conciseness.

B generates parameters and the master key s, then sends them to A_{II}. B also maintains H_1^{list} and H^{list}, then chooses a ID_j as the challenged identity. B provides all the aforementioned oracles except SEM-Extract, User-Extract and SEM-Sign because they are redundant.

During querying the oracles, B sets $Q_j=bP$, $X_j=aP$ and $Y_j=s(aP)$. After A_{II} outputs his forgery, with forking lemma B can get two signatures $\sigma^*=(R,U)$ and $\sigma'=(R,U')$, where $U=aR+ahD_j$ and $U'=aR+a\,h'\,D_j$. B computes

$$U-U'=(h-h')sabP \qquad (5)$$

and then

$$abP=(U-U')s^{-1}(h-h')^{-1}. \qquad (6)$$

B successfully solves the CDHP.

7 Conclusions

The primary goal of introducing the SEM into CLS is to provide an instantaneous revocation so that the signing procedure should involve the SEM in presenting some information which contains his partial secret value. In this paper, we show that Yang et al.' MCLS scheme is insecure because of neglecting the fact that there is no authenticity on the public key. We then improve their scheme and prove it secure against existential forgery under fully-adaptive chosen message attack in the random oracle model.

References

[1] D. Boneh, X. Ding, G. Tsudik, et al., "A method for fast revocation of public key certificates and security capabilities," Proceedings of the 10th USENIX Security Symposium, Jun 25-30, 2001, Boson, MA, USA. Washington DC, USA: USENIX Association, 2001:297-308

[2] S. S. Al-Riyami, K. G. Paterson. "Certificateless public key cryptography," Proceedings of CRYPTO 2003 (LNCS 2894), Nov 30-Dec 4, 2003, Taipei, China. Berlin, Germany: Springer-Verlag, 2003: 452-473

[3] H. S. Ju, D. Y. Kim, D. H. Lee, et al., "Efficient revocation of security capability in certificateless public key cryptography," Proceedings of KES 2005 (LNAI 3682), Sep 14-16, 2005, Melbourne, Australia. Berlin, Germany: Springer-Verlag, 2005: 453-459

[4] X. Huang, W. Susilo, Y. Mu, et al., “On the security of certificateless signature schemes from Asiacrypto 2003, “ Proceedings of CANS 2005 (LNCS 3810), Dec 14-16, 2005, Xiamen, China. Berlin, Germany: Springer-Verlag, 2005: 13-25

[5] C. Yang, W. P. Ma, X. M. Wang, “Secure mediated certificateless signature scheme,” The Journal of China Universities of Posts and Telecommunications, 2007, 14（2）: 75-78

[6] W. S. Yap, S. S. M. Chow, S. H. Heng, et al., “Security mediated certificateless signatures,” Proceedings of ACNS 2007 (LNCS 4521), Berlin, Germany: Springer-Verlag, 2007: 459-477

[7] S. S. M. Chow, C. Boyd, J. M. G. Nieto. “Security-mediated certificateless cryptography,” Proceedings of PKC 2006 (LNCS 3958). Berlin, Germany: Springer-Verlag, 2006: 508-524

[8] M. H. Au, J. Chen, J. K. Liu, et al., “Malicious KGC attack in certificateless cryptography,” Proceedings of the 2nd ACM symposium on Information, computer and communications security-ASIACCS 2007, New York: ACM,2007,302-311

[9] D. Boneh, M. Franklin. “Identity-based encryption from the Weil pairing,” Proceedings of CRYPTO 2001 (LNCS 2139). Berlin, Germany: Springer-Verlag, 2001: 213-229

[10] J. K. Liu, M. H. Au, W. Susilo. “Self-generated-certificate public key cryptography and certificateless signature/encryption scheme in the standard model,” Proceedings of the 2nd ACM symposium on Information, computer and communications security-ASIACCS 2007, New York: ACM,2007,273-283

[11] D. Pointcheval, J. Stern. “Security proofs for signature schemes,” Proceedings of EUROCRYPT ’96 (LNCS 1070), May 12-16, 1996, Zaragoza, Spain. Berlin, Germany, Springer-Verlag 1996: 387-398

The Definition and Implementation of XML Document Update Language Based on Xquery

Hongcan Yan* [1,2] Minqiang Li[2] Baoxiang Liu[1] Dianchuan Jin[1] Wei Gao[3]

1 College of Sciences, HeBei Polytechnic University, Tangshan, Hebei Province, 063000 China

2 School of Management, TianJin University，Tianjin , 300072 China

3 Department of Computer Science, TangShan Normal University, Tangshan, 063000 China

Email: yanhongcan@heut.edu.cn

Abstract

In order to fully evolve XML into a universal data representation and sharing format, we must allow users to specify updates to XML document and must develop techniques to process them efficiently. Update capabilities are important for modifying XML documents. The World Wide Web Consortium has developed a standard for XML query language, called XQuery, according to it, begin with defining XML document update granularity, design a XML update language XUL in this paper, propose a set of constructs for expressing six updating operations in both an ordered and unordered XML data model, and map this constructs into the syntax of the XQuery language. In the end, we provide a logical frame for implementing XUL within an XML repository based on a relational database system.

Keywords：Primitive operation; Update granularity; XML document update; Xquery; Auto-Trigger

1 Introduction

The Characteristics of good data storage format, extensibility , being highly structured and easy network transmission determines XML (eXtensible Markeup Language) outstanding performance. The information exchange between enterprises through internet has become an important means to carry on e-commerce. With the wider use of e-commerce, XML has become the de facto standards for the exchange of information. Effective and efficient access to XML corresponding information has become increasingly important. To achieve this, we must have a query language to obtain the necessary information accurately and Update XML data sources. W3C (World Wide Web Consortium) put forward XQuery which has many advantages in the Query language combining with XML mainstream Query Language. But XQuery standards do not provide a lasting XML document on updating mechanism. No commercial system to support XML Query right through XML updating. But for many e-business information systems, an efficient updating operation is indispensable. This paper aims to design a XML update language, called XUL, express the syntax and semantics of updated statement on the basis of XQuery data model, and proposes logic implementation of XML metadata storage model and update mechanism.

This paper is structured as follows. We begin with a description of related work in Section 2. In Section 3, we define XML document update granularity -six basic elements and six sub-update operations, Section 4 describe XUL operational semantics by examples. Section 5 provides our implementations strategies. We present conclusions and future directions in Section 6.

* This work was supported by Natual Scientific Fund Project of Hebei Province (F2006000377), the Ph.D. Programs Foundation of Ministry of Education of China (No. 20020056047)..

2 Related Work and Analysis

Although a lot of storage and query optimization measures[1-3] have been provided, but it is unrealistic that lack updating function in a complete XML database management, and at present, the research in this area is very rare. Some database management systems based on object-oriented manager such as Lorel[4] has provided updating operation, but the operating sequence of programming language is very complicated because its involving multi-objects inserting and deleting. Reference [5] proposed that the XML - RL (XML Rule-based Query Language) language's updating definition consists of two parts: inquiry and modify, which supports multi-variant binding and orderly multi-documents modification, but it does not express metadata storage model and implementation mechanism. Reference [6] not only gives the grammar definition of modifying XML Document, but also analyzes the performance of various modifications strategy according to storage in relation to database. But it does define basic operation object or update granularity. An update language itself should be succinct and directly user-oriented besides the basic requirements of query and updating document contents [7-9]. Moreover, it must support the technique of XML Namespace and Xpath. This paper will emphasize the succinct language definition and open mechanism realization. XUL not only define XML document update granularity, but also express UPDATE clause in the form of FLWR(For-Let-Where-Return).

3 XQuery-Based Extended XML Update Language-XUL

The definition of Xquery-based extended XML Update Language considers both ordered XML document and unordered document, meanwhile supports IDREFS operation,

makes phrasing rule and XQuery FLWR expression compatible as possible as it can, simultaneously closes to SQL, make it become SQL of XML.

3.1 XQuery data model

XQuery data model view XML files as node tree with labels[1-2]. Figure 2 is a document tree corresponding XML document of Figure 1. Document Update granularity is the component of document contents, including the following six elements:

（1）Attribute: Attribute is expressed the ordered pair as (attribute name, attribute value). To distinguish between elements and attributes, add symbol "@" to the former, such as (@UnitPrice,26.50).

（2）Elements: Element is expressed as a list sequence serial of element name and sub-elements or attributes, the symbols ">" is used to describe the relationship between the elements and sub-elements, such as (LineItem->ProductID, @ UnitPrice).

（3）CDATA: refers to the type of CDATA text data.

（4）Processing-Instruction: expressed as (process instruction name, process instruction information), such as <?cocoon-process type =" xsp "?> , express this instruction as (cocoon-process, "xsp").

（5）IDREF, IDREFS : IDREF type attribute values invoke ID type attribute values, implied his relationship with his father, noted as (@IDREFS attribute name->@ID attribute value 1, @ID attribute value 2 ...).

（6）Comment: similar to the text of the string value.

```
<Customer>
  <CustomerID>ALFKI</CustomerID>
  <Order>
    <OrderID>10966</OrderID>
    <LineItem>
      <ProductID>37</ProductID>
      <UnitPrice>26.50</UnitPrice>
    </LineItem>
    <LineItem>
      <ProductID>56</ProductID>
      <UnitPrice>38.00</UnitPrice>
    </LineItem>
  </Order>
</Customer>
```

Figure 1 XML document fragment

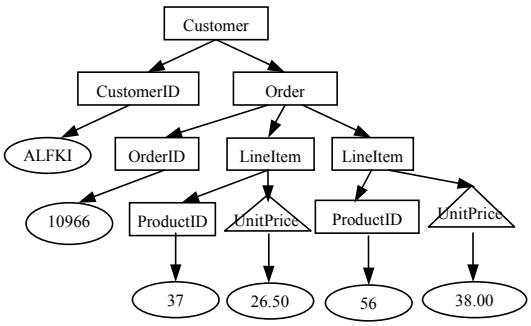


Figure 2 XML document tree

3.2 XUL operations

XUL provides six sub-update operations, an update is a sequence of primitive operations of the following types:

（1）Delete (node): Delete the target node. the node can be one of the above six elements. If the node is a composite element, elements or attributes will be removed automatically-first; If an ID cited, it will also be deleted; If IDREF, only to remove the reference IDREF from the IDREFS.

（2）Rename (node, name): Rename target node. Rename can not operate on the value of IDREFS in CDATA and IDREF, but can rename the whole IDREFS.

（3）InsertBefore (ref, content) : only applied to order documents. If the “ref” is an element of target object or CDATA (not an attribute because attributes are disordered), then “content” must also be an element or CDATA. Insert “content” directly before “ref”; If “ref” is IDREFS, then “content” must be an ID. Insert before the {ref}.position (i). Position (i) can be a user-defined or system-defined function.

（4）InsertAfter (ref, content) : similar to definition of InsertBefore (ref, content).

（5）Append (content) : only used for disorder documents. Append content to the end of target object. Content can be above one of the six types of content. But when inserting attribute, the “content” name can not be same as attribute name.

（6）Replace (node, content) : for order documents, equivalent to the first InsertBefore (node, content) operation and then Delete (node); for disorder documents, equivalent to the first implementing Append (content) and then Delete (content) operation.

4 XUL Operational Semantics and Examples

The basic set of operations presented in Section 2 express XUL logically, the next step is to map above six meta-language of sub-operation into the XQuery language syntax.

4.1 Basic form of XUL

We extend XQuery with FLWR syntax , use EBNF to describe as:

FOR $binding1 IN XPath-expr,…
LET $binding := XPath-expr,…
WHERE predicate1,…
UPDATE $binding1 {subOp {, subOp}*}

SubOp here refers to above six meta-language operations, described as follows:

DELETE $node |
RENAME $node TO name |
INSERT content BEFORE|AFTER $node |
APPEND content|
REPLACE $child WITH $content|
FOR $binding' IN XPath-subexpr, …
WHERE predicate1, . . . updateOp

In order to carry out multi-layers update operation in the compound XML structure, the UPDATE definition allows nesting.

4.2 XUL extensions in detail and examples

Example 1: delete operation. Figure 1, assuming that document fragment comes from the file ware.xml, the document root element is Customer. Now delete the LineItem elements whose UnitPrice is less than 30.

FOR $item IN
document(“ware.xml”)//LineItem
LET $price: = $item/UnitPrice
WHERE $price<30.00
UPDATE $item {

DELETE $item->ProductID
DELETE $item->@UnitPrice
DELETE $item }

Example 2: Insert operation. The statement below is to insert an attribute (@PO, "9572658") into the element Customer whose ID is "ALFKI".

FOR $c IN document("ware.xml")/;
Customer[CustomerID="ALFKI"]
UPDATE $c{
APPEND new_attribute(@PO,"9572658")}
(continuation row character: ;)

Example 3: Insert one LineItem element shown in Figure 3 and ensure ID ordered.

FOR $item IN
document("ware.xml")//LineItem
LET $p:=$item/ProductID
WHERE $p>=40.00
UPDATE $item{
new_element(ProductID,40)
new_attribute(@Unitprice,32.45)
INSERT (LineItem-> ProductID , @Unitprice) BEFORE {$item }.first()}

We can use INSERT AFTER operation, Screening all the LineItem elements whose ProductID<40, and insert after the last element, thus to complete an order insertion.

Example 4: Rename operation. The statement below is to rename all OderID elements to OderNo.

FOR $o IN document("ware.xml")/;
Customer/Order/OrderID
UPDATE $o{
RENAME $o TO OderNo}

Example 5: nested application. Sometimes we need to update multi-layers elements at a time, so we need to use XUL nesting. Such as:

FOR $c IN
document("ware.xml")/Customer
$order IN $c/Order
UPDATE $c {
new_ref(@PONo->t001)
APPEND new_attribute(@PO,@PONo)
FOR $item IN $order/LineItem[2]
UPDATE $item{
REPLACE 40.00 WITH $item->@UnitPrice}}

This statement completed inserting a reference attribute @PO which point to ID attribute @PONo, it's value is "t001", and meantime, modifying the UnitPrice attribute of second LineItem to 40.00.

5 Logic Realization Based on Trigger Mechanism

In order to ensure data integrity and consistency, update operation must be "safe" and "effective". This need the lock principles[7, 8] of update granularity layer, consistency check-up of data type and data path structure. DTD[W3C 1998] and XML Schema[W3C 2001b] have defined the XML file structure as effective criterion of model. This paper carried out the process of updating check up from two facets, the logic chart shown in Figure 4.

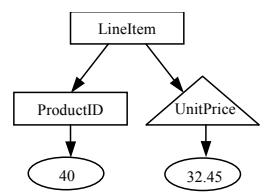


Figure 3 Iinsert elements LineItem

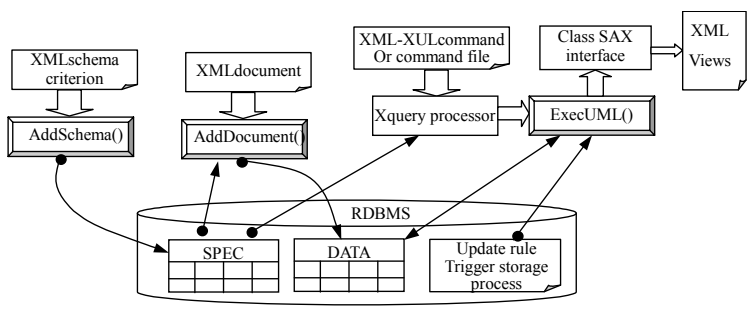


Figure 4 Logic structure of XML Document update realization

5.1 Parsing of XQuery processor

XQuery Processor mainly completes parsing sentence of XUL, and checks up the validation of data type and path structure matching, such as the number of occurrences of Elements, the type matching of predicate operator, legality of search path.

5.2 Auto-Trigger rules of update executing

According to the mapping rules of XML documents reflect to relational database[10-14], the update triggers confirm[15-17] update granularity. If the file is mapped to multi-table, the lock granularity is the table; if the file mapped to single table (which can be divided into multi-table horizontally according to the type of element value), the lock granularity is recorded. The author has discussed how to encode the nodes using Li-Moon coding[3] and storage data in single table in paper "Schema-based Constrained XML data indexing and storage Technique ".

The ECA (Event Condition Action) rules which are similar to relation rules of update are developed according to storage mapping. We defined these rules as triggers which can trig automatically when update statement executes. For example, when we try to delete a composite element with attributes or elements, the sub-element should be delete first; at the same case, when we try to delete the identifier reference (IDREF), this IDREF should be only removed from the list IDREFS. Especially for nested Update operation, some operators have strict order.

6 Conclusion and further research

We defined XUL language in user logic layer by means of extending XQuery, described the semantic of update operations in detail, and showed the syntax of multi-layers update of XML document by examples. This paper also addressed the logic frame of update realization based on relation database storage strategy[10-12]. The next step is to establish the prototype system of Figure 4 by adopting SQL Server and Java technology. Certainly, we need to define many assistant functions [18-21]by means of object technology which are integrated into module ExecUML(), and translate XUL language into SQL-3[14]. We don't consider the problem of index maintenance for highly efficient search. Of course, XML document update will bring index update. How to keep index order becomes our next research work.

References

[1] Floreseu D. KossmanD. Storing and Querying XML Data using a RDBMS. IEEE Data Engineering Bulletin, September 1999, 22（3）

[2] J. Shanmugasundaram, K. Relational Databases for Querying XML Documents：Limitations and Opportunities, In：Proc. of 25th Int1. Conf. on Very Large DataBases, Edinburgh, Scotland, UK 1999

[3] Quanzhong Li, Bongki Moon, Indexing and Querying XML Data for Regular Path Expression[C]. Proceedings of the 27th InternationalConference on Very Large Database, Roma, Italy, 2001: 361-370

[4] S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J. L.Winer. The Lorel query language for semistructured data. In Proceedings of International Journal on Digital Libraries, volume 1（1）, pages 68{88, April 1997

[5] Mengchi Liu, Li Lu and Guoren Wang, A Declarative XML-RL Update Language. ER 2003, LNCS 2813, pp. 506–519, 2003.Springer-Verlag Berlin Heidelberg 2003

[6] Tatarinov, I., Ives, Z.G., Halevy, A.Y., Weld, D.S.: Updating XML. In Proceedings of 2001 SIGMOD Conference, Santa Barbara, CA, USA (2001)

[7] Liu, M.: A Logical Foundation for XML. In Proceedings of the 14th International Conference on Advanced Information Systems Engineering (CAiSE '02), Toronto, Ontario, Lecture Notes in Computer Science, Vol. 2348, Springer 2002. (2002)568–583

[8] Andreas Laux,XML Update language, Working Draft 2000-09-09,http://www.infozone-group.org/lexusDocs/html/wd-lexus.html#N488c3376

[9] Lars Martin, XML Update Language Requirements, Working Draft - 2000-11-24, http://xmldb-org.sourceforge.net/ xupdate/xupdate-req.htm

[10] Tian Feng,DeWittDJ,et al. The Design and Performance Evaluation of Alternative XML Storage Strategies. SIGMOD Record,March 2002,31（1）

[11] Abiteboul S, Cluet S, et al. Querying and updating the file In:Proc. of 19th International Conference on Very Large Data Bases,Dublin, Ireland 1993

[12] Kanne C,Moerkotte G. Efficient storage of XMI, data In: Proc.of the 16th International Conference on Data Engineering, 28February-3 March, 2000, San Diego, California, USA IEEE Computer Society 2000

[13] Arenas M, Libkin L. A normal form for XML documents. ACM Trans. on Database Systems, 2004,29（1）:195-232

[14] Christophides V, Cluet S, Moerkotte G, Siméon J. On wrapping query languages and efficient XML integration. In: Chen W, Naughton J, Bernstein P, eds. Proc. of the 2000 ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM Press, 2000. 141-152

[15] ZHang Wen-chao, ZHang Jing, LI Jun-huai, New model of an active database based on triggers, Computer Applications, Vo1. 26 No. 10 Oct. 2006

[16] Li Jin-tang, Zhao yong, Yang wen-wei, Study of Multi-system Integration Based on Using XQuery to Simulate Database'S Stored Procedure, Journal of GuangDong Univercity of Technology ,V01.23 No.2 June 2006

[17] Zheng Hua, A Solution Of The Java Persistence Based On Stored Procedure

[18] Ludascher B, Papakonstantinou Y, Velikhov P. Navigation-Driven evaluation of virtual mediated views. In: Zaniolo C, Lockemann P, Scholl M, Grust T, eds. Advances in Database Technology-EDBT 2000, 7th Int'l Conf. on Extending Database Technology. Berlin, Heidelberg: Springer-Verlag, 2000. 150-165

[19] Fankhauser P, *et al.* XQuery 1.0 and XPath 2.0 formal semantics.2005. http://www.w3.org/TR/query-semantics/

[20] Chamberlin D, *et al*. XQuery 1.0: An XML query language. 2005. http://www.w3.org/TR/xquery/

[21] Buneman P, Fernandez M, Suciu D. UnQL: A query language and algebra for semistructured data based on structural recursion. The VLDB Journal, 2000,9（1）:76−110

Optimize FIR Digital Filter based on CSD Arithmetic

Xia Zhu[1] Yulin Zhang[2] Zhilei Chai[3] Wenbo Xu[3]

1 School of Information Technology, Southern Yangtze University, Wuxi, Jiangsu ,214122, China

Email: snail2024@163.com

2 School of Communications and Control Engineering, Southern Yangtze University, Wuxi ,Jiangsu, 214122, China

3 School of Information Technology, Southern Yangtze University, Wuxi ,Jiangsu, 214122, China

Abstract

The motive of digital filter research is that it is becoming a major method for digital signal processing. The performance of digital system is determined by multiplication. In this paper, a set of high-efficient multipliers for FIR, based on CSD coding, is presented, in which various optimized technique for digital filter is used. The experimental results show that CSD arithmetic improves the performance of multiplication, and reduces the use of resources. The results also show that CSD arithmetic only requires 26.7% of the LEs of 2C arithmetic and 40.7% of LEs of DA arithmetic in the optimal condition.

Keywords：digital signal processing ; FIR; CSD; DA

1 Introduction

Digital filtering is a part of digital signal processing. Digital signal processing mainly research these Figure s or symbols in order to indicate signal waveforms, in a sense, they are changed into a clear form, in order to estimate parameters of the signal, or weaken the excess signal and increase the useful signal component. All these signal filtering, transform, modulation, demodulation, balanced and enhanced compression, valuation, identification, production of processing, can all be included in the field of digital signal processing.

Now digital filter mainly uses DSP hardware or FPGA chip. Compared with the DSP, FPGA has the parallel arithmetic structure[1], which can significantly increase the data throughput filters. FIR filter based on FPGA is becoming a trend. Design of FIR filter has been continuously improved. Firstly, we began with the simple continually deconvolution, then Croisier put forward DA algorithm to replace convolution algorithm continually multiply, and then S.A.White[2] brought forward LUT-based DA algorithm and H.Yoo[3] gave LUT-less DA algorithm, respectively. What they do is to improve the speed of optimization and optimize the areas.

In this paper, we improve the FIR digital filter based on coding. This is because data are expressed in the form of common 2C (Two's complement), which is the traditional method. While there are some non-traditional expressions (such as CSD coding), and in some specific applications or in order to solve particular problems, we can improve efficiency. This is because: non-zero elements can usually be adopted to estimate multiplication efficiency. The CSD coding is different from the traditional binary coding ,that is, it has triple valued: 0, 1, -1,-1 usually is written as $\bar{1}$ .Application of CSD coding can reduce the number of non-zero elements, in addition, reduce the time of operations , improve the speed and reduce the areas . This improvement is based on the basic rules of the CSD which is advanced by RMM Oberman and LE Turner[8-10], and we apply it to the FIR digital filter design.

2 The basic principle of FIR filter

The FIR filter can be expressed as following mathematical expression. Eq.（1）describes an FIR filter of length N:

$$y(n)=\sum_{n=0}^{N-1}x(n)h(N-n) \qquad (1)$$

Where:

x and y are two vectors of size N that represent the input and transformed data , respectively.

h is the set of constant coefficients of the filter.

N is the number of taps of the FIR filter.

We can see that the FIR filter is made of an adder and a multiplier .Each sample x(n) requires N consecutive multiplication and (N-1) adder operating. The impulse response coefficient h is real numbers. The design of an FIR filter of length N can be accomplished by finding either the impulse response sequence h or N samples of its frequency response $H(e^{jw})$ .Also, to ensure a linear-phase design, the condition $h(n)=\pm h(N-n)$ must be satisfied, then, we can only use half of the coefficients to reduce the number of multipliers. The digital filter involves to the massive convolutions operation, which may take the massive resources. In order to avoid this kind of situation, we can make full use of FPGA with lookup table structure[4], which will make it easy.

3 Distributed arithmetic

Distributed Arithmetic (DA), which was as early as proposed in 1973 by Croisier[5], has carried on the promoted work by Peled and Liu. Until the emergence of LUT-based FPGA architecture, we can make full use of this algorithm.

Distributed Arithmetic appeared as a very efficient solution especially suited for LUT-based FPGA architectures. This technique is a multiplier-less architecture that is based on an efficient partition of the function in partial terms using 2's complement binary representation of data. The partial terms can be pre-computed and stored in LUTs. The flexibility of this algorithm on FPGAs permits everything from bit-serial implementations to pipelined or full-parallel versions of scheme, which can greatly improve the design performance.

In a bit-serial DA scheme .assuming that the input x to the filter is represented in M-bit 2's complement binary numbers with the sign bit to the left, we have:

$$x_i=\sum_{n=0}^{M-2}2^n x_i(n)-2^{M-1}x_i(M-1) \qquad (2)$$

Replacing this result in Eq.（1）, we obtain:

$$\begin{aligned} y_i&=\sum_{i=0}^{N-1}h_i x_i=\sum_{i=0}^{N-1}h_i[\sum_{n=0}^{M-2}2^n x_i(n)-2^{M-1}x_i(M-1)] \\ &=\sum_{i=0}^{N-1}h_i\sum_{n=0}^{M-2}2^n x_i(n)-\sum_{i=0}^{N-1}h_i 2^{M-1}x_i(M-1) \\ &=\sum_{n=0}^{M-2}2^n\sum_{i=0}^{N-1}h_i x_i(n)-2^{M-1}\sum_{i=0}^{N-1}h_i x_i(M-1) \end{aligned}$$

4 CSD algorithm

Multiplication is a basis operation of digital filter whose performance is determined by multiplications. It is not a difficult thing for any high languages. However, it is very hard for hardware design, because of its complicated Logical relations, resource usage and lower speed. To resolve this issue, several multipliers-less methods[6-7] can be proposed over years. This type of multiplier-less technique is the conversion-based approach ,in which the coefficient are transformed to other numeric representations whose hardware implement or manipulation is more efficient than the traditional binary representations. Example of such techniques are the Canonic Sign Digit (CSD) method which a save the usage of adders.

Regarding the coefficient multiplier design, a variable with a constant coefficient multiplication operation can be part of the shift product and then combined to achieve. Therefore, uses the DA algorithm and the CSD code may effectively reduce the product quantity, thus reduce the hardware complexity of multiplier unit.

（1）CSD coding

Each digit in the CSD word is allowed to take on a value of 1, 0,-1. A CSD coded coefficient will have most one-half as many non-zero digits as the equivalent binary coded coefficient. This translates directly into a reduction in the size of the multiplier implementation. A CSD coded bit-serial multiplier made from bit-serial adders, substrates and shifters are implemented for each

different multiplier coefficient in the FIR filter.

From the lowest effective bit of binary code, we replace all a greater than or equal to 2 in a series with $10\ldots0\bar{1}$. This classic CSD is a unique coding, whose another feature is that there is 1 zero between 0, $1,\bar{1}$ at least.

Eq.（1）can be done as the following changes based by CSD coding:

$$y_i = \sum_{i=0}^{N-1} h_i x_i = \sum_{i=0}^{N-1} x_i \sum_{j=0}^{M-1} (h_i(j)) \qquad (3)$$

$$= \sum_{i=0}^{N-1} x_i (2^{M-1} h_i(M-1) + 2^{M-2} h_i(M-2) + \ldots\ldots + 2^1 h_i(1) + 2^0 h_i(0))$$

$$= \sum_{i=0}^{N-1} (2^{M-1} x_i h_i(M-1) + 2^{M-2} x_i h_i(M-2) + \ldots\ldots + 2^1 x_i h_i(1) + 2^0 x_i h_i(0)) \qquad (4)$$

From Eq.（4）we can see, with the increasing of 1 sequence, the addition will be followed increasing frequency. The application of CSD can reduce the number of non-zero elements in the multiplication operation in addition to reducing the amount of computation, help improve the speed and reduce the computing resources of the occupied. For example, X is the input, Y is the output, and H is coefficients. Let us assume $H=31_{10}$,then, $Y=H*X=X*31=X*(011111)_2=X*(2^4+*2^3+*2^2+*2^1+*2^0)=X*(10000\bar{1})_{CSD}=X*(2^5-2^0)$。

Figure 1 gives a multiplication operator realization of the combination of sequence structure, Figure 2 shows optimization of the above sequence.

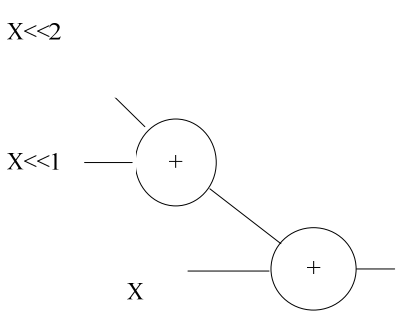


Figure 1 sequence structure before optimization

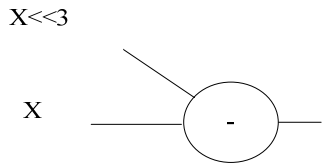


Figure 2 sequence structure after optimization

Given the complexity of the hardware constraints, classic CSD coding it is not always capable of producing the best binary coding. Therefore, there is the need to amend the classical coding method and turn it into the best coding method.

There is the best coding method:

1．From the lowest effective bit of binary code, replace all series whose value is greater than or equal to 2 with $10\ldots0\bar{1}$

2．let $110\bar{1}$ take palace of 1011;

3．From the highest effective bit of binary code, let O11 take place of $10\bar{1}$.

Best CSD coding can make the number of non-zero elements to minimum; it can reduce the use of the multiplier resources. In the following table, some comparisons are given. We can see from the table, using multiplication CSD code optimization design, not only made the filter to reduce the use of resources and reduce the computation time.

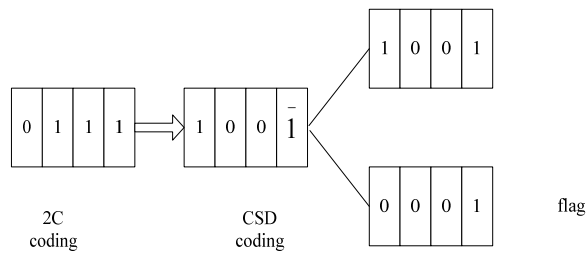


Figure 3 describe the conversion between 2C and CSD coding

5 Design process

There are three steps:

（1）Using MATLAB FDATLOOL tools to get coefficient, and quantize;

Filter order: 8 coefficient bit: 8 bit

$h_0= h_7=1$ $h_1= h_6=15$ $h_2= h_5=19$ $h_3= h_4=128$

（2）Read quantization coefficient, generate CSD coding and the corresponding symbol flag [M];

Table 1 the generated CSD coding and Flag [M]

Coefficient	CSD	CSD Hardware Description	Flag[M]
$h_0=h_7=1=2^0$	00000001	00000001	
$h_1=h_6=15$	$0001000\bar{1}$	00010001	00000001
$h_2=h_5=1$	$0001010\bar{1}$	00010101	00000001
$h_3=h_4=12$	10000000	10000000	

（3）Using Eq.（4）and flag [M], computation results.

Y(n)=A*X(n)+B*X(n-1)+C*X(n-2)+D*X(n-3)+

A*X (n-7) +B*X (n-6) +C*X (n-5) +D*X (n-4)

Where ： $A=2^0=1$； $B=2^4-2^0=15$ (the last bit of h_1 and h_6 Flag [] is 1);

$C=2^4+2^2-2^0=19$ (the last bit of h_2 and h_5 Flag [] is 1); $D=2^7=128$.

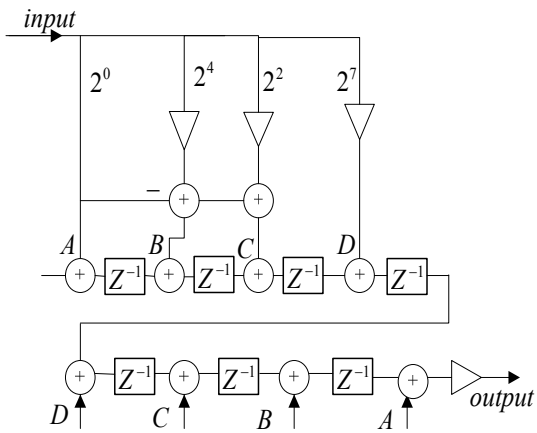


Figure 4 Block diagram of FIR filter based CSD coding

6 Results and analysis

S.A.White implemented the LUT-based DA version : the scheme gets a partial term beginning with the LSB of the input and the shifts this to the right to add it to the next partial result. H.Yoo implemented the LUT-less DA version to compare area performance with typical LUT-based DA schemes. That is: the new scheme first begins with the MSB of the input and shifts each partial result to the left, avoiding the logic necessary to manipulate the LSB. H.Yoo used an Altera Stratix EP1S80F1508C6 FPGA device and tested the implementation with different FIR filter length (N) .The results for 4, 16, and 64-tap filters are shown in Table2.

Table 2 Area requirements of LUT-based and LUT-less DA schemes presented in [2] for different filter length implementations on an Altera FPGA device

N	LUT-based DA LE	LUT-less DA LE
4	272	210
16	551	367
64	1639	887

In the present work, the results obtained using the Xilinx compilation tool are shown in Table 3. The device selected for all the implementations is Virtex-E XCV400E FPGA of Xilinx family.

Table 3 Area requirements of LUT-based and LUT-less DA schemes presented in [2] for different filter length implementations on a Xilinx FPGA device

N	LE(DA)	CPU(DA)
4	233	13.45/14.47
8	384	14.45/15.47
16	489	19.28/23.70
32	695	21.17/22.39

In a bit-serial DA scheme, assuming the input x_i to the filter is represented in 2's complement binary numbers. From the other hand, we can also do with h_i with the same method. In the normal condition, h_i is represented in 2's complement binary numbers. Now we apply CSD coding to function（1）, and we can get the expression in function（4）.Table 4 shows that the requirements of the proposed CSD architecture. In terms of memory requirements, the results also show that CSD arithmetic only requires40.7%~75% of LEs of DA arithmetic in the optimal condition. In terms of operation time, the savings are about 63.8%~82.4% in each implementation.

Table 4 Area requirements of DA and CSD coding for different filter length implementations on a Xilinx FPGA device

N	LE(DA)	CPU(DA)	LE(CSD)	CPU(CSD)
4	233	13.45/14.4	95	8.59/9.63
8	384	14.45/15.47	187	9.67/10.72
16	489	19.28/23.70	371	15.84/16.86
32	695	21.17/22.39	538	17.28/18.53

Because of the inherent characteristics of CSD coding, the efficiency of the function（4）is determined by the length of the 1 series. As we can see, Table（5）shows the area performance in the optimal condition (the length of 1 series of h_i is biggest) and the worst condition (h_i is alternating with 011).

Table 5 Area requirements of CSD coding and 2C coding for different filter length implementations on a Xilinx FPGA device

	N=4 max	N=4 min	N=8 max	N=8 min	N=16 max	N=16 min
LE (2C)	332	212	684	444	1388	908
CPU (2C)	11.81 /12.84	9.72 /10.7	17.97 /20.5	12.34 /13.3	41.42 /42.5	19.52 /20.5
Elapesd (2C)	12/13	9/10	18/20	12/13	41/42	20/20
LE (CSD)	95	207	187	423	371	855
CPU (CSD)	8.59 /9.63	10.38 /11.4	9.67 /10.7	13.81 /14.9	15.84 /16.8	20.89 /21.9
Elapsed (CSD)	9/9	10/11	9/10	13/14	16/17	21/22

As we can see, in terms of memory requirements, the results also show that CSD arithmetic requires 50% of LEs of 2C arithmetic in normal condition, especially only requires26.7% in the optimal condition and requires 95% in the worst condition.

7 Conclusion

We have successfully implemented high-efficient FIR filters, using CSD arithmetic, which has been further improved in area performance with the efficient usage. In the optimal condition, the occupier of resources and the effect of speed are particularly evident. Compared table（4）and table（5）, the performance is slightly superior in the CSD approach to DA arithmetic. But in the worst case, DA algorithm is superior to CSD. However, in the normal condition, in terms of the use of resources alone, (CSD coding) < (DA) << (2C coding).

References

[1] Parhi KK. “A systematic approach for design of digit-serial signal processing architectures”. IEEE J SolidState Circ 1992; 27: 29-43

[2] S.A.White,”Application of distributed arithmetic to digit signals processing: A tutorial review”, IEEE ASSP Magazine, vol.6.pp.4-19, July 1989

[3] H.Yoo and D.Anderdon,” Hardware-Efficient Distributed Arithmetic Architecture for High-Order Digit Filters”, in proc.IEEE International Conference on Acoustics ,speech, and Signal Processing ,2005,vol.5,pp.125-128

[4] M.A.Soderstrand, L.G.Johnson, H.Arichanthiran, M.Hoque, and.Elangovan, ”Reducing Hardware Requirement in FIR Filter Design”, in proc.IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000, vol.6, pp.3275-3278

[5] A.Croisier,D.J.Esteban,M.E.Levilion, and V.Rizo, ”Digit Filter for PCM Encoded Signals”, U.S.Patent No. 3,777, 130, issued April, 1973

[6] Zhao, Q; Tadokoro, Y. A. “Simple Design of FIR Filters with Powers-of-Two Coefficients”, IEEE Transactions on Circuits and Systems. 35,5 (May, 1988)

[7] Meyer-Baese, U.“Digital signal processing with field programmable gate arrays” (Springer-Verlag, Berlin, Heidelberg, 2001)

[8] R.M.M. Oberman. “Digital Circuits for Binary Arithmetic”. Macmillan Press Ltd., 1979

[9] L.E. Turner, P.J.W. Graumann and S.G. Gibb, “Bit-serial FIR Filters with CSD coefficients for FPGAs”, 1994

[10] Hartley R. Subexpression sharing in filters using canonic signed digit multipliers[J]. IEEE Transactions on Circuits and Systems II, 1996 43（10）：677—688

Development of Automatic Control Principles Virtual Experimental Platform Based on Matlab

Jianjun Zhu[1] Xingquan Gao[2]

1 Department of Automation, Jilin Institute of Chemical Technology, Jilin, 132022, China

Email:1 zjj099@163.com ; 2 gxq2008@163.com

Abstract

The virtual experimental technology is generated with the development of computer technology, multimedia and network technology.Especially with the development of distance education, remote experimental teaching is gradually taken by the people.

This paper researches virtual experimental development based on Matlab Web Server, presents the technology based on Matlab and Web.The paper introduces structure and development of the platform.At last the process of developing the virtual lab is introducted by one typical virtual experiments. The platform has lower software and hardware requirement of client, has advantages of easy expansion, maintenance and upgrading, also it has beautiful interface and simple operation.

Every experiment has objection,theroy and steps,you can input parmaeters in parameter boxes and click submit button to get the results.The platform breaks the restrictions of the time and space,students can do experiments at any time and any where,alter paremeters repeatedly,it can deepen the understanding of knowledge and enhance the sets and the ability of innovate.

Keywords：virtual experiment; distance education; Web; Matlab web Server; virtual lab

1 Introduction

Traditional simulation approach is generally based on single-computer, which is poor in interoperability and portability. As far as simulation of the engineering system is concerned, the cost is quite high and the learning period is long. Along with the development of net technology, the simulation approach based on Web system came into being. With this approach, the application program of server is separated from the setting of client, so that the simulation is realized as long as the IE is installed at the client. Compared to the traditional approach, it has advantages of low-cost, good maintainability and high integration.

As the most excellent numerical calculation software in the world, Matlab has powerful calculation function, ample and convenient graphs and full function automation control software toolkit. And therefore it is applied in simulation and calculation by technical staff. The present paper adopts the Matlab Web Server of Matlab 6.5 to develop a simulation experimental platform of control system based on Web by combining Matlab's calculation and Web remote access [1].

2 Principle of actualization

MATLAB Web Server applications are a combination of M-files, Hypertext Markup Language (HTML), and graphics. Knowledge of MATLAB programming and basic HTML are the only requirements.

The application development process requires a small number of simple steps:

1) Create the HTML documents for collection of the input data from users and display of output. You can code the input documents using a text editor to input HTML directly, or you can use one of the commercially

available HTML authoring systems, such as Front Page from Microsoft, PageMill from Adobe, or HoTMetaL from SoftQuad.

2) List the application name and associated configuration data in the configuration file matweb.conf.

3) Write a MATLAB M-file that:

a. Receives the data entered in the HTML input form.

b. Analyzes the data and generates any requested graphics.

c. Places the output data into a MATLAB structure.

d.Calls htmlrep to place the output data into an HTML output document template. The maximum amount of HTML data you can receive from MATLAB is 256 KB.

Figure 1 shows how Matlab operates over the Web [2].

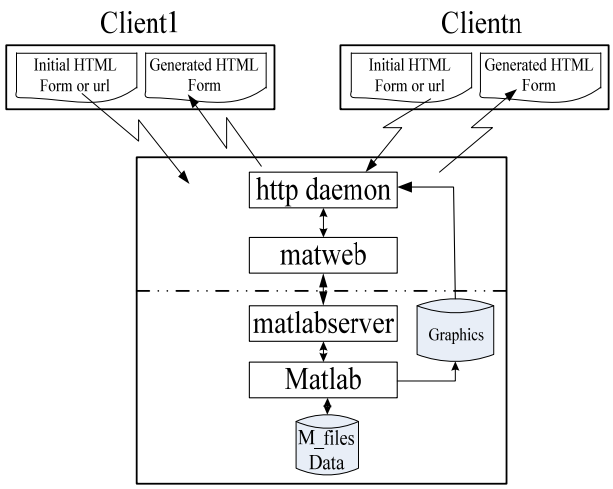


Figure 1 Matlab on the Web

3 Realization of simulation platform

3.1 The configuration of Web server

3.1.1 Web server Configuration

You need to install Web Server software(http or similar)on the system where Matlab is running or on a machine that has network access to the machine where Matlab is running. After installing Matlab Web Server, Matlab Web Server needs some necessary configuration to make it fuse with www service of the system. Apache 2.0 is applied to design the server computer into Web server, and then the specific access should be opened for CGI so as to make Apache has an access to CGI. Locate <Directory "C:\ Apache \cgi-bin">…< /Directory> in the document named http.conf of Apache, make "Options None" into "Options ExecCGI", and then restart Apache Web Server to put it into effect. In this way CGI program can be operated under the virtual catalogue "\cgi-bin"[3,4].

3.1.2 Configuration of matweb.conf

To connect with malabserver, matweb requires information stored in the configuration file matweb.conf. Take the control system of tank liquid-level PID for instance:

[webroot]

mlserver=127.0.0.1

mldir=C:/Apache2/htdocs/kongzhi

[webroot] is the name of M document employed by Matlab while simulation of tank liquid-level PID; mlserver=127.0.0.1 means to conFigure the name of the server of IP(the illustration is used to the single test). mldir=C:/Apache2/htdocs/kongzhi means to set the path of Matlab program, meanwhile it also reads and writes the document catalogue, i.e. the storage catalogue of webroot.m, once appointed the system will automatically add the path of the catalogue to the path of the Matlab system [5,6].

3.2 Creating Output Documents

At the client, users hook up with system through webpage explorer, conFigure the system parameter and submit them to the server. The main structure of the system contains the connection of all the modules while in the framework of specific module choices such as parameters of the configuration, watching signals and simulation result should be provided.

This file looks like:

...<!—this line establishes communication with

```
Matlab by matweb.exe-->
    <FORM action="/cgi-bin/matweb.exe" method="post" target="outputwindow">
    <INPUT type="hidden" value="kongzhi" name="mlmfile">
    <p><span class="style7">Configuration Parameters </span></p>
    <P align=center><SPAN class=style7>end: Tf = <INPUT size=10 value=20 name=Tf>Set point:Ys= <INPUT size=10 value=2 name=Ys> </SPAN></P><P align=center>
    <SPAN class=style7>parameters: Kp = <INPUT size=10 value=5 name=Kp> Ti = <INPUT size=10 value=0.5 name=Ti> </SPAN></P>...
```

After making the HTML document submitted to the server, it is stored under the virtual catalogue C:\Apache2\htdocs\kongzhi. By keying in the host name or IP address as well as the title of the page, the page could be browsed individually.

3.3 Matweb M-File

Matlabserver uses the value of mlmfile obtained from the matweb M-file, matweb.m, to run the Matlab application[7,8].Server and finally m document deals with the simulation calculation and drawing graph. Take the control simulation of tank liquid-level PID as an example, its file M is named kongzhi.m, which corresponds to the parameter value of the input form variable mlmfile.

The 'kongzhi' web application M-file is named kongzhi.m. This file helps to transfer the input from the kongzhi.html to the MATLAB function kongzhi.m. The kongzhi.m is then doing the necessary calculations. On the final stage the kongzhi.m transfers data computed by the kongzhi.m into the HTML output file.This file looks like:

```
function out = kongzhi(in,outfile)

Kp = str2double(in.Kp);
Ti = str2double(in.Ti);
r = str2double(in.Ys);
Tf = str2double(in.Tf);
A=2;
h0=1.5;
Q1=0.1;
Q2=0.15;
Q3=0.25;
u1=0.5;
C1=Q1/u1;
C3=Q3/sqrt(h0);
x=h0;
Ts=0.2;
 h=0.01;
 Fig = Figure ('visible','off');
t=0:h:Tf;
M=t(length(t))/Ts;
N=Ts/h;
u=0;
e=0;
n=1;

for k=2:M+1
     for k0=1:N
          y1(n)=x;
          k1=kz(t(k-1),x,u1,C1,Q2,C3,A);

k2=kz(t(k-1)+0.5*h,x+0.5*h*k1,u1,C1,Q2,C3,A);

k3=kz(t(k-1)+0.5*h,x+0.5*h*k2,u1,C1,Q2,C3,A);
          k4=kz(t(k-1)+h,x+h*k3,u1,C1,Q2,C3,A);
          x=x+h/6*(k1+2*k2+2*k3+k4);
          n=n+1;
      end
    y(k-1)=y1(n-1);
    e(k)=r-y(k-1);
    u=u+Kp*((1+Ts/Ti)*e(k)-e(k-1));
    u1=u+0.5;
end

T=0:Ts:(M-1)*Ts;
hold on;
plot(T,y,'m')
```

```
plot(T,r,'r')
hold off;

set(gcf,    'PaperPosition', [.5 .5 16 6]);
cd(in.mldir);
filename ='kongzhi';
icondir = '../icons/';
  icondirhttpdname = '../matlab/icons/';
  jpegname = sprintf([filename '%s.jpeg'], in.mlid);
fulljpegname = [icondir jpegname];
drawnow;

  wsprintjpeg(Fig, fulljpegname);
close(Fig);
  templatefile = [filename '2.html'];
s.GraphFileName = [icondirhttpdname jpegname];
if nargin == 2
    out = htmlrep(s, templatefile,outfile);
else
     out = htmlrep(s, templatefile);
end

  cleanname = [filename 'ml*.jpeg'];
wscleanup(cleanname, 1, icondir);
```

The explanation of main functions is:

（1）Function of sprintf([filename '%s.jpeg'], in.mlid) creates a name for a jpeg file. If mild has the value ml00005 and filename has the value kongzhi, for example, the jpeg file will be named kongzhiml00005. jpeg.

（2）Function of htmlrep is to obtain variable data from out and replace the corresponding variables as the form of &picName& on the final page. In this way the title of the simulation imagery file is transmitted. The variable that represents the graphic output is found in the line <img border=0 src="$GraphFileName$"> in kongzhi2.html.

Figure 2 shows the final outputs document both the input and output frames of the control simulation of tank liquid-level PID.

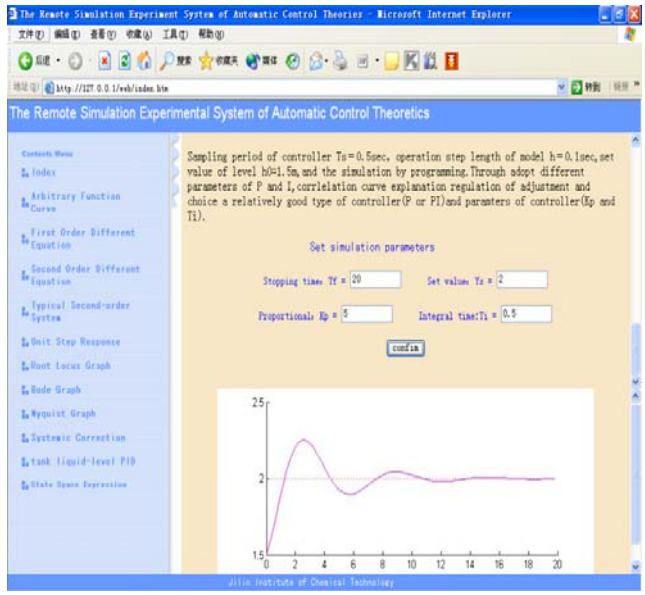


Figure 2 simulation of tank liquid-level PID

4 Conclusions

The subject constructs Automatic Control Principles virtual lab using Dreamweaver powerful Web function, Flash powerful graphic processing function and MATLAB powerful drawing, computing and web functions. The construcetion of this lab includes web servere on figuration, virtual experimental circuit design, input documents creation, M-Files creation and output doeuments creation. The paper introduces structure and development of the lab. The lab has twelve experiments, covering time and frequency domain analysis of linear system, root locus and series compensation of linear system, an alysis of discrete and nonlinear system. The lab has lower software and hardware requirement of client, has advantages of easy expansion, maintenanee and upgrading, also it has beautiful interface and simple operation. Every experiment has objection, theory and steps,also circuit diagram and virtual circuit, you can input parameters in parameter boxes and click submit button to get the results. The lab breaks the restrictions of time and space, students can do experiment sat any time and anywhere, alter parameters repeatedly, it can deepen the understanding of knowledge and enhance the sets and the ability of innovate.

References

[1] The Math Works.Matlab Web Server[M].USA Massachusetts: The Math Works Inc,1999

[2] Douglas E.Comer. Computer Networks and Internets, Second Edition [M]. American: Prentice Hall Companies, 1999

[3] Veith T L.World Wide Web_based Simulation [J].Int.J. Engng Ed, 1988, 14（5）:316-321

[4] Matlab Web Server[DB/OL].The Mathworks, Inc, Version 1.2,2002

[5] Shuang H Yang,James L Ahy. Development of a distributed simulator for control experiments through the Internet [J]. Future Generation Computer Systems，2002，18:595-611

[6] Gu Hui,Li Yan shan，Chen Qi，Zhang Yun tao.One distributed virtual experiment platform model and the cooperation mechanism[J].IEEE Computer Supported Cooperative Work in Design，2004，14-17

[7] Hui qin Jia, Jun hua liu. Developing remote virtual instrument laboratory(RVIL) based on browser/server pattern.IEEE Info-tech and Info-net，2001:267-272

[8] Zhao Ai ping. Matlab Web Server and it's a pplication in remote collaborative design of magnetic bearing systems [J]. Chinese Journal of Meehanieal Engineeeing. 2001,14（2）: 179 - 183

[9] Burdea G,Coiffet P.Virtual reality technology. NewYork: JohnWile&Sons,Inc,1994,5-10

[10] ZENG Jian-iiang,CHEN Wen-liang,DING Qiu-ling.A Web-based CAD system[J]. Journal of Materials Proeessing Technology.2003,139:229-232

General Register Design

Kui Yi Yuehua Ding Xin Du

Department of Computer Science and Information Engineer, WuHan Polytechnic University, Wuhan, Hubei, 430023, China

Email: ykll1903@126.com

Abstract

This paper discusses a kind of 32-bit high performance microprocessor instruction set structure, and introduces design and implementation of general register. We do research on logic of register file and circuit optimization design. Furthermore, we design port share and controller-circuit with stagger read-write circuit. Emphase of our research is implementation of general register file, which includes two pipeline processes: launch and acknowledgement. We design one clock signal、one write-enable signal、one write-target register port and one write data port in launch stage; We design two read-source register ports and two output data port in acknowledgement stage. Finally this paper also verifies function of register file.

Keywords: MIPS; RISC; VHDL; FPGA; CPU; General Register

1 Introduction

With rapid development of micro-electronic technology, people's life has been changed constantly. FPGA/CPLD has been replaced by ASIC market gradually. FPGA/CPLD has so many advantages, such as high performance、use repeatability、small amount production、short development cycle, more and more IC designers are willing to use FPGA/CPLD.

CPU based on RISC instruction system pipeline is designed by field FPGA/CPLD rapid prototyping to construct real computer system. We completed these function design of CPU: micro-system design、logic design and integration、 FPGA function verification、 Linux kernel replant etc. This project realizes pipeline structure design(5-stage pipeline: single issue、dynamic scheduling, rename, branch prediction, inference execution)、 memory system design(cache、memory access management、data transform、virtual memory)、arithmetic component design: Write micro-architecture C simulator which is to sign exactness; Run tiny test program and benchmark test program; Check correctness of design and micro-debug structure configuration parameter[1]. In our project, we used VHDL hardware language to write module Executable Register-Transfer-Level process model. Furthermore, we completed logic synthesis and achieved RTL processor model which can execute binary program. We used EDA tools to produce test programs for running these test programs on RTL model, and then, we compared state of RTL model to state of referenced processor model. The state of two model architecture must match in every cycle. Target of function verification is that RTL model describes processor in faith.

This paper discusses design and implementation of general register file. This register has two parts: data store and data output. Data Store part has one clock sign、one write-enabled signal、one write destination register port and one write data port. Data output part has two read source register ports and two output data ports. Figure 1 shows general register file structure.

How to implement MIPS instruction architecture is emphasis of our project. Based on existing computer instruction system MIPS, the instruction applied to this system is to be acknowledged in depth and completeness and correctness of the instruction system

is considered. Difficulty of the project is how to realize the rationality of operation code field and address code field assignment, and how to make computer recognize instruction from user correctly. The main function of general register design is how to assure the correctness of data store and output. Because register is the most important unit of CPU and register read or write data constantly, the emphases in our research includes that: how to assure data write to specified register, how to ensure correctness of read data, and register is how to connect to other unit etc.

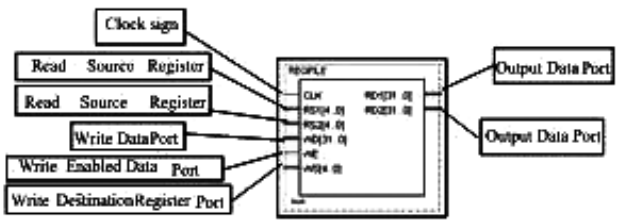


Figure 1 General Register File Chip

2 Microprocessor architecture

2.1 CPU Structure

Function of CPU includes instruction fetch、instruction decoder、instruction execution、data write-back. In the project, the general register file has 32 32-bits registers[2]. Read operand from random-access register group and write result to register file(Read operand from register file and write result to register file). General register file permits instruction access register through specified register address and leaves the register value unchanged while reading.

CPU architecture shows as Figure 2. Instruction store sub-system and data store sub-system is mainly composed of memory. Program counter and instruction queue compose of instruction address generation unit and fetch instruction unit. Instruction decoder unit is composed of instruction decoder. Instruction is composed of arithmetic unit and memory buffer. Controller constitutes center controller Unit. When CPU powers on, control unit sends instruction to instruction address generate unit, And then, the unit generates instruction address and commands instruction store system to send instruction. The instruction was put to fetch instruction unit. The instruction is decoded by instruction decode unit. After decoder, instruction execution unit operates on operand depending on decoder result under central controller unit's control. Result is put into write-back unit, and write-back unit puts final result into register file based on controller signal from controller unit and operation selected by instruction. Finally, judge whether put data into memory, and instruction fetch unit fetches next instruction to run at the same time.

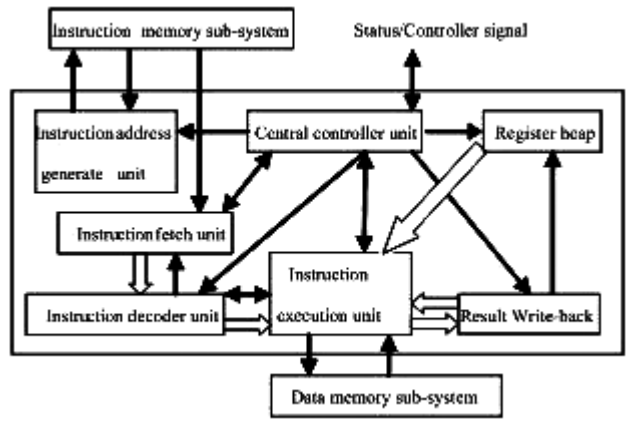


Figure 2 CPU Architecture

2.2 Instruction Set Architecture

Instruction format is usually composed of operation code field and address code field. Different instruction is represented by different operation code field. Each code expresses one kind of instruction, but address code field expresses location of operand in computer[3].

2.2.1 Three Type Instruction

R- Format Instruction: R-Format instruction is the most common instruction. Typical R-Format instruction format shows as Table 1:

Table 1 R-Format Instruction

Instruction Code *OP*	Register RS	Register RT	Register RD	Shift Offset Shamt	Function func
D31~D26	D25~D21	D20~D16	D15~D11	D10~D6	D5~D0
6 bits	5 bits	5 bits	5 bits	5 bits	6 bits

R-Format instruction has 6-bit operation code(*opcode*) and 15-bit operand. There are three operands which are all put in three specified register. Each addressing breadth of register are 5-bit in R-Format, which means there are 32 registers in register file. The third 5-bit register *RD* is destination register, and the first 5-bit register *RS* and the second 5-bit register *RT* are source register. The last 6-bit *func* in R-format instruction is used as function field, which denotes what kind of operation is to be executed, such as add、subtraction. *Shamt* expresses shift offset.

I-Format instruction: it is similar to R-format instruction. The difference between them is that the third destination register *RD*、*Shamt* field and last function field *func* of R-format is replaced by immediate of I-Format instruction. Table 2 illustrates typical I-Format instruction. Each I-Format operation code has only one corresponding instruction, because it has no function field *func* as R-Format instruction.

Table 2 I-Format Instruction

Instruction Code *OP*	Register RS	Register RT	Immediate Value
D31~D26	D25~D21	D20~D16	D15~D0 (Immed RAdr)
6 bits	5 bits	5 bits	16 bits

J-Format instruction: it includes register *RS*、register *RT*、 instruction code and corresponding 16-bit immediate. J-Format is only used in jump instruction modality. Because instruction code has 6 bits, the other 26 bits is used as jump specified value. Eight bits compose one byte in program memory unit, and 32 bits are 4 bytes. Program memory is increased by degrees with 4. Program counter of immediate addressing 26 bits must multiply 4 to become 28 bit. *A1*、*A0* which are lowest bits in J-Format are always 0. The immediate addressing 26 bits express *A27~A2* in PC, and moreover, *A31~A28* which are the highest 4 bits maintain original value. Typical J-Format instruction shows as Table 3.

Table 3 J-Format Instruction

Instruction Code *OP*	26-bit immediate address
6 bits	A31, A30, A29, A28 (A27~A2, [A1, A0 =00])

2.2.2 Instruction Set

According as foregoing statement, instruction system adopted in our design shows as Table. 4:

Table 4 Instruction Set

Operation	Instruction	Format	Function Statement	Shamt	func
111111	HALT	R	Halt	11111	111111
000000	NOP	R	No Operation	00000	000000
000000	ADD R1, R2, R3	R	Add Word (R3 = R1 + R2)	00000	100000
000000	ADDu R1, R2, R3	R	Add Unsigned Word (R3 = R1 + R2)	00000	100001
000000	SUB R1, R2, R3	R	Subtract Word (R3 = R1 – R2)	00000	100010
000000	SUBu R1, R2, R3	R	Subtract Unsigned Word (R3 = R1 – R2)	00000	100011
000000	SLT R1, R2, R3	R	Set on Less Than (R3=0 if R1<R2 else R3=1)	00000	101010
000000	SLTu R1, R2, R3	R	Set on Less Than Unsigned	00000	101011
000000	AND R1, R2, R3	R	And(R3 = R1 and R2)	00000	100100
000000	OR R1, R2, R3	R	Or(R3 = R1 or R2)	00000	100101
000000	XOR R1, R2, R3	R	Exclusive-Or (R3 = R1 xor R2)	00000	100110
000000	NOR R1, R2, R3	R	Not Or (R3 = R1 nor R2)	00000	100111
000000	JR Rs	R	Jump Register	00000	001000
000000	SLL R1, R2, R3	R	Shift Word Left Logical (R3 = R1 shifted by R2)	shamt	000000
000000	SRL R1, R2, R3	R	Shift Word Right Logical (R3 = R1 shifted by R2)	shamt	000010
000000	SRA R1, R2, R3	R	Shift Word Right Arithmetic(R3 = R1 shifted by R2)	shamt	000011
000010	Jmp Adr	J	A31,A30,A29,A28(PC) Adr(A27~A2) A1,A0=00		
000100	BEQ R1,R2,data16	I	Branch on Equal; If R1 = R2 GOTO data16		
000101	BNE R1,R2,data16	I	Branch on Not Equal; If R1 ≠ R2 GOTO data16		
001001	ADDIu R1, R2, data16	I	Add Immediate Unsigned Word (R2 = R1 + data16)		
001000	ADDI R1, R2, data16	I	Add Immediate Word : (R2 = R1 + data16)		
010011	LW R1, R2 data16	I	Load Word		
011011	SW R1, R2, data16	I	Store word		

Based on foregoing instruction set, we write one simple assembly language program showing as below:

```
LW      R0, R1, data16   ;  R1 <= RO + data16
LW      R0, R2, data16   ;  R2 <= R0 + data16
ADDu    R1, R1, R3       ;  R3 <= R1 + R1
ADDu    R3, R1, R4       ;  R4 <= R3 + R1
SUBu    R2, R1, R5       ;  R5 <= R2 - R1
SRA     R1, R4, R3       ; R3 <= R1 shifted by R4
JR      R0               ; jump to address R0 pointed
```

Instruction fetch unit、register file unit、branch jump and pre-fetch operand logic unit、arithmetic unit and interface between CPU and main memory must be connected to compose whole CPU model. After program counter fetch the first instruction from location 0, every instruction adds 4 bytes to get the next instruction physical address and execute program orderly. The program examples shows: after load operand into register group, do arithmetic operation on operands, such as add and subtraction, and do shift operation on operands. The last instruction is jump instruction which is to jump to location 0 of memory. And then, continues next execution. The assembly language program shows one infinite loop.

3 General register file design

3.1 Micro-processor Design

1) R-Format instruction: CPU fetches instruction from memory according to program counter specifying address and load instruction into instruction queue of instruction register. When the instruction is decoded, the instruction is broken down to two parts: operation code and operand. Operation signal is sent to operation port of ALU, and select arithmetic operation type to do operation. The operand is sent to register group in CPU. R-Format instruction fetches operand from *RS* register and *RT* register. The operands are put to ALU. After arithmetic is over, write back result to the *RD* register. R-Format instruction data path shows as Figure 3.

2) I-Format instruction: it is similar to R-Format instruction. Comparing the two type instruction, the first register is the same source register *RS* , but the second register is write destination register *RD* and the rest bit is replaced by immediate. Load operation code of instruction to ALU, and load the data appointed by first register into ALU as first operand. Extend 16-bit immediate to 32-bit signed number or unsigned number and load it into ALU. After arithmetic is over, the answer is sent to register specified by the second register *RT*. I-Format instruction data path shows as Figure 4.

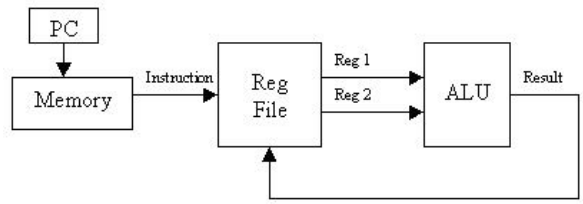


Figure 3 R-Format Instruction Data Path

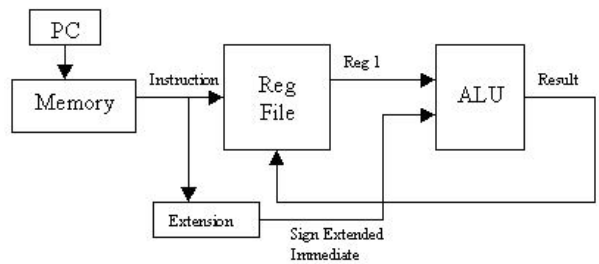


Figure 4 I-Format ALU Instruction Data Path

3) J-Format instruction: one register compares to constant 0 in register jump data path. If instruction is jump if zero instruction and content of register is 0, load value of the second register into program counter. If value of register is not 0 and instruction is jump if zero instruction, load next program counter value and execution instruction continues flowing. Jump if not zero instruction is similar to jump if zero instruction. The J-Format ALU instruction data path shows as Figure 5.

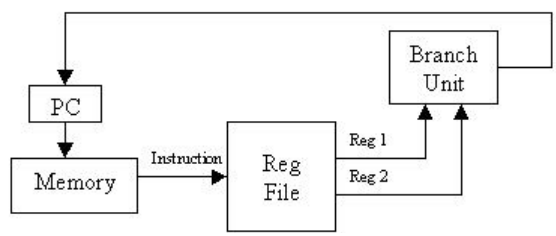


Figure 5 J-Format Instruction Data Path

3.2 General Register Design

Design of microprocessor general-purpose register file shows as Figure 6. Function of register group is to realize write data and read data and ensure correctness of write and read, moreover, another function of register file is to assure store data and send data stability. *Clk*(1 bit) is clock sign. Register receives operands from instruction register. The first operand is from5-bit source register *rs1*. The second operand is from 5-bits source register *rs2*, and the third operand is from 5-bit destination register *ws*. Furthermore, *wd*(32 bits) expresses the data to be written and *we*(1bit) expresses the write-enabled signal. Register file sends destination data *rd1*(32 bits) and *rd2*(32 bits) to two input ports of ALU respectively. Data of *rd1* and *rd2* are fetched form register appointed by value of *rs1* and *rs2*. The working theory of general register file shows as following: when R-Format instruction comes, fetch data from register appointed by signal of *rs1* and *rs2,* and sent the two data to *rd1* and *rd2*. Send the data to ALU. After arithmetic is over, send answer to register appointed by *ws* signal and *wd* receives the answer. The *we* judges whether loads data into the register file.

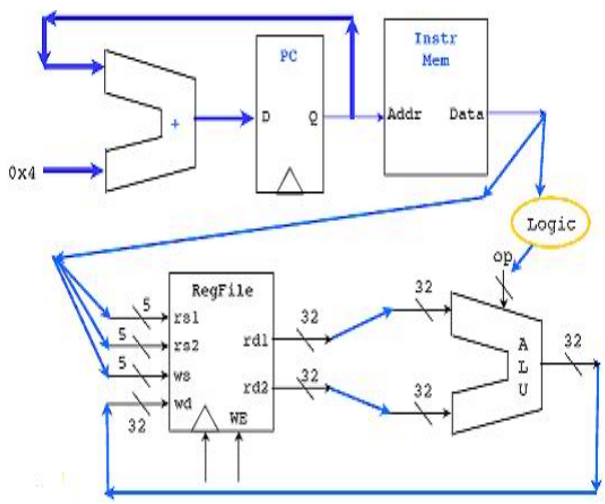


Figure 6 Unit Connection Graph

The internal structure of general file shows as Figure 7. The register group is composed of 32 32-bits D-triggers. The register has read part and write part according to function. When write data: by judging *ws* and *we, demus* logic unit generates write-enabled signal *en* of register and send the signal *en* to every 32-bit register. Only write the data to selected register whose *en* signal is 1. *Clk* sign controls time sequence. The *wd* sign inputs the new data which is to written to register. The purpose of design is to assure the correctness of write data. When positive clock edge comes and *we* signal and *ws* signal are all 1, write data to register to store. This process assures stability of data in register. When data read: all 32-bit D-triggers are all connected to the input port controlled by two *mux* logic units, which compose a 32-bit multi-path selector. The multi-path selector chooses corresponding register by judging signal of *rd1* and *rd2* and outputs data saved by the register to realize read function.

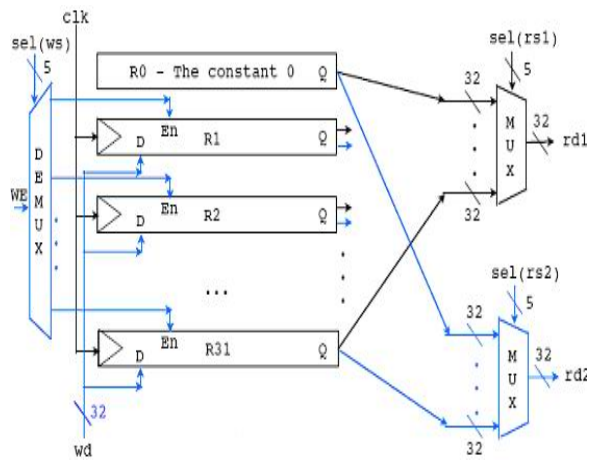


Figure 7 *RegFile* Design

It is known from Figure 8 there are different type instruction in CPU, such as R-Format instruction and I-Format instruction. I-Format instruction is immediate instruction, so it outputs value register appointed by *rs1* to *rd1* and extends 16-bit number to 32-bit number. And then, the two operands are all sent to two input port of ALU respectively. Finally, send result to *wd* port of register file and save the result to register file. The saving destination is determined by *ws*. According to colligate the forgoing two different type instruction, content of *rt* and *rd* can all input to *ws* port as operand part. We use *regdest* logic unit to choose different input signal by judging operation code with decoder. When *regdest* sign is 1, choose *rd* of three operands to *ws*. When r*egdest* sign is 0, choose *rt* of two operands to *ws*. In addition, immediate can extend to 32-bit number by *ext* logic unit.

The *ext* logic unit is controlled by *extop*(1 bit) signal. When *extop* is 0, it is zero-extend that the upper 16 bits are all set up 0; When *extop* is 1, it is sign-extend that the expansion upper 16 bits are all set up 1.

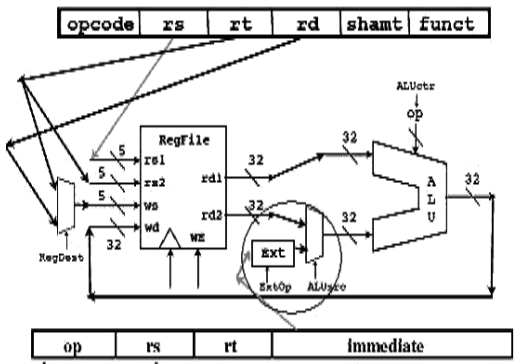


Figure 8 Theory Graph of Input Instruction into Register File

Figure 9 shows an entire general register file structure diagram. We implement operand write function and read function. of register group function. For example, when power is on CPU begins running. Program counter generates location 0, and then, instruction fetch unit fetches instruction from 0 address of main memory based on the 0. It is supposed that assembly language code is *ADDu R1,R2,R3*. Machine code of the instruction was sent to instruction register, and then it was sent to instruction decoder. if the highest 6 bits of the instruction are all constant 0 by decoder judging, the instruction is R-Format with no doubt. Then, parse three operands、*shamt* offset and *func* function bit in R-Format instruction. According to the *func* function bit *100001*, the instruction is add unsigned word instruction. After instruction resolution, the instruction was sent to ALU. This instruction is to realize *R3<=R1+R2*. Fetch operands saved in the former two register *RS*、*RT* and send the numbers to ALU. ALU do add unsigned word operation depending on the signal. The register is sent to destination register appointed by *R3*. Next instruction was executed in sequence after the instruction completed. Repeat all these steps in cycle and CPU just work like this. It is assumed that highest 6 bits of instruction are *000100*, which means the instruction is J-Format instruction. CPU sends jump address to program counter and instruction fetch unit fetches new address from program counter. And then, fetch instruction from memory and continues. As the same theory, if the instruction is recognized as I-Format instruction, the working theory of the I-format instruction is same as R-format.

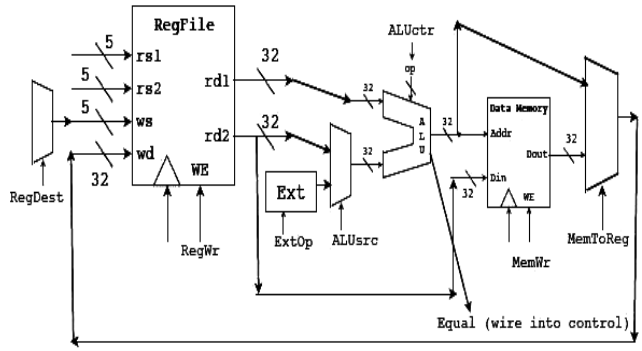


Figure 9 General Register File

4 Emulation verification

4.1 Signal Unit Waveform

1) *Demux* unit shows as Figure 10. *WE*(1bit) is write-enabled signal. *SEL*(5 bit)is input signal *ws* of register, *TOEN*(32 bit) is write-determined signal sent to each *en* port of 32-bit D-trigger. Figure 11 is *demux* unit waveform. Generate write signal of register 0、register 1、register 2、register 3 expressed by binary based on *WE* signal and *SEL* signal on every preceding four cycles. *TOEN* signal generates write signal sent to each 32-bits register. Enabled signals to input register 0、register 1、register 2 are generated in the 2nd cycle、the 3rd cycle and the 4th cycle and are expressed by hexadecimal in waveform.

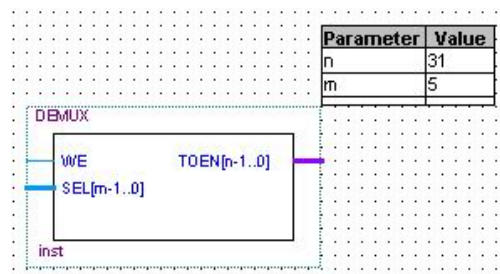


Figure 10 *demux* Unit

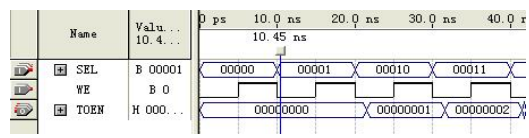


Figure 11 *demux* Unit Waveform

2) *Ext* unit shows as Figure 12. *EXTOP*(1 bit) receives immediate extend signal from controller. *DATAIN*(16 bit) receives 16-bits immediate. *DATAOUT* is extended to 32-bit immediate for input to ALU. Figure 13 shows *ext* unit waveform. According to *EXTOP* signal and *DATAIN* signal, immediate is zero-extended to 32-bit immediate in the 1st and the 3rd cycle , but immediate is sign-extended to 32-bit immediate in the 2nd and the 4th cycle . The *DATAOUT* expresses immediate extend result in the 1st cycle、the 2nd cycle、the 3rd cycle and the 4th cycle. In the waveform, the *DA*TAIN and DATAOUT are expressed by hexadecimal, but *EXTOP* is expressed by binary.

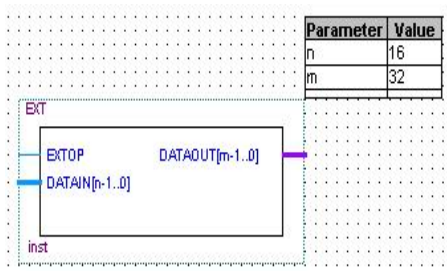


Figure 12 *ext* Unit

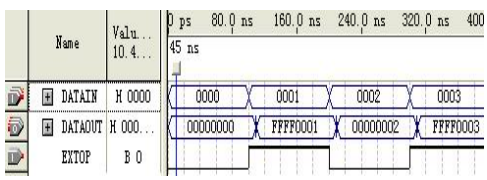


Figure 13 *ext* Unit Waveform

3)Register 0 unit shows as Figure 14. Depends on rules of RISC, N0.0 register of register group is to save constant 0 and it is not to input but for reading constant 0, so there is no input sign in but output sign Q in No. 0 register. It continues to output constant 0. Figure 15 shows as No. 0 register waveform. It is known that read data is always 0 from the data in the waveform.

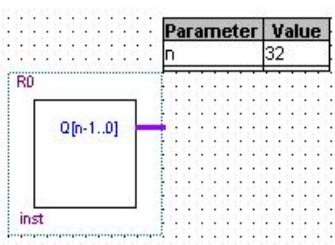


Figure 14 No.0 Register

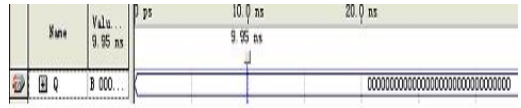


Figure 15 No.0 Register Waveform

4) Figure 16 shows as register group unit No. 0 register. Register includes *CLK*、*EN*、*D*、*Q*. *CLK*（1 bit）is clock sign. *EN*(1 bit) is write-enabled port for each register. Only if *EN* signal is generated by *demux* unit, data can be written in the register. The *D* (32 bit) is data port for write register and receives data from *wd*. The *Q*(32 bit) is output sign to send data to *rd1* or *rd2*, which sends data to ALU. Figure 17 is register waveform. Depending on *CLK* sign, only if positive clock edge comes and *EN* is 1, data can be written to register. Data store in register shows from the fifth cycle to the ninth cycle. *CLK* and *EN* are all expressed by binary, but *D* and *Q* are all expressed by hexadecimal.

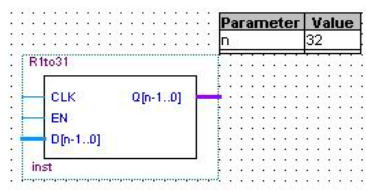


Figure 16 General Register

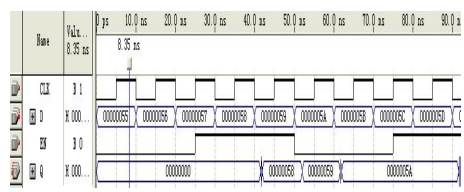


Figure 17 General Register Waveform

4.2 RegFile Unit Waveform Integration

Figure 16 shows encapsulation unit of register file. There is one 32-bit unit besides forgoing introduced units in register. The unit is automatically generated by Quartus II 4.1 system. *Rs1* and *rs2* are select input port which selects register to output data. *CLK*(1 bit) is clock signal. *RS1*(5 bit) receives *rs1* signal. *RS2*(5 bit) receives *rs2* signal. *RS1* and *RS2* value data which is read from specified register; *WD*(32bit) receives *wd*

signal expressing data to write; *WE*(1 bit) is write-enabled signal. *WS*(5 bit) receives *ws* signal, which specifies target operand register for saving answer; *RD1*(32 bit) and *RD2*(32bit) data that read from appointed register are all put to ALU. Figure 18 is waveform of register file. *CLK*、*RS1*、*RS2*、*WE*、*WS* are all expressed by binary. *WD*、*RD1*、*RD2* are expressed by hexadecimal. The meaning of the waveform is that: input different data into No.1st ~No.32rd register of register file in sequence in odd number cycle of *CLK* . The input data is decided by *WD*. In continuous two cycles, such as from first cycle to second cycle、from third cycle to fourth cycle and from fifth cycle to sixth cycle. Value of No.0th~ No.31st 32-bit register is read in sequence. Time for each signal to interact is given enough in testing, because contrast of each unit and internal circuit results in different delay slot. If the interact time is not enough, the result is usually wrong and waveform verification can not acquired..

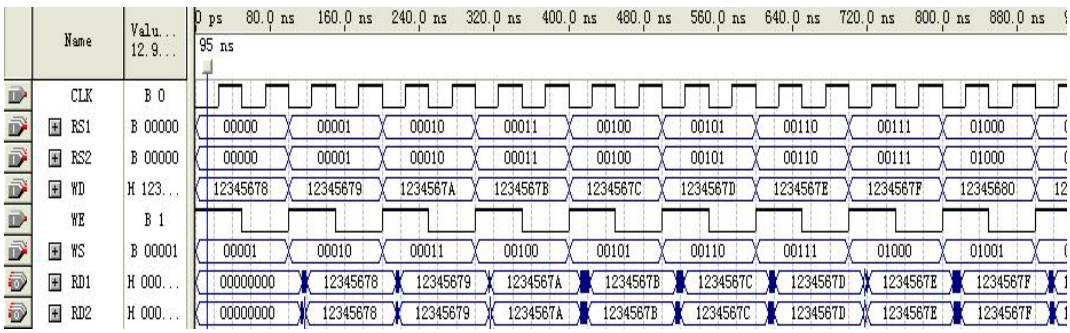


Figure 18 Register File Waveform

5 Conclusion

In this paper, we design general register of 32-bit embedded RISC microprocessor with SOPC (system on program chip) technology. The purpose of general register is to simplify embedding 32-bit RISC microprocessor, which is designed by oneself, into FPGA chip. Advantage of using SOPC to design embedded system shows up. The whole system can be accomplished in one chip. Not only simplify circuit design, but also interface speed is not bottleneck any more. Design of register file simulates、integrate and routes on Quartus II 4.3, and result indicates that design complete prospective function.

References

[1] Zheng-WeiMin, Tang-ZhiZhong. Computer System Structure (The second edition), Tsinghua University Press, 2006

[2] Pan-Song, Huang-JiYe, SOPC Technology Utility Tutorial , Tsinghua University Press, 2006

[3] MIPS32 4KTMProcessor Core Family Software User's Manual , MIPS Technologies Inc

[4] Mo-JianKun, Gao-JianSheng,Computer Organization, Huazhong University of Science and Technology Press, 1996

[5] Bai-ZhongYing, Computer Organization, Science Press, 2000. 11

[6] Zhang-XiuJuan, Chen-XinHua, EDA Design and emulation Practice [M]. BeiJing, Engine Industry Press. 2003

[7] "IEEE Standard of Binary Floating-Point Arithmetic" IEEE Standard754, IEEE Computer Society, 1985

[8] John L. Hennessy David A. Patterson Computer Organization &Design--The Hardware/Software Interface, Engine Industry Press, 1999.9

[9] Aldec Active-HDL the Design Verification Company Online Help

[10] Yi-Kui, Ding-YueHua, Application of AMCCS5933 Cont-roller in PCI BUS, DCABES2007, 2007.7

[11] Yi-Kui, Xiong-Pin, Ding-YueHua , 32 bit Floating-Point Addition and Subtraction ALU Design, DCABES2007, 2007.7

[12] Ding-YueHua, Yi-Kui, 32 bit Multiplication and Division ALU Design Based on RISC Structure, DCABES2007, 2007.7

Instruction Fetch Module Design of 32-bit RISC CPU Based on MIPS

Yuehua Ding Kui Yi Ping Sun

Department of Computer Science and Information Engineer, WuHan Polytechnic University, Wuhan, HuBei, 430023, China

Email: ykll1903@126.com

Abstract

In this paper, we analyze MIPS instruction format、instruction data path、decoder module function and design theory based on RISC CPUT instruction set. Furthermore, we design instruction fetch(*IF*) module of 32-bit CPU based on RISC CPU instruction set. Function of *IF* module mainly includes fetch instruction and latch module、address arithmetic module、 check validity of instruction module、synchronous control module. Function of *IF* modules are implemented by pipeline and simulated successfully on QuartusII.

Keywords: MIPS; Data Flow; Data Path; Pipeline

1 Introduction

Because memory was expensive in old days, designer of instruction enhanced complication of instruction to reduce program length. Tendency of complication instruction design brought up one traditional instruction design style, which is named “Complex Instruction Set Computer-CISC” structure. But great disparity among instructions and low universal property result in instruction realization difficulty and long running-time cost. Comparing to CISC, RISC CPU have more advantages, such as faster speed 、 simplified structure 、 easier implementation. RISC CPU is extensive use in embedded system. Developing CPU with RISC structure is necessary choice.

2 MIPS instruction system

2.1 MIPS Processor

Full name of MIPS is microcomputer without interlocked pipeline stages. Another informal full name is Millions of instructions per second. MIPS has already been pronoun of MIPS instruction set and MIPS instruction set architecture [1].

2.2 MIPS Instruction Set

ISA(Instruction Set Architecture) of processor is composed of instruction set and corresponding registers. Program based on same ISA can run on the same instruction set. MIPS instruction has been developed from 32-bit MIPSI to 64-bit MIPSIII and MIPSIV since it was created. To assure downward compatibility, every generation production of MIPS instruction directly extends new instruction based on old instruction but not abnegates any old instruction, so MIPS processor of 64-bit instruction set can execute 32-bit instruction.

All MIPS instructions are all 32-bit specified instruction and instruction address is word justification. MIPS divides instructions into three formats: immediate format(I-Format)、register format(R-Format) and jump format(J-Format)[2]. Three instruction format shows as Figure 1. Meaning of every instruction field as following:

- *OP*: 6-bit operation code;
- *rs*: 5-bit source register;

- *rt*: 5-bit temporary (source/destination)register number or branch condition;
- *immediate*: 16-bit immediate, branch instruction offset or address offset;
- *destination*: 26-bit destination address of unconditional jump;
- *rd*: 5-bit destination register number;
- *shamt*: 5-bit shift offset;
- *funct*: 6-bit function field;

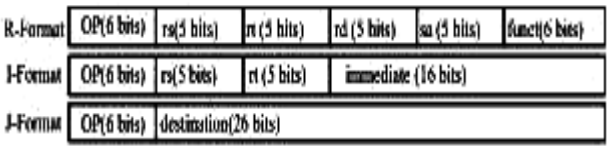


Figure 1 MIPS Instruction Format

MIPS instruction decoder or MIPS instruction execution is very high performance because of three type format with given length. Several simple MIPS instructions can accomplish complicated operation by complier [3].

3 Data flow

Data flow is determined by hardware data path, which express data flow process. There is no clear difference between data and control. Operation code、operand、memory address and value、register address and value、jump destination address and content are usually included in data, but control composes of control signal of unit、time sequence control signal and interrupt control signal, and these signals are not always defined clearly and strictly.

3.1 R-Format Data Path

In R-Format data path, fetch instruction from memory and analyze instruction into different parts. Two register specified by instruction fetch data from register file and ALU execute instruction command. Finally, after ALU outputs answer write the answer to register file. Figure 2 shows R-Format data path.

For example, *ADD R1,R2,R3* instruction, which is add signed word instruction(*R1= R2+ R3*). Data flow of this instruction shows as following: PC fetches *ADD R1,R2,R3* instruction from memory. At first, the instruction access two registers *R2* and *R3* and value of the two register is put to ALU. After arithmetic is over, ALU write back result to R1 register. And then, data flow is in the end.

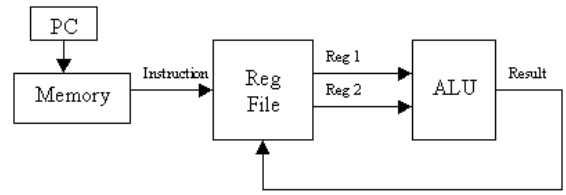


Figure 2 R-Format Instruction Data Path

For another example, *SRL R1,R2,R3* instruction, which is shift word right logical instruction. Data flow of this instruction shows as below: PC fetches *SRL R1,R2,R3* instruction from memory. At first, the instruction access two register *R2* and *R3* and value of the two register is put to ALU. After arithmetic is over, ALU write back answer to *R1* register, And then, data flow is in the end.

3.2 RI-Format Data Path

RI-Format instruction is similar to R-Format instruction[4]. The difference between them is that the second read register of R-format instruction is replaced by immediate of RI-Format instruction. The immediate is 32-bit signed number which is extend by 20-bit number, and put to ALU as the second operand. Finally, write-back result to register file. RI-Format data path shows as Figure 3.

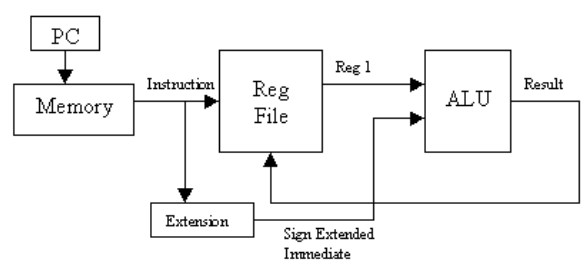


Figure 3 RI-Format Instruction Data Path

Format includes *ADDI R1, R2, data6* instruction ，

SUBI R1, R2, data6 instruction etc.

When *ADDI R1, R2, data6* instruction executes, PC fetches *ADDI R1, R2 data6* instruction from memory and register *R2* value is put to ALU. At the same time, immediate *data6* is extended to 32-bit signed number and put to ALU Finally, after ALU completes add of the two operands, ALU writes back answer to *R1* register. The difference data flow between *SUB R1,R2 data6* instruction and *ADD R1,R2,data6* instruction is that the former instruction do subtration.

3.3 Load Word Data Path

Load word data path is similar to I-Format data path. The difference between the two data path is that result is written to memory in load word data but result is written to register in I-Format. In load word data path, fetch data from memory and load it to register file. Load word data path shows as Figure 4.

LW R1, R2, data6 instruction is the only one instruction in load word data path. It works shows as below: PC fetch *LW R1, R2, data6* instruction from memory. *R1* register is to load data. Firstly send *R2* register value to ALU, at the same time, extend *data6* immediate to 32-bit and send it to ALU. The answer of

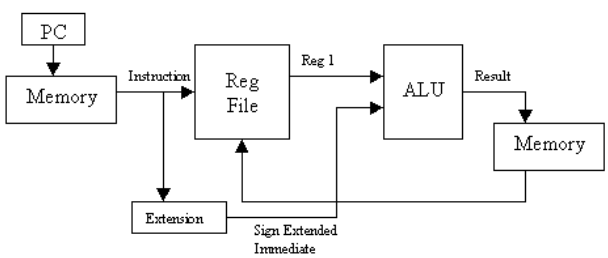


Figure 4 Load Word Data Path

adding the two numbers is memory address，And then, copy content of the memory address to *R1* register.

3.4 Memory Word Data Path

Memory word data path is similar to load word data path, but target which register is to write is memory but not register file.

There is only *SW R1, R2, data6* instruction in load word instruction. PC fetches *SW R1, R2, data6* instruction from memory. *R1* register stores data which is to be stored. Firstly, send *R2* register value to ALU, at the same time, extend *data6* immediate to 32-bit and send it to ALU. The result of adding the two numbers is memory address. Memory instruction data path shows as Figure 5.

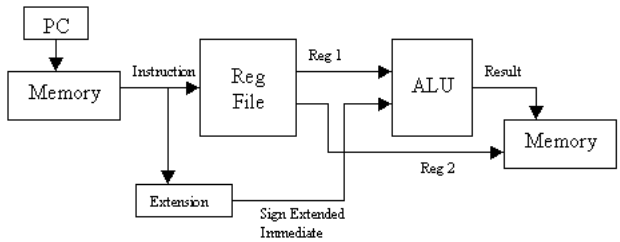


Figure 5 Memory Instruction Data Path

3.5 Register Jump Data Path

In register jump data path, one register compares to 0. When jump instruction is jump if zero instruction and register value is 0, the second register value loads to program counter. When jump instruction is jump if zero instruction and value in register is not 0, the next program counter value is loaded and instruction execution continues. Jump if not zero instruction is similar. Figure 6 shows jump instruction data path.

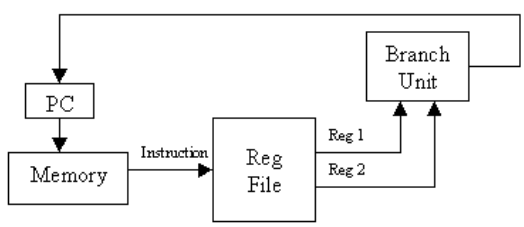


Figure 6 Jump Instruction Data Path

Register jump instruction includes two instructions: *BZ R1, R2* instruction and *BNZ R1, R2* instruction.

BZ R1, R2 instruction expresses if it is equal to constant 0 jump. Program counter fetches *BZ R1, R2* instruction from memory, and instruction accesses *R1* register and *R2* register. And then, send value of the two registers to branch unit. Branch unit judges whether *R1* value is equal to 0. If *R1* value is equal 0, send value of register *R2* to program counter. If *R1* value is not equal 0, PC adds 1 and program continues executing orderly.

BNZ R1, R2 instruction expresses if it is not equal to constant 0 then jump. Program counter fetches *BNZ R1, R2* instruction from memory, and instruction accesses *R1* register and *R2* register. And then, send value of the two registers to branch unit. Branch unit judges whether *R1* value is equal to 0. If *R1* value is not equal 0, send value of register *R2* to program counter. If *R1* value is equal 0, PC adds 1 and program continues executing in sequence.

4 Pipeline design

Pipeline decomposition enhances throughput rate of instruction. Clock cycle is decided by the slowest stage running time. In general words, pipeline includes five stages: instruction fetch(*IF*) 、 instruction decoder(*ID*)、execution(*EXE*)、 memory/ IO(*MEM*)、write-back(*WB*).

4.1 Instruction Fetch(*IF*)

Instruction fetch (*IF*) stage is request for instruction which is fetched from memory. Main component of *IF* stage shows as Figure 7. Instruction and PC is memorized in *IF/ID* pipeline register as temporary memory for next clock cycle.

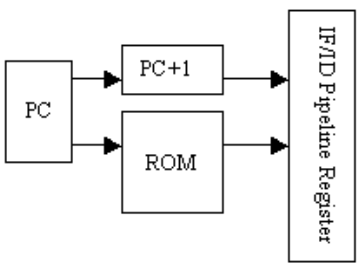


Figure 7 *IF* Stage

IF stage mainly depends on program counter(PC) current value. CPU fetches instruction from ROM based on PC value and PC adds 1 automatically. Finally, send all these information to *IF/ID* pipeline register to decoder.

4.2 Instruction Decoder(*ID*)

ID stage sends control command to other units of processor based on decode of instruction. Figure 8 shows *ID* stage structure. Instruction is sent to control unit and decoded here. Read register fetches data from register file. Branch unit is also included in *ID* stage.

Input of *ID* stage is from *IF* stage. *ID* stage decodes instruction to control signals and prepared operand. For example, if instruction is I-Format instruction, extend immediate to 32-bit data and access register file. If instruction is J-Format instruction, *EXE* stage comes after branch unit process completes.

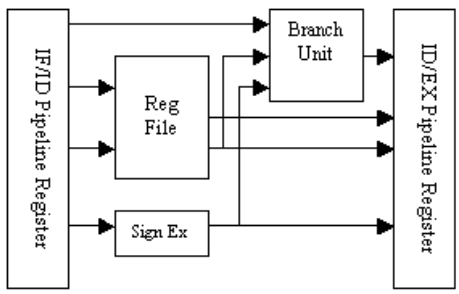


Figure 8 *ID* Stage

4.3 Execution（*EXE*）

EXE stage executes arithmetic. Main component of *EXE* stage is ALU. Arithmetic logic unit and shift-register compose of ALU. Figure 9 shows *EXE* stage structure. Function of *EXE* stage is to do operation of instruction, such as add and subtraction. ALU sends result to EX/MEM pipeline register before entering *MEM* stage.

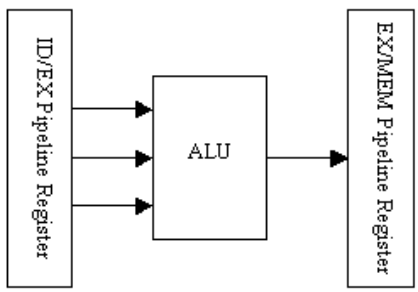


Figure 9 *EXE* Stage

4.4 Memory and IO (*MEM*)

Function of *MEM* stage is to fetch data from memory and store data to memory. Another function is to input data to processor and output data. If instruction is not memory instruction or IO instruction, result is sent to *WB* stage. *MEM* stage structure shows as Figure 10.

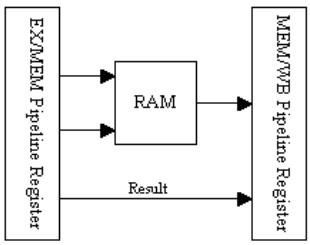


Figure 10 *MEM* Stage

Storing data in register is main function after result is calculated. Some result may be not stored in RAM definitely, and some result can be written to register directly. Give an example, some temporary variable is not memorized in RAM because of low execution efficiency. However, some data must be stored in RAM. Memory data in RAM or register depending on demands in *MEM* stage. There is a data copy in *MEM/WB* pipeline register.

4.5 Write-Back (*WB*)

WB stage charges of writing result、store data and input data to register file. The purpose of *WB* stage is to write data to destination register. For example, *ADD R1, R2, R3* instruction memories result in *R1* register to make program run faster. Figure 11 shows *WB* unit instruction.

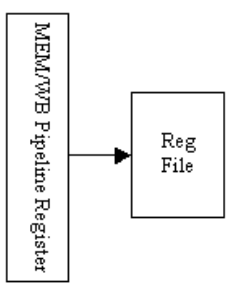


Figure 11 *WB* Stage

5 Instruction fetch stage design

5.1 Function Statement

Function of instruction fetch(*IF*) stage shows as below:

1) Fetch instruction and latch. Fetch instruction from instruction register depending on PC value and send the instruction to IF/ID pipeline register to latch.

2) Address arithmetic. Based on value of *sel[3..0]* in *pcselector*, select next value of PC from four address jump sources. These address jump sources are *incPC*、*branchPC*、*retiPC and retPC* .

- If instruction in *WB* stage of pipeline is jump instruction or successful branch instruction, select *branchPC* value and destination address of program jump acts as address arithmetic result;
- If instruction is not jump instruction or fail branch instruction, PC adds 1 automatically and points to next instruction in instruction register;
- If instruction is interrupt-return instruction, select *retiPC* value;
- If instruction is subprogram return instruction, select *retPC* value.

3) Check validity of instruction. Check operation code and function code validity based on definition of instruction set. If instruction is wrong, an exception is thrown.

4) Synchronous control. Use *CLK* to control synchronous of external sign.

5.2 Module and Implementation

IF stage includes five modules: *incPC*、*lpm_rom0*、*progc*、*pcselector* and *ifid*. Figure 12 shows connection of each module.

Their function shows as below:

- *incPC*: PC adds 1 automatically. PC points to address of next instruction;
- *lpm_rom0*: application store program;

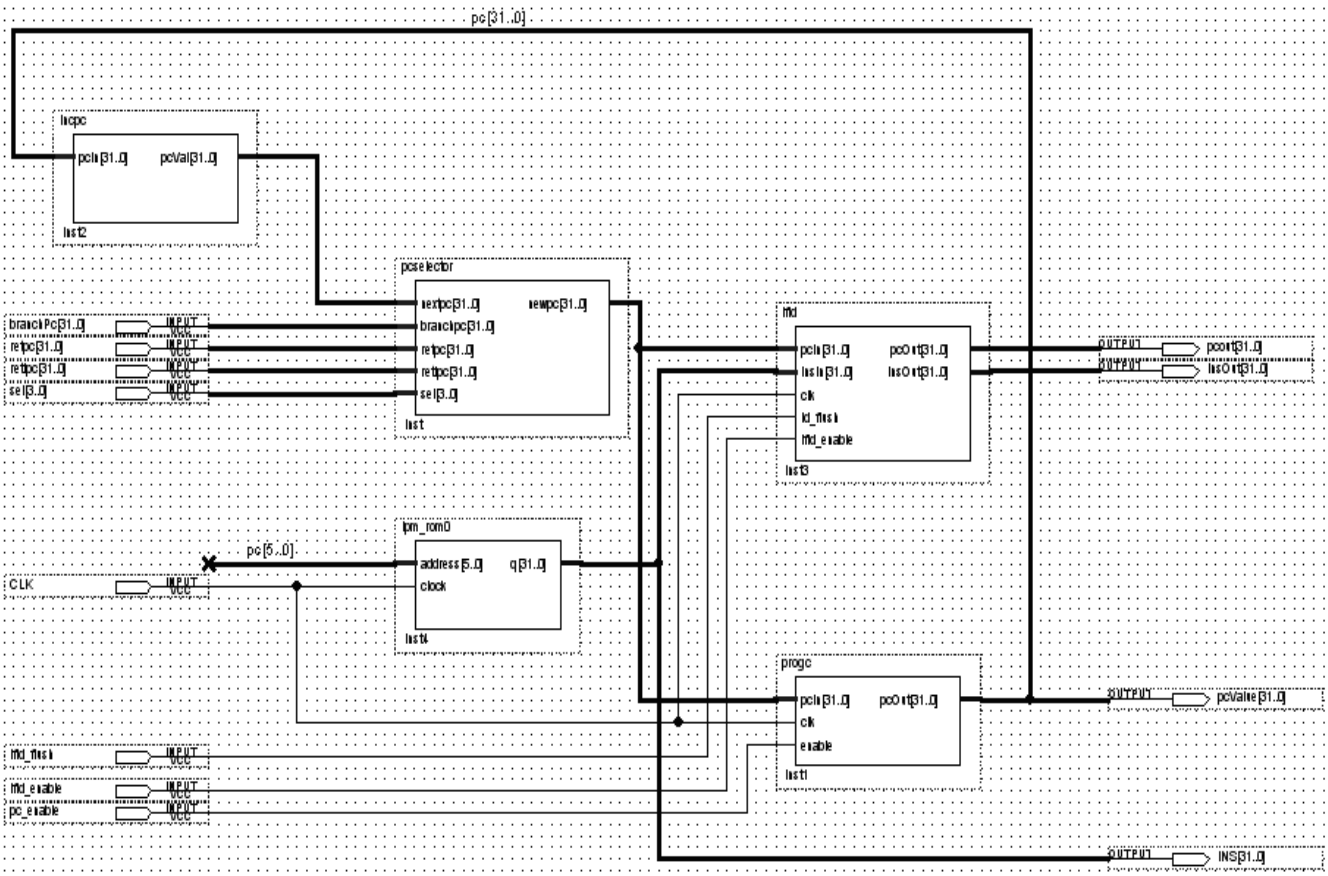


Figure 12 *IF* Circuit Diagram

- *progc*: program counter;
- *pcselector*: control next instruction selection;
- *ifid*: pipeline latch.

Every module uses VHDL to describe. Input signal of *IF* stage includes *branchPC*、 *retPC*、*retiPC*、*sel*、*clk*、*ifid_flush*、*ifid_enable* and *pc_enable*. Their function shows as below:

- *branchPC*: jump address of branch signal
- *retPC*: subprogram return address signal
- *retiPC*: interrupt return address signal
- *sel* : selection signal from *pcselector* in *EXE* stage
- *clk*: clock signal
- *ifid_flush*: data signal
- ifid_enable、pc_enable: control signal

Output signal of *IF* stage includes *ins[31..0]*、*pcvelue[31..0]* 、 *insOut[31..0]*and *pcout[31..0]*. Their function shows as below:

- *ins[31..0]*: instruction code fetch from instruction register;
- *pcvelue[31..0]*: PC value in *IF* stage;
- *insOut[31..0]*: instruction code which is to sent to next stage and comes from pipeline register *ifid*;
- *pcout[31..0]*: program counter value.

Module Implementation shows as below:

（1） *pcselector* module. Input port includes *nextpc[31..0]* 、 *branchpc[31..0]* 、 *retpc[31..0]* 、 *retipc[31..0]* and *sel[3..0]*. Output port includes *newpc[31..0]*. Select data from four source data as next instruction address determined by *sel[3..0]*. The four source data are *nextpc[31..0]* 、 *branchpc[31..0]* 、 *retpc[31..0]* and *retipc[31..0]*.

Input signal are *nextPC*、*branchPC*、*retPC*、*retiPC* and *sel.* Output signal are *newPC.* Function of input signal shows as below:

- *nextPC*: next instruction address;
- *branchPC*: address of branch jump signal;

- *retPC*: subprogram return address signal;
- *retiPC*: interrupt return address signal;
- *sel*: selector signal.

Time sequence simulation waveform of *pcselector* shows as Figure 13. Input different address sign into *nextpc*、*branchpc*、*retpc*、*retipc* ports, and *newpc* selects one of the four input signal to output depending on value in *sel[3..0]*.

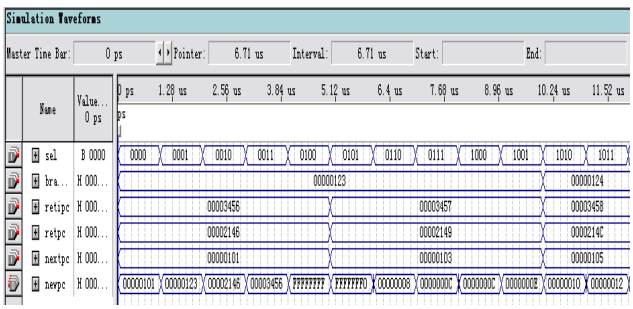


Figure 13 *pcselector* Stage Simulation Waveform

（2）*progc* module. Input port includes *pcin[31..0]*、*clk* and *enable*. Output port includes *pcout[31..0]*. The function of the module is to communicate with instruction memory. When positive clock edge comes, send value of address bus *pcin[31..0]* to instruction memory and fetch next instruction from *ins[31..0]*. *ins[31..0] is* output of instruction memory. Send instruction out when negative clock edge comes. Figure 14 shows *progc* module simulation waveform.

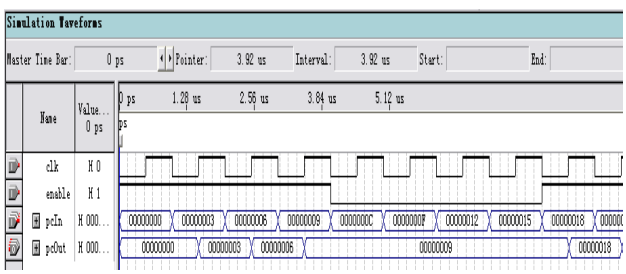


Figure 14 *progc* Module Simulation Waveform

3) *incPC* module. Input port includes *pcin[31..0]* and output port includes *pcout[31..0]*. The function of *incPC* module is to PC add 1 and the new PC cat as one optional value. When negative clock sign comes, PC value is sent to *pcselector* module. Figure 15 shows *incPC* module entity structure and RTL structure. Figure 16 shows simulation waveform of *incPC* module. We can know *pcIn* value adds 1 and send result to *pcVal* from waveform.

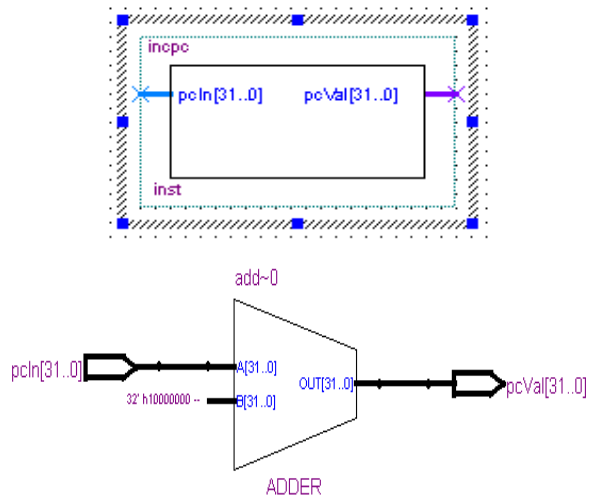


Figure 15 *incPC* Entity Structure and RTL Structure

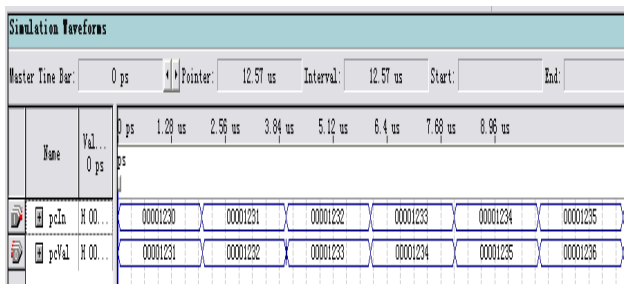


Figure 16 *incPC* Module Simulation Waveform

4) *lpm_rom0* module. Input port includes *address[5..0]* and *inclock*. Output port includes *q[31..0]*. Function of the module is to memory program machine code. Access memory location which is specified by address bus *address[5..0]*, moreover, fetch next instruction from memory and send out the instruction by instruction bus *q[31..0]*.

lpm_rom0 module can be implemented EAB of FPGA by calling macro function module. Adopt *lpm_rom* structure in macro function library to realize the module. Parameter configuration is that address bus *address* is 6-bit and output bus *q* is 32-bit. Process of *lpm_rom0* is described as following: when positive *inclock* edge comes, latch *address[5..0]* and output the data pointed by value of *address[5..0]* to output port *q[31..0]*. Set up data in *lpm_rom0* by memory initialization file (*.mif*), or edit、update and reload data on debugging by system memory editor tool.

5) *ifid* module. Input port includes *pcin[31..0]*、

insin[31..0]、*clkid_flush* and *ifid_enable*. Output port includes *pcout[31..0]* and *insout[31..0]*. Function of *ifid* is to latch *PC* and *instr* of Statge1 and send them to next stage.

Time sequence simulation waveform of *Ifid* module shows as Figure 17. We can see fact that when *ifid_enalbe* is high level and *id_flush* is low level, data are not relative in pipeline. When positive edge of *clk* comes, values of *insOut* and *pcOut* are same to *insIn* and *pcIn* respectively; When *ifid_enable* and *id_flush* are all high level, data is relative in pipeline. When positive edge of *clk* comes, *insOut* changes to "0000H", but *pcOut* maintains its original value ; After pipeline conflicts, *insOut* and *pcOut* returns to normal working state; if *ifid_enable* is low level, pipeline stops working and *insOut* and *pcOut* maintain its original state.

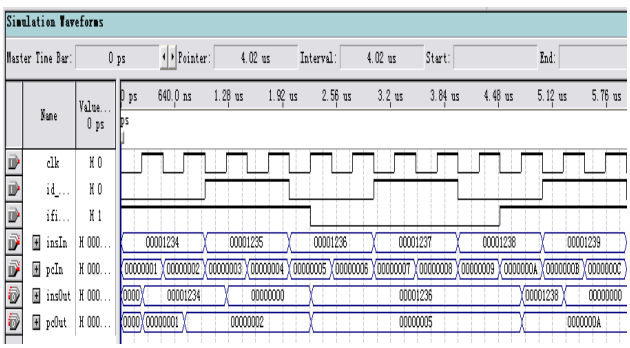


Figure 17 *ifid* Module Simulation Waveform

6 Conclusion

In this research, we adopt top-down design method and use VHDL to describe system. At first, we design the system from the top, and do in-depth design gradually. The structure and hierarchical of design is very clear. It is easy to edit and debug. Design of instruction fetch (*IF*)stage simulates、integrate and routes on Quartus II 4.3. The result indicates *IF* stage completes prospective function.

References

[1] Bai-ZhongYing, Computer Organization, Science Press, 2000.11

[2] Wang-AiYing, Organization and Structure of Computer, Tsinghua University Press, 2006

[3] Wang-YuanZhen, IBM-PC Macro Asm Program, Huazhong University of Science and Technology Press, 1996.9

[4] MIPS Technologies, Inc. MIPS32™ Architecture For Programmers Volume II: The MIPS32™ Instruction Set，June 9, 2003

[5] Zheng-WeiMin, Tang-ZhiZhong. Computer System Structure (The second edition), Tsinghua University Press,2006

[6] Pan-Song, Huang-JiYe, SOPC Technology Utility Tutorial , Tsinghua University Press,2006

[7] MIPS32 4KTMProcessor Core Family Software User's Manual , MIPS Technologies Inc

[8] Mo-JianKun, Gao-JianSheng,Computer Organization, Huazhong University of Science and Technology Press, 1996

[9] Zhang-XiuJuan, Chen-XinHua, EDA Design and emulation Practice [M]. BeiJing, Engine Industry Press. 2003

[10] "IEEE Standard of Binary Floating-Point Arithmetic" IEEE Standard754, IEEE Computer Society, 1985

[11] Yi-Kui, Ding-YueHua, Application of AMCCS5933 Controller in PCI BUS, DCABES2007, 2007.7

[12] Yi-Kui, Xiong-Pin, Ding-YueHua , 32 bit Floating-Point Addition and Subtraction ALU Design, DCABES2007, 2007.7

[13] Ding-YueHua, Yi-Kui, 32 bit Multiplication and Division ALU Design Based on RISC Structure, DCABES2007, 2007.7

Implementation and Simulation of A QoS Signaling Protocol for Mobile IP Networks*

Gang Nie

College of Computer Science, Wuhan University of Science & Engineer ,Wuhan, Hubei, 430073, China

Email: ng@wuse.edu.cn

Abstract

QoS signaling protocol is one of the key components in Internet QoS architectures to establish, maintain, and remove reservation states in mobile environments. This paper identifies the key issues on the interaction between Mobile IPv6 and QoS signaling, and proposes desired QoS signaling functions for mobile environments. The key features of the protocol include crossover node discovery and local repair to achieve seamless services. Our simulation and experimental results show that the proposed protocol works well in mobile environments.

Keywords: QoS; Signaling protocol; Mobile IPv6

1 Introduction

With the rapid increase of portable devices such as laptops, PDAs, hand-held computers, and a variety of wireless devices, mobile computing applications have become more practical. Real-time services such as Internet telephony, video conferencing, and video-on-demand should be available in the mobile computing environments. It is important for the mobile Internet environment to provide QoS guarantees in the near future.

The IETF Next Steps in Signaling (NSIS) working group is standardizing a signaling protocol suite with QoS signaling as the first use case [1]. The lower general layer in the NSIS signaling protocol suite, called the NSIS Transport Layer Protocol (NTLP), is intended to provide a general transport service for signaling messages. The actual signaling messages are generated within upper layer signaling applications, each having its own NSIS Signaling Layer Protocol (NSLP). The main functionality of the NTLP is to discover appropriate NSIS nodes and to deliver the signaling messages to them. The general description of the NSIS protocol suite, including its two-layer architecture, can be found in [2].

One of the important features that the NSIS protocol needs to provide is mobility support [3]. In highly mobile environments, frequent handovers may result in a significant degradation of QoS performance if the mobile access network is unable to provide enhanced solutions for prompt QoS reestablishment. Especially, how QoS signaling interacts with Mobile IP may have a significant impact on QoS performance.

This paper mainly identifies the key issues on the interaction between Mobile IPv6 and QoS signaling, and proposes desired QoS signaling functions for mobile environments. The rest of the paper is organized as follows: Sections 2 and 3 describe the proposed protocol called mobility-aware QoS signaling protocol (MQSIG), Section 4 presents the simulation results. Section 5 makes the final considerations.

2 Impact of mobility on QoS signaling protocols

IP-based mobility itself includes topological changes due to the change of network attachment point. Topological changes entail the change of routes for data packets sent to or from a mobile node (MN) and may

* This research was supported by the Educational Ministry of Hubei Province, China under Grant D200717005.

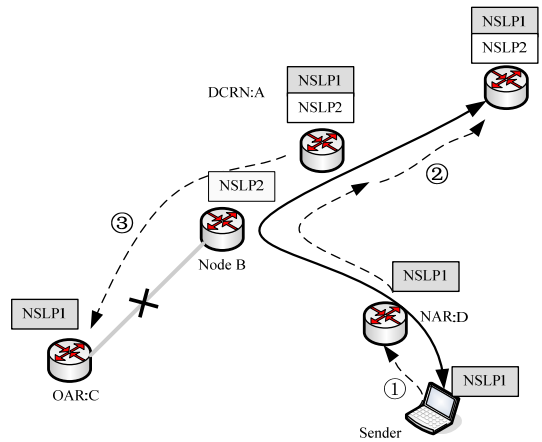
lead to the change of host IP addresses. These changes of route and IP addresses in mobile environments are typically much faster and more frequent than traditional route changes and may have some significant impact on QoS signaling protocols.

Although the well-known resource reservation protocol, RSVP [4], is able to setup resource reservation for real-time traffic in the wired Internet, it is not adequate to reserve resources in mobile networks. For example, a change in the location of the MN may make the reserved resources on the old path useless and a new reservation on the new path has to be established while maintaining the old reservation. This results in the inefficient use of network resources and also introduces an additional de-lay due to end-to-end signaling.

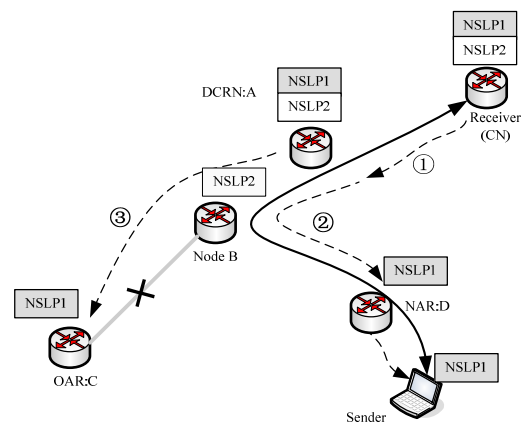
To overcome such drawbacks of RSVP, many solutions have been proposed, which are mostly based on the modification or extension of RSVP [5, 6, 7]. However, most of RSVP-based solutions do not have yet appropriate QoS mechanisms for preventing service disruption during handover [8].

To develop a mobility-aware QoS signaling protocol which solves the problems, it is necessary to analyze the key differences between generic route changes and mobility [9, 10]. The generic route changes occur due to load sharing, load balancing, or a link (or node) failure, but the mobility is associated with the change of the network attachment point. These will cause divergence between the old path where QoS state has already been installed and the new path where data forwarding will actually happen. Although the mobility can be considered similar to the route changes, the main difference is that the flow identifier may not change after the route changes while the mobility may cause the change of flow identifier by having a new network attachment point. Since the reservation session should remain the same after a mobility event, the flow identifier should not be used to identify any signaling application session [2].

In general, a mobility event results in creating a common/unchanged path, an old path, and a new path. The old and new paths converge or diverge according to the direction of each signaling flow as shown in Figure 1. Such topological changes make the signaling state established on the old path useless, and thus it should be removed, the existing QoS state should be re-established along the new path as fast as possible and updated along the common path.



(a) Downstream flows



(b) Upstream flows

Figure 1 The topology for NSIS signaling caused by mobility

To minimize the impact of mobility on seamless QoS service and to improve the scalability of signaling, QoS signaling for state re-establishment should be localized within the affected area. This localized signaling procedure is referred to as local repair in this paper. The major issue in this case is to find a node which performs the local repair. One of the most appropriate nodes for the local repair is the crossover node (CRN) where the old and new session paths meet..

3 An IP mobility-aware QoS signaling protocol

In this section, we propose an efficient QoS signaling protocol (MQSIG) which operates adaptively in IPv6-based mobile networks. The key features of MQSIG include CRN discovery and local repair.

CRN Discovery

The CRN discovery can be performed according to whether the discovery is coupled with the transport of signaling application messages (i.e., a coupled approach and an uncoupled approach). Generally, the coupled approach would be preferred to the uncoupled approach to reduce the signaling delay. In this paper, the CRN discovery and local repair are based on the coupled approach. We also assume that the CRN discovery is considered as an extension to the peer discovery at the NTLP layer to reduce overall processing overhead [3].

1) Identifiers for CRN Discovery: To discover the CRN in a fast and efficient manner, the following basic identifiers are required: session identifier (Session_ID), flow identifier (Flow_ID), signaling application identifier (NSLP_ID), and NSLP branch identifier (NSLP_Br_ID).

The Session_ID is contained in the NTLP message and used to easily identify the involved session because it remains the same while the Flow_ID may change after handover. Note that the uniqueness of Session_ID is one of the keys features to solve the double reservation problem. On the other hand, the Flow_ID is used to specify the relationship between the address information and the QoS state re-establishment. In other words, the change of Flow_ID indicates topological changes, and therefore it represents that the state along the common path should be updated after the CRN is discovered.

The NSLP_ID is used to refer to the corresponding NSLP at the NTLP level, and in the context of CRN discovery it helps to discover an appropriate NSLP CRN which has installed the corresponding QoS state using the NTLP peer discovery message.

The NSLP_Br_ID is a virtual branch identifier and used to establish or delete NSIS associations between peer nodes. It is also used as an identifier to determine the CRN at the NTLP layer. The NSLP_Br_ID consists of the location information of peer nodes with the corresponding NSLP_ID by the procedure of NTLP message association [8]. For instance, as shown in Figure 1 (a), for the downstream direction (i.e., the direction from a data sender towards the destination) NSLP1 (state) of node A requires a messaging association for sending its messages towards node D after a route changes. In this case, NSIS entity (NE) A creates an NSLP_Br_ID for NSLP 1 to-wards node D and increases the NSLP_Br_ID counter to locally distinguish each virtual interface identifier between adjacent NSLP peers. That is, the corresponding NSLP_Br_ID is 1-D-#2: 1, D, and #2 indicate NSLP_ID-flow, flow directions (Downstream (D) or Up-stream (U)), and the counter number of branch, respectively. Note that this identifier would be more useful when the physical merging point of the old and new paths is not an NSLP CRN after the route changes as shown in Figure 1.

Optionally, the Mobility identifier as an object form can be used to inform of the handover status or a route change and therefore to expedite the CRN discovery. The Mobility object is defined in the NTLP (e.g., in GIMPS payload) or NSLP messages to notify of any mobility event explicitly, and it contains various mobility-related fields such as mobility_event_counter (MEC) and handover_init (HI) fields. The ‘MEC’ field is used to detect the latest handover event to avoid any confusion about where to send a reservation confirmation message and to handle the ping-pong type of movement. The ‘HI’ field is used to explicitly inform that a handover is now initiated for fast state re-establishment.

2) The Procedures for CRN discovery: When a mobility event occurs, the CRN can be recognized by com-paring the existing stored identifiers with the identifiers included in the NSIS peer discovery message initiated by an NSIS initiator (NI) (e.g., an MN or a CN). If an NSIS message is routed to an NSIS peer node, the node should check the following information to determine whether it is a CRN:

- Whether the same NSLP ID exists
- Whether the corresponding CRN has already been discovered
- Whether the same Session_ID and Flow_ID exist
- Whether the NSLP_Br_ID has been changed; for example, as shown in Table I, for NSLP 1 it is changed to 1-D-#2 from 1-D-#1
- Optionally, whether any Mobility identifier exists; for example, the Mobility object may be examined to find out which message is sent due to the latest handover by checking the MEC field.

The CRN discovery can be further divided into downstream CRN (DCRN) discovery and upstream CRN (UCRN) discovery (owing to asymmetric routing) ac-cording to which node is a signaling initiator (by up-stream or downstream), or whether the MN is a data sender. The procedure of DCRN discovery is similar to the creation procedure of the routing table of node N as shown in Figure 1 (a), and the procedure of UCRN discovery is similar to Figure 1 (b). Note that since the UCRN is determined by examining whether the outgoing path diverges or not, the UCRN discovery is more complex than the DCRN discovery [3].

Local repair Procedures

The CRN discovery procedures are different according to the direction of signaling flows in mobility scenarios, and therefore the procedures for local repair also differ depending on the direction of signaling flows: downstream local repair and upstream local repair.

For either type of local repair, the NSIS protocol needs to interact with mobility signaling protocols (e.g., Mobile IPv6), if any (during or posterior handover), to achieve fast re-establishment of the NSIS states along the new path. For example, the signaling protocol should interact with the binding process of Mobile IPv6 through several methods to immediately perform CRN discovery and the local repair [3].

In the downstream local repair, if resource availability is assured (after detection of mobility by Binding Update (BU) message), the MN initiates NSIS signaling for state setup towards a CN along the new path, and the DCRN discovery is implicitly done by this signaling (Figure 1 (a)-①). The node where the old and new logical session paths converge realizes that it is a DCRN (e.g., node A) through the CRN discovery procedures described above, and afterward it sends a response message toward the MN to notify of the installed NSLP state. The DCRN then sends a refresh message towards the signaling destination to update the changed Flow_ID on the common path (Figure 1 (a)-②), and it also sends a teardown message towards the old AR to delete the NSIS states on the obsolete path (Figure 1 (a)-③). In case of upstream local repair, the CN (or a HA) sends a refresh message toward the MN to perform local repair (Figure 1(b)-①). The UCRN is discovered implicitly by the CN-initiated signaling along the common path, and the node from which the common path begins to diverge into the old and new logical paths realizes that it is a UCRN (e.g., node N). In this case, the CN should be informed of the mobility event by detecting a change in its binding entry (BU message). After the UCRN is determined, it may send a refresh message to the MN along the new path while establishing the NSIS association between the updated peers (Figure 1 (b)-②), and afterward the UCRN may send a teardown message toward the old AR to delete the NSIS state on the obsolete path (Figure 1 (b)-③).

One of the goals of local repair is to avoid double reservations on all paths described in Section II. The double reservation made along the common path can be torn down by establishing a signaling association using the unique Session_ID and by updating packet classifier/flow identifier. In this case, the NSLP state should be shared for flows with different flow identifiers. After re-establishment of the NSIS state along the new path, the state on the obsolete path needs to be quickly re-moved by the local repair mechanism to prevent waste of resources (and resource allocation problem by call blocking). Although the release of the state on the old path can be accomplished by the timeout of soft state, the refresh timer value is quite long (e.g., default value of 30 s in RSVP [4]) and the maintenance of the obsolete state in mobile environments may not be necessary. Therefore, the transmission of a teardown message is particularly preferred to the use of refresh timer to delete

the old state in a fast manner.

The release of old state on the obsolete path is also accomplished by comparing the existing and the new NSLP_Br_IDs. This will prevent the teardown message from being forwarded toward along the common path. However, whether the teardown message can be sent toward the opposite direction to the state initiating node is still for further study. This also leads to authorization problem because a node which does not initiate signaling for establishing the NSIS state can delete the state.

4 Performance analysis

In this section, we mainly evaluate the performance of MQSIG in terms of resource re-reservation delay on the new path after handover. Then, we compare the related performance of existing signaling protocols such as RSVP and RSVP-MP [6]. The experimental result is also provided to demonstrate that the proposed signaling protocol works well in MIPv6-based mobile environments.

We have performed simulation studies to measure the performance of RSVP, RSVP-MP, and MQSIG in terms of resource re-reservation delay. Figure 2 depicts a simulation topology where there are 8 MNs. The number of hops from the MN and the CN is 7, and every MN may generate UDP traffic. It is assumed that the refresh period of RSVP and RSVP-MP is 30s. Initially, only one MN which communicates with the CN generates UDP traffic. The traffic load increases when MNs other than the current MN begins to generate UDP traffic. For example, the amount of added traffic load is 0.1 when another MN starts to generate traffic. Our simulation model is based on Marc Greis' RSVP model implemented in ns-2.1b3 and Rui Prior's RSVP model implemented in ns-2.26 which is an updated version of Marc Greis' model.

As depicted in Figure 3, the proposed signaling protocol shows better performance in terms of signaling delay for resource re-reservation, compared to RSVP and RSVP-MP even when the traffic load increases. This is because MQSIG performs CRN discovery for localized signaling after handover and MIPv6 binding process is closely associated with QoS signaling for fast resource re-reservation. Furthermore, only QoS state update (not re-reservation) is performed on the common/unchanged path to minimize the signaling delay through local repair procedures.

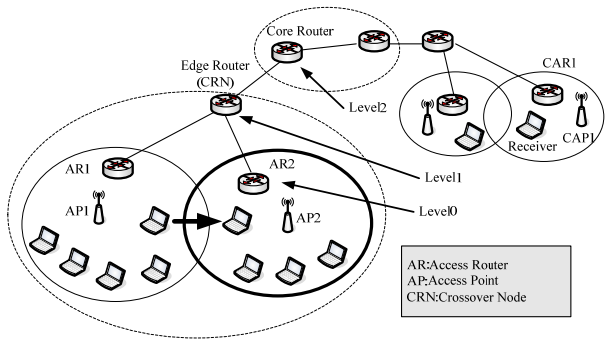


Figure 2 Simulation Topology

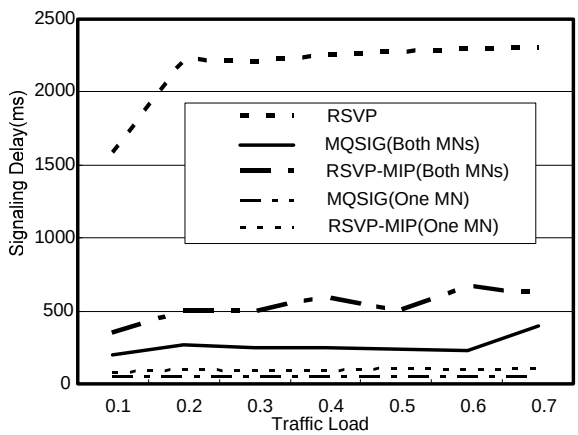


Figure 3 Signaling delay for re-reservation vs. traffic load

5 Conclusions and future work

In this paper, we identified some crucial issues including double reservation and end-to-end signaling problems which may occur when QoS signaling interacts with macro-mobility management protocols (e.g., Mobile IPv6). Based the analysis, we proposed a mobility-aware QoS signaling protocol (MQSIG). We also demonstrated that the proposed signaling protocol reduced the resource re-reservation delay by fast localized CRN discovery and local repair mechanisms.

References

[1] G. Karagiannis et al., "Next Steps in Signaling (NSIS): Framework", RFC4080, June 2005

[2] R. Hancock, "Next Steps in Signaling: Framework", draft-ietf-nsis-fw-04, September 2003

[3] S. Lee, S. Jeong et al., "Applicability Statement of NSIS Protocols in Mobile Environments", draft ietf nsis applicability mobility signaling-00.txt, October 2004

[4] B. Braden, L. Zhang et al., "Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specification", RFC 2205, September 1997

[5] A. Talukdar, B. Badrinath, and A. Acharya, "MRSVP: A Resource Reservation Protocol for an Integrated Services Packet Network with Mobile Hosts", Technical report DCS-TR-337, Rutgers University, 1997

[6] W. Chen and L. Huang, "RSVP Mobility Support: A Signaling Protocol for Integrated Services Internet with Mobile Hosts", Proc. IEEE Conference on Computer Communications, Vol. 3, pp.1283-1292, 2000

[7] I. Mahadevan and K. M. Sivalingam, "architecture and Experimental Results for Quality of Service in Mobile Networks using RSVP and CBQ", ACM Wireless Networks 6, pp. 221-234, July 2000

[8] Terzis, A., Srivastava, M., Zhang, L., "A Simple QoS Signaling Protocol for Mobile Hosts in the Integrated Services Internet," Proceedings of IEEE INFOCOM, Vol. 3, p. 1011-1018, March 1999

[9] López, A., Velayos, H., Manner, J., "Reservation Based QoS Provision for Mobile Environments," 1st IEEE Workshop on Services and Applications on the Wireless Public Interface, Volume 7, July 2001

[10] Chaskar, H., Koodli, R., "QoS support in Mobile IP version 6," IEEE Broadband Wireless Summit, May 2001

Community Structure in B. thuringiensis Metabolic Network

Dewu Ding[1,2] Yanrui Ding[1,*] Wenbo Xu[1,*] Kezhong Lu[2] Shouwen Chen[3]

1 School of Information Technology, Jiangnan University, Wuxi, Jiangsu ,214036, China

Email: dwding@yahoo.com.cn; yr_ding@yahoo.com.cn; xwb_sytu@hotmail.com

2 Department of Mathematics and Computer Science, Chizhou College, Chizhou, Anhui ,247000, China

Email: luke76@163.com

3 State Key Laboratory of Agricultural Microbiology, Huazhong Agricultural University ,Wuhan, Hubei ,430070, China

Email: chenshouwen@mail.hzau.edu.cn

Abstract

A number of recent studies have focused on the community structure (i.e. dense node-node links within communities but sparser links between them) in complex network systems such as social networks and biological networks, and it suggested that community structure could helpful for understanding the organizational principles of these networks. In this article, we firstly investigate the basic structure in reconstructed metabolic network of *B. thuringiensis* based its "bow tie" structure characteristic. Subsequently, the most important part giant strong component (GSC) is extracted and decomposed based on simulated annealing algorithm, and the results suggested that these divisions are functional significant for metabolism by compared to pathway information in KEGG.

Keywords: Bow Tie; Community Structure; Modularity; Simulated Annealing Algorithm

1 Introduction

The discovery of fundamental organizational principles that underlie complex biological systems is a prime goal in the emerging systems biology [1]. Advances in complex networks, especially in biology networks in recent years fuelled this expectation[2-5], global topological structural properties are achieved[6-7] and network-based pathway analysis methods are proposed [8-9] for the purposes.

However, these global properties such as so-called "small-world", "scale-free" etc only reflect one aspect of the global topological structure of the networks, while network-based pathway analysis methods are hardly applied to genome-scale metabolic networks due to combinatorial explosion of pathways. It is suggested that metabolic networks should have modularity [10-12] which is similar to other complex networks, such as social networks, Internet, Worldwide Web etc. Thus, to discover functional information involved in metabolic networks, we need identify functional modules in networks.

As defined in social networks, the basic principle for defining community in biology networks is also that dense node–node links within communities but sparser links between them [13]. A simple paradigm network with such a community structure shown in fig.1 is used to illustrate this characteristic. There are three communities of dense connected nodes in the network, with a much sparser links (gray lines) between them. Community structure detection has attracted considerable attention in biology, physics, computer science and other fields recent years [13-15], and could obviously have practical applications. For instance, they often correspond to functional units such as cycles or pathways in metabolic networks.

* Corresponding authors

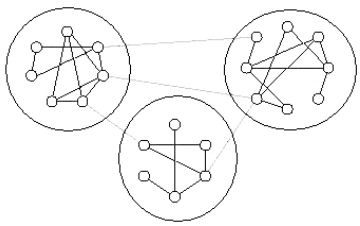


Figure1 A simple paradigm network with three communities

In this paper, we first introduce community structure in complex networks. Subsequently, structural and functional analysis of B. thuringiensis metabolic networks based "bow tie" are explained and discussed, emphasis is placed on giant strong component (GSC). At last, the GSC is decomposed based on simulated annealing algorithm, which tries to find the optimal community structure by maximizing the network modularity [16-17].

2 Materials and methods

2.1 Data Acquisition

We first obtained all metabolic reactions involved in metabolic network of *B. thuringiensis* from KEGG database [18]. Subsequently, all of the reactions are revised based a KEGG-based database developed by Ma and Zeng [19]: 1) confirmed the reversibility of every reaction; 2) excluded the current metabolites and small molecules such as ATP, ADP, NADH and H_2O etc, with the purpose of reflecting biologically meaningful transformations. At last, the metabolic network reconstructed is represented by so-called metabolite graph in which the nodes are metabolites and the links are reactions. For example, the irreversible reaction, C00064+C00026 → C00025 is represented by two directed arcs C00064 → C00025 and C00026 → C00025. The metabolic network of *B. thuringiensis* contains 830 nodes and 1132 links.

2.2 Clustering Coefficient and Modularity

One of the most important properties of nodes to form community structure might be clustering, which is the property that two nodes have a heightened probability of link if they were both connected to the same third node. This effect is quantified by the clustering coefficient C, defined as following [20]:

$$C_i \equiv \frac{2E_i}{k_i(k_i-1)} \qquad (1)$$

where E_i is the number of links between neighbours of node i, k_i is the degree of node i. The value of C_i is between 0 and 1. According to the definition, it is 1 on a fully connected graph and has typical values in the range of 0.1 to 0.5 in many real-world networks [14]. It is clear that the clustering coefficient of node i measures the extent of its neighbours to form a community.

Another important property related to community structure is modularity. For a presumptive partition of the nodes of a network into communities, the modularity M of this partition is defined as following [13, 21]:

$$M \equiv \sum_{s=1}^{r}\left[\frac{ls}{L}-\left(\frac{ds}{2L}\right)^2\right] \qquad (2)$$

where r is the number of communities, l_s is the number of links between nodes in communities, d_s is the sum of the degrees of the nodes in community s, and L is the total number of links in the network. It is suggested that maximization of the modularity function would yield the most accurate results for random networks and widely used for identification of communities [16-17].

2.3 Community Detection

Since Ma and Zeng proposed [22] the "bow tie" structure of metabolic networks, it is increasingly recognized as being a conserved property of complex networks [23-25], and the results suggest that this structure property is functional meaningful for metabolism, disease and the design principle of biological robustness.

Generally speaking, a network with the "bow tie" structure could be decomposed into four parts: giant strong component (GSC), substrate subset (S), product subset (P), and isolated subset (IS) [22]. The GSC is the biggest strongly connected components of a metabolic network and determined structure of the entire network

at a certain extent, and thus it is extracted and decomposed based on simulated annealing algorithm developed by Guimera and Amaral [16-17] herein.

Simulated annealing is a stochastic optimization technique that could find 'low cost' configuration without getting trapped in 'high cost' local minima. As mentioned above, the method based on simulated annealing tries to find the optimal community structure by maximizing the network modularity [16-17], and thus the cost is $C=-M$ herein, where M is the modularity defined in Eq.(2). At each temperature T, some random updates are performed and accepted, and there would be $n_i = fS^2$ nodes individual movements from one community to another, and $n_c = fS$ nodes collective movements, where S is the number of nodes in the network, and f is iteration factor with the recommended range of 0.1 to 1. At each certain temperature T, the system would be cooled down to $T'= cT$, where c is cooling factor with the recommended range of 0.990 to 0.999.

3 Results and Discussion

The whole metabolic network of *B. thuringiensis* is decomposed into four parts based the "bow tie" structure. We noted that the metabolites and reactions involved in the most interested part (i.e. the GSC) is clearly much less than the whole network (14.2% and 23.7%, respectively), and would be used to reduce the complexity of applying other pathway analysis methods such as extreme pathways and elementary modes [8-9], while most nodes in S, P and IS part are connected by some single link.

All of the 268 metabolic reactions in GSC are compared to KEGG pathways, and show that they are mainly concentrated on carbohydrate metabolism and amino acid metabolism (table 1). The reactions of carbohydrate metabolism accurately correspond to glycolysis, TCA cycle, pentose phosphate pathway, and partly correspond to pyruvate metabolism and butanoate metabolism. From the point of view of network topological, the results show that metabolites in carbohydrate metabolism (in particular glycolysis, TCA cycle and pentose phosphate pathway, i.e. the central metabolism) have the higher probability of much more links and stronger robustness in network, and thus might have higher attack tolerance despite external cues, genetic variation and stochastic noise. While reactions of amino acid metabolism are mainly concentrated on urea cycle and metabolism of amino groups, arginine and proline metabolism, and glycerophospholipid metabolism, these might reveal the nutrient requirement in *B. thuringiensis*.

Various of decomposed results of the giant strong component of *B. thuringiensis* metabolic network based on simulated annealing algorithm are obtained due to different iteration factor (f) and cooling factor (c), at last we chosen the best decomposed result after a number of computing. The result gives clearly partition and the modularity in the partition of the network is 0.752183. Then the decomposed result is reaffirmed by compared to KEGG pathways, i.e. most communities are mainly corresponding to one or two KEGG pathways (table 2). For instance, 11 of 15 within links in community 1 are corresponding to butanoate metabolism, 11 of 12 within links in community 4 are corresponding to Glycero phospholipid metabolism, etc.

Table 1 Reactions in GSC of *B. thuringiensis* metabolic network

Corresponding pathway in KEGG	No. of reactions	Percentage
Carbohydrate Metabolism	140	52.2%
Amino Acid Metabolism	84	31.3%
Energy Metabolism	24	9.0%
Lipid Metabolism	8	3.0%
Others	12	4.5%
Total	268	100%

4 Acknowledgements

The authors would like to thank Dr. Guimera R and Dr. Amaral LAN for providing us the software NetCarto; Dr. Ma HW and Dr. Zeng AP for providing us their database; and the anonymous reviewers for their helpful

comments on the manuscript. Support for this work is provided by fund from Jiangnan University Innovation Teams (JNIRT0702).

Table2 The decomposed results of GSC of *B.thuringiensis* metabolic network are compared to KEGG pathways

Community	Pathways in KEGG
1	butanoate metabolism
2	pyruvate metabolism
3	glycolysis, ppp, carbon fixation
4	glycerophospholipid metabolism
5	——
6	pyruvate metabolism, amino acid biosynthesis
7	arginine and proline metabolism, urea cycle and metabolism of amino groups
8	——
9	TCA cycle

References

[1] H. Kitano, "Computational systems biology", Nature, Vol. 420, No. 6912, 2002, pp.206~210

[2] H. Jeong, B. Tombor, R. Albert, et al, "The Large-scale Organization of Metabolic Networks", Nature, Vol. 407, No. 6804, 2000, pp.651~654

[3] U. Alon, "Biological Networks: the Tinkerer as an Engineer", Science, Vol. 301, No. 5641, 2003, pp.1866~1867

[4] A.L. Barabasi, and Oltvai ZN, "Network Biology: Understanding the Cell's Functional Organization", Nat. Rev. Genet., Vol. 5, No. 2, 2004, pp.101~113

[5] S. Boccaletti, V. Latora, Y. Moreno, et al, "Complex networks: Structure and dynamics", Physics Reports, Vol. 424, No. 4-5, 2006, pp.175~308

[6] D.J. Watts, and S.H. Strogatz, "Collective dynamics of 'small-world' networks", Nature, Vol. 393, No. 6684, 1998, pp.440~442

[7] A.L. Barabasi, and R. Albert, "Emergence of scaling in random networks", Science, Vol. 286, No.5439, 1999, pp.509~512

[8] C.H. Schilling, D. Letscher, and B.O. Palsson, "Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective", J. Theor. Biol., Vol. 203, No. 3, 2000, pp.229~248

[9] S. Schuster, D.A. Fell, and T. Dandekar, "A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks", Nat. Biotechnol., Vol. 18, No. 3, 2000, 326~332

[10] E. Ravasz, A.L. Somera, D.A. Mongru, et al, "Hierarchical organization of modularity in metabolic networks", Science, Vol. 297, No. 5586, 2002, pp.1551~1555

[11] A.W. Rives, and T. Galitski, "Modular organization of cellular networks", PNAS, Vol. 100, No. 3, 2003, pp.1128~1133

[12] J.A. Papin, J.L. Reed, and B.O. Palsson, "Hierarchical thinking in network biology: the unbiased modularization of biochemical networks", Trends Biochem. Sci., Vol. 29, No. 12, 2004, pp.641~647

[13] M.E.J. Newman, and M. Girvan, "Finding and evaluating community structure in networks", Phys. Rev. E, Vol. 69, No. 2, 2004, 026113

[14] M. Girvan, and M.E.J. Newman, "Community structure in social and biological networks", PNAS, Vol. 99, No. 12, 2002, pp.7821~7826

[15] P. Gleiser, and L. Danon. "Community structure in jazz", Advances in Complex Systems, Vol. 6, No. 4, 2003, pp.565~573

[16] R. Guimera, and L.A.N. Amaral, "Functional cartography of complex metabolic networks", Nature, Vol. 433, No. 7028, 2005, pp.895~900

[17] R. Guimera, and L.A.N. Amaral, "Cartography of complex networks: modules and universal roles", J. Stat. Mech.-Theory Exp., No. P02001, 2005, pp.1~13

[18] M. Kanehisa, and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes", Nucl. Acids Res., Vol. 28, No. 1, 2000, pp.27~30

[19] H.W. Ma, and A.P. Zeng, "Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms", Bioinformatics, Vol. 19, No. 2, 2003, pp.270~277

[20] M.E.J. Newman, S.H. Strogatz, and D.J. Watts, "Random graphs with arbitrary degree distributions and their applications", Phys. Rev. E, Vol. 64, No. 2, 2001, 026118

[21] R. Guimera, M. Sales-Pardo, and L.A.N. Amaral, "Modularity from fluctuations in random graphs and complex networks", Phys. Rev. E, Vol. 70, No. 2, 2004, 025101

[22] H.W. Ma, and A.P. Zeng, “The connectivity structure, giant strong component and centrality of metabolic networks”, Bioinformatics, Vol. 19, No. 11, 2003, pp.1423~1430

[23] M. Csete, and J. Doyle, “Bow ties, metabolism and disease”, Trends Biotechnol., Vol. 22, No. 9, 2004, pp.446~450

[24] H. Kitano, “Biological robustness”, Nat. Rev. Genet., Vol. 5, No. 11, 2004, pp.826~837

[25] J. Zhao, L. Tao, H. Yu, et al, “Bow-tie topological features of metabolic networks and the functional significance”, Chinese Science Bulletin, Vol. 52, No. 1, 2007, pp.47~54

Study of Sensor Management Based on DWPSO Algorithm

Dingguo Jiang Xiaoliu Zhu Baoguo Xu

School of Communication and Control Engineering, Jiangnan University, Wuxi, Jiangsu, 214122, China

Email:zhangyulin@hyit.edu.cn

Abstract

Sensor optimal management is very important to the performance of multi-sensor system in target assignment problem. To solve it effectively, main old algorithms for solving this problem are discussed first, based on Dynamic Weight Particle Swarm Optimization algorithm, a new method is put forward then. The algorithm simulation is given in the end; the result shows the new target assignment method is valid, especially for the problem of large scale target assignment the new method shows faster convergence rate and higher precision.

Keywords: sensor management; multi-sensor system; data fusion; efficiency function; Dynamic Weight Particle Swarm Optimization algorithm

1 Introduction

With the development of multi-sensor information fusion technology, the requests to the coordination and management between multi-sensors are further enhanced. As a feedback module, Sensor optimal management enables the multi-sensor data fusion system to constitute a closed-loop control system, thus enhanced system's timeliness and the whole optimization. [1-2]

At present, there are several methods on sensor management and dispatch. They are based on the plan (linear programming, dynamic programming and nonlinear programming), based on the information theory method, based on the fuzzy logic and the neural network method, based on the stochastic set theory method and based on expert system's method [3-6] and so on. Among them, the method based on the plan with the one based on the information theory are two research focuses.

The particle swarm optimization algorithm (PSO) [7-10] is a kind of global searching algorithm, which is based on community intelligence. It was proposed by Kennedy and E.berhart in 1995. For its prominent merits, the algorithm has been widely applied into numerous domains at present. The ability to parallel search and the merit on convenience processing constraint of PSO adapt completely the question of sensor optimal management.

2 DWPSO Algorithm

A standard particle swarm optimization algorithm (SPSO) is depends on the speed - position search model, in the supposition community the particle i in the D Uygur solution space position is $X_i = (x_{i1}, x_{i2}, ..., x_{id})$, the speed $V_i = (v_{i1}, v_{i2}, ..., v_{id})$, the current time's individual extreme value records is p_{ibest}, the overall situation extreme value records is g_{best}. In each time iterates, the particle track individual optimum value, the global optimum and the preceding time's condition adjusts the current time the position and the speed, the iterative formula is as follows:

$$v_{ij}(t+1) = wv_{ij}(t) + r_1c_1(p_{ij} - x_{ij}(t)) + r_2c_2(g_j - x_{ij}(t)) \quad (1)$$

$$x_{ij}(t+1) = x_{ij}(t) + v(t+1) \quad (2)$$

r_1 and r_2 are between [0, 1] random numbers, c_1 and c_2 are the acceleration factor, w is the inertia factor. In standard PSO algorithm, because which takes

the inertia factor as the preset parameter, and usually is smaller than 1, which has the obvious deficiency: The particle speed getting smaller, even stopping moving, which will have precocious restraining.

This paper proposed a kind of dynamic inertia factor DW(Dynamical Weight) which changes along with the particle evolution process the improvement strategy, enables the particle to have the auto-adapted change speed, can adjust the granule auto-adapted along with the evolution process the multiplicity, thus jumps out effectively is partially most superior, avoids restraining precociously. From the standard PSO algorithm, may see the granule optimization the process is the process which evolves unceasingly, may divide into it two parts: The particle velocity evolution degree and the granule polymerization's degree changes unceasingly process.

Defines 1: If in the t generation of population the particle i uses the expression $x_i(t)\,|\,(i \in \{1,2,...,N\})$, this particle current optimal-adaptive value may express is f_i, the current time's individual extreme value records is $f_{ibest}(t)$, then the $(t-1)$ generation of individual extreme value is $f_{ibest}(t-1)$, $\alpha(x)$ expressed particle velocity evolution degree:

$$\alpha(x) = f_{ibest}(t) / f_{ibest}(t-1) \qquad (3)$$

Defines 2: With defines 1, and f_{avg} expressed that the current grain of subgroup's average adaptation value, $\beta(x)$ is called the particle polymerization the degree:

$$\beta(x) = f_{ibest}(t) / f_{avg}(t) \qquad (4)$$

The defines 1 and defines 2 are clear reflection grain of subgroup's optimization process, the inertia factor size changes along with the particle velocity evolution and the particle extent of polymerization change: When $\alpha(x)$ is big the evolutionary rate is quick, the algorithm may continue in the big space to search, namely granule in wide range optimization; When $\alpha(x)$ is small, may reduce causes the particle to search in the small scope, thus found the optimum value quickly. When $\beta(x)$ is small, that is ,when the particle is quite scattered, the particle is not easy to fall into partially most superior, along with $\beta(x)$ increases the algorithm easily to fall into the partial optimum value, this time needs to increase w, thus increases the search the space, enhances a grain of subgroup overall situation optimization ability. In summary, w increases along with $\alpha(x)$ reducing, w increases along with $\beta(x)$ increasing, therefore w with $\alpha(x)$ between $\beta(x)$ functional relations may express the equation below:

$$w = f(\alpha,\beta) = w_0 - 0.5 * \alpha(x) + 0.1 * \beta(x) \qquad (5)$$

w_0 is w starting value, generally takes w_0=0.9; By, the definition $\alpha(x)$ and $\beta(x)$ may know $0 < \alpha(x) \le 1$ and $0 < \beta(x) \le 1$, therefore $w_0 - 0.5 < w < w_0 + 0.1$, has guaranteed $w < 1$ restraining request.

3 Application of DWPSO Algorithm in Sensor Management

The sensor management the most primary mission is basis certain most superior principle determines the goal, and then chooses sensor's type and sensor's search pattern and the operational parameter. Its goal is the use limited sensor resources, carries on the effective assignment under the most superior criterion to the sensor resources.

3.1 Introduction Comprehensive potency function

Firstly, the goal priority sorting function and the sensor (combination) with the goal pairing function are introduced, defines sensor *i* pair of goal *j* the potency function, thus extracts all sensors in the system to the complete goal comprehensive potency function, in addition the biggest track capacity restraint and to the goal cover restraint, may realize to the sensor resources optimized management.

Supposed has the goal which *m* basic sensor and needs to track is *n*. Because the identical time possibly has many sensors to assign for the same end, this time may synthesize several basic transducer units a sensor combination establishment to contain the basic sensor combination (literature [1] to be called a false sensor). Like this in assigns in the time, a goal “the sensor” (basic or false sensor) carries

on the observation on only then one, this time "sensor" the number increased *2m-1* from *m* Supposes sensor *i* pair of goal *j* the potency function is E_{ij}, it is definite by the sensor flight path data covariance. Simultaneously needs to consider that the sensor search and track capacity's limit, namely within certain amount of time cannot surpass the sensor own biggest track capacity. Makes the basic sensor *i* most to be possible to assign approves the goal, but each batch of goals must and can only assign for a sensor (basic or false sensor); $U = \{U_1, U_2, ..., U_{2m-1}\}$, U_i assign for the *i* sensor's goal set; *J(i)* contains the serial number constitution integer set which the basic sensor *i* all sensors combine. Then the goal optimization assignment problem's comprehensive potency function expression is:

$$\max E = \sum_{U_i \in U} \sum_{j \in J_i} E_{ij} \qquad (6)$$

$$\text{Constraints:} \quad \sum_{j \in J(i)} \left| U_j \right| \le l_i \qquad (i = 1, 2, ..., m)$$

$$\sum_{j \in U} \left| U_j \right| = n$$

$$U_i \bigcap_{i \ne j} U_j = \varnothing \qquad (i = 1, 2, ..., 2^m - 1;\ j = 1, 2, ..., 2^m - 1)$$

In the formula expression set element integer. the type (6) obtains the comprehensive potency function to each kind of sensor goal assignment combination, and selects the greatest achievement optimization assignment result. In constraint's first had guaranteed assigns the goal number which tracks for the sensor will not surpass each basic sensor's biggest track capacity, second and the third constraints guaranteed a goal to, and only could assign for a sensor (including false sensor).

3.2 Sensor management based on DWPSO algorithm

In this paper, the potency function is taken as sensor-goal pair target, which has realized the pair target automatic renewal, then has realized in the multi-sensor-multi-objective tracking system multi-sensor's automatic optimized management. Applies the sensor which and the goal pair question when the DWPSO algorithm manages in the sensor, must first make some suppositions: First, the sensor regards as the adaptation value which to the goal assignment's comprehensive potency function the granule searches; Second, the granule decomposes in the flight way choice into the goal to sensor's way selective rule and the goal to the goal way selective rule, the goal to the goal way choice is the random selection strategy; Third, the comprehensive potency function optimum value uses the overall situation renewal and the partial renewal principle, maintains the partial renewal the goal is suitable reduces the granule group flight the centralism too to cause the stop search. From this, in the sensor and the goal first in the known situation, optimizes the management based on the DWPSO algorithm's sensor the flow to be as follows:

1) Define the number of particle swarms, comprehensive potency function starting value, biggest permission iteration number of times, and initialization each granule position and speed;

2) Take the position vector of each particle as the synthesis potency function the partial optimum value, compute global optimum and the partial optimum value of all particles.

3) Compared current optimum value of a particle with its own best optimum value, if good, takes it as this granule current best position;

4) Each particle's own best optimum value compares with the grain of subgroup's current best optimum value, if good, takes it as a grain of subgroup the best position;

5) Once achieved the constraints or the biggest iterative number of times stops the search, this time grain of subgroup's most superior position namely for synthesis potency function maximum value, otherwise continues the step 6);

6) If the iteration number of particle swarms increases 1, according to (1) - (5) the formula renewal granule's speed and the position, continues to search.

4 Algorithm Simulation Example

This paper shows with a simple simulation example based on the DWPSO algorithm validity. The supposition tracking system (s_1, s_2, s_3) is composed of

three basic sensors, tracks the airborne ten surveillance goal ($t_1,...,t_{10}$), each sensor are most may monitor four goals. For simplicity between the supposition three basic sensors does not have the mutual influence, namely may combine willfully, and thus considers four false sensors (s_4, s_5, s_6, s_7), and $s_4 = \{s_1, s_2\}$, $s_5 = \{s_1, s_3\}$, $s_6 = \{s_2, s_3\}$, $s_7 = \{s_1, s_2, s_3\}$ 。Information increase as shown in Table 1 by the fore-mentioned algorithm.

Table 1 Forecast information increase

Sensors	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
s_1	1.946	1.740	2.015	2.542	2.347	1.738	1.609	2.141	1.638	2.169
s_2	1.135	0.636	0.638	1.273	1.478	1.852	2.378	3.001	2.351	1.376
s_3	2.478	1.779	2.003	3.137	1.739	1.485	0.984	1.249	1.622	1.241
s_4	3.081	2.377	2.651	3.817	3.787	2.742	2.728	3.648	2.939	2.829
s_5	3.494	2.743	3.032	4.255	2.082	3.583	2.445	3.719	2.578	2.545
s_6	2.948	2.137	2.361	3.629	3.546	2.139	2.861	3.232	2.354	2.364
s_7	3.967	3.104	3.392	4.748	4.472	3.003	3.249	4.749	3.262	3.187

After the experiment, the following parameters may established, the number of particle swarms is 10, the iterative total is 100, the suppressive condition is smaller than 0.000001 for the adaptation value and the optimal solution difference, its assignment result as shown in Table 2, 1 expressed that the sensor (group) assigns for this goal, 0 expressions have not assigned.

Table 2 Assignment result

Sensors	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
s_1	0	1	0	0	0	0	0	0	0	1
s_2	1	0	0	1	0	0	0	1	1	0
s_3	0	0	0	0	0	1	1	0	0	0
s_4	0	0	0	0	0	0	0	0	0	0
s_5	0	0	1	0	1	0	0	0	0	0
s_6	0	0	0	0	0	0	0	0	0	0
s_7	0	0	0	0	0	0	0	0	0	0

In Figure1, a 20-time average optimal-adaptive value change curve is shown, which a grain of subgroup searched, on the diagram of curves obvious has been possible to see after the granule underwent 30 iterative searches, probably obtained the synthesis potency function optimum value, had the quick convergence rate, and had the high restraining precision. To the computation result, the algorithm is always satisfying the limit in the optimization process the condition as far as possible to assign the goal to the sensor combination, thus saved the limited sensor resources. The above simulation experiment showed this algorithm's sensor management is effective, and has certain rationality.

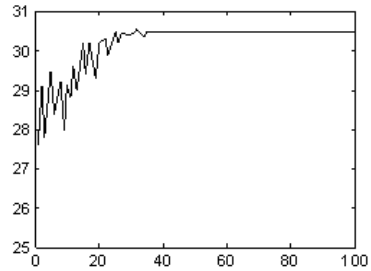


Figure1 Synthesis potency function 20 average optimum value evolution chart

5 Conclusions

In this paper, the potency function is taken as the most superior criterion, a method of sensor information optimal management is proposed based on the DWPSO algorithm. And is not easy using the improvement grain of subgroup algorithm's parallel search to fall into partially the most superior characteristic, enables this kind of optimized algorithm to have the rapid convergence and the restraining precision high merit. The simulation result demonstrates that this method is effective and reasonable.

References

[1] NASH J M. Optimal allocation of tracking resource [A]. Proceedings IEEE Conference on Decision and Control [C], 1177-1180

[2] HINTZ K J. A Measure of information gain attributable to cueing [A].IEEE Trans. On System, Man and Cybernetics[C] Vol. 21, March 1991, 434 - 441

[3] WAYNE S. Information based sensor management [A]. Signal Processing, sensor fusion, and target recognition II, Proceedings of the 1994, IEEE International Conference on Neural Networks[C], Vol. 5, Orlando, FL, June 27 - July 2, 1994, 3403- 3408

[4] LU D , ZENG Y, YAO Y. Sensor Management Based on Cross - entropy [A]. Instrumentation and Measurement Technology Conference, Proceedings of the 20th IEEE[C], 2003, Vol. 2. 1555 - 1557

[5] St romberg D. A platform - based data fusion and sensor management node [A]. RADAR 2002 [C]. 2002, 483 - 487

[6] Tubaishat M, Madria S. Sensor networks: an overview [J]. IEEE Potentials, 2003, 22 (2): 20 - 23

[7] Kennedy J, Eberhart R C. Particle swarm optimization [A]. Proc IEEE International Conference on Neural Networks[C]. USA,IEEE Press,1995，4. 1942-1948

[8] Eberhart R C, Kennedy, J. A new optimizer using particle swarm theory [A] Proc. of the sixth international Symposium on Micro Machine and Hunan Science [C] Nagoya .Japan: [s.n.], 1995，39-43

[9] R.Eberhart, Y.Shi. Particle Swarm optimization: delve- lop-ment, applications and resource [J].IEEE Int conf on evolutionary Computation, 2001 pp81-86

[10] F van den Bergh. An analysis of particle swarm optimizers [D]. PhD thesis. Department of computer science, University of Pretoria, South Africa, 2002.81:83

A Reliable and Efficient Communication Mechanism for Mobile Agents

Shengjun Xue Xianju Zhou

Department of Computer Science and Technology, Wuhan University of technology, Wuhan, Hubei, 430063, China

Email：chrysanthmu@126.com

Abstract

The communication mechanism is the foundation of Mobile Agent systems, but there are some problems such as reliability and efficiency exist in former communication mechanisms. In this essay, I bring forward a mobile agent communication mechanism AREM. It includes Mobile agents' creation, addressing and communication mechanism, I use AgentID to give Agents unique names, adopt name resolver to achieve transparent addressing, and make use of mailbox to ensure message cache and transmission. Based on the above features we can solve the problem of invalid communication, so as to achieve mobile agent system's reliability and efficiency.

Keywords：Communication; Mobile Agent; Addressing; Moving Process

1 Introduction

With the development of Internet, Mobile Agent which is derived from Internet and mobile computing application has attracted more and more attention, and has been applied to many areas, such as e-business, personal assistant, security proxy, workstream[5].

Communication is an indispensable part of mobile agents, both the cooperation between them and the control of them are based on communication. So a reliable communication mechanism is needed to ensure efficient communication between agents[7]. While, mobile agents have mobility and autonomy, the communication mechanism is different from normal communication in a large part. In the communicating process of mobile agents, there exist phenomena such as communication exception and message pursuit. The communication between mobile agents is an infrastructure of mobile agent system. How to communicate efficiently and reliably is a hot topic in Internet environment.

2 Current conditions and problems

In recent years, researchers have brought about a large number of communication mechanisms to solve the problem of Mobile agent communication[4]. Mole system has adopted the session-oriented mechanism, and addressing method like DNS, it discards invalid messages, at the same time, sends back error information and saves invalid messages. When target agent comes back, it delivers messages again. In fact, Mole system doesn't solved the problem of communication invalidation, and doesn't offer a reliable communication protocol.

Mogent[7] system has introduced the concept of state to agents, each node in Mogent doesn't only have a communicator to take charge of communication details, but also designs a Home, these two module record the agent states and mails messages respectively. Mogent takes charge of the two signals to make sure that sender only sends messages to static agent, and agent can move when there is no message sent to it. Mole system has settled the problem of communication invalidation, but it restricts agents' movement and messages sending, as a result, it affects agents' independence and movement.

The communication mechanisms above both have advantages and disadvantages, to solve the problem better, I put forward a reliable and effective

communication model AREM, in this model, I divide the whole internet into several parts based on domain, use Domain Name Resolver(shortened as DNR) and communication components to realize mobile agents communication; moreover, I import agent state, message overtime and priority to solve the problem of communication invalidation.

3 Description of communication in AREM model

As we can see in figure 1, the whole Internet has been divided into many Domains[3], each Domain has a DNR, which is responsible for the name service of agent, such as agent enrollment, cancellation, inquiry and security authentication. Mobile agent communication is mainly accomplished by DNR and Communicator.

3.1 Data structures in DNR

In this model, there are two tables in each DNR, Host Table(shortened as HT) and Visitor Table(shortened as VT), the data structures are as below:

HT:<AgentID，Domain，HNaddr，Service>AgentID is the only global symbol of mobile agent, Domain is the IP that domain agent belonged to, HNaddr is the Home address which is established by mobile agent, Service is the kind of service.

VT：<AgentID,Mailbox,State>Mailbox is the current mailbox address of agent, State is the current state.

3.2 Communication service component —— communicator

In each domain there are some physic nodes, each of which is configured a mobile Agent Server to manage the current moving agents. The most dominant system component which offers communication in Server is Communicator, its main function is to offer message sending and receiving between agents, at the same time, it registers to DNR, multiple Communicators can cooperate, it consists two parts:

Message pool: there is a message pool in each Domain, it is the transferring station of agents in Domains.Mailbox: it is the cache space for agents.

There is an Inner Mapped Table(shortened as IMP) in Communicator.

IMT:< AgentID,Mailbox,State>Mailbox is the mail address of Agent.

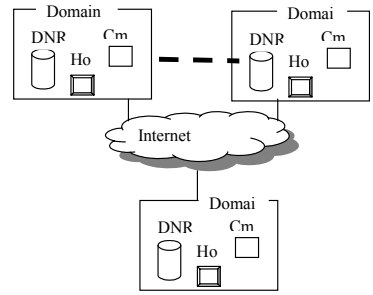


Figure 1 AREM communication model

4 The description of communication process in AREM model

4.1 The creation process of mobile agents

When mobile agent is created, the system allocates a unique AgentID to it according to the naming regulation. Communicator allocates a Mailbox to it, and stores information in the IMT, at the same time, Communicator sends a logging reply to the DNR inside its own domain, stores agent's information in HT, and IMT in VT, what's more, it sets its state as static.

4.2 The moving process of mobile agents

It is distributed into two kinds:

First one is inner-domain movement, as we can see in figure2-----movement1. When mobile agent A moves from Server1 to Server 2, the records of A in HT are not changed, and only the Mailbox address in VT is changed, so we just have to change the correspondence between AgentID and Mailbox in VT.

Second one is inter-domain movement, as we can see in figure2-----movement2. Server3 accepts this Agent successfully, informs DNR2 to enroll for the Agent. DNR2 adds A's record to VT; at the same time,

DNR2 makes requests to DNR1, in order to update A's Domain parameters. Besides, DNR2 makes requests to DNR1, deletes the enrolling information in VT.

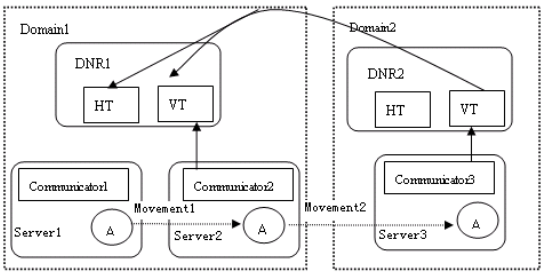


Figure2 Moving process of mobile agents

4.3 The addressing and communication functions of Mobile agents

In mobile agent system, if A wants to communicate with B, A and B are mobile agents, the basic communication process is as figure3:

(1) Sending AgentID of B and the communication messages to Communicator.

(2) Communicator checks that whether B is in the same Domain, if so, puts the message in its mailbox, so as to implement effective local communication;

(3) else, sending part resolves B's host Domain according to AgentID, then the DNR in host Domain finds out B's current Domain according to HT, and makes a request to current Domain, so as to find out B in current Domain's VT.

(4) If it finds out, returns B's mailbox address, Communicator sends messages to this address, and message is captured by the destination Communicator and puts in B's mailbox, so as to complete a communication process.

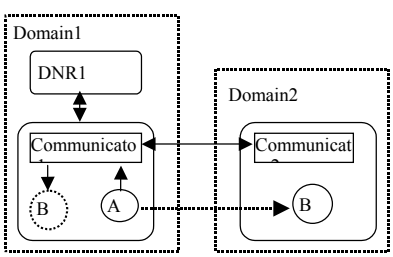


Figure3 Communication of mobile agents

4.4 Solution to invalid communication problems

In this communication model, if A is sending messages to B, message is sent to B's cache components---Mailbox, Normally, B will take message inside the Mailbox by turn and deal with them. But when B is moving during the process, Mailbox will also receive the messages. If the moving process is failed, B will use the former Mailbox as a message cache components; if the moving process is successful, although the records of B in IMP before moving will be deleted, the old Mailbox will not be cancelled until it sends the messages to B's new Mailbox by turn. As the messages are treated by turn, we can solve the problem of message pursuit.

4.5 The improvement on message

In this model, I add two parameters------- timeout and message priority. Timeout is timeout period set by message sender when establishing messages. Priority is used to record priority index.

As for the sender, when the sender sends message to the receiver, it sends message timeout and priority and destination AgentID to Communicator. Then the sender waits for answer, if the time is as long as the Timeout and there is still no answer, sender will withdrawal from this delayed event.

As for the receiver, when the receiver's Communicator receives messages, it will count the priorities according to message priority index, and then insert the message to destination Agent's message queue in Mailbox. Agent takes the messages out one by one from high priority to low, judges that whether the message is timeout, if so, it does nothing with it, else it makes corresponding answer. Besides, before the old Mailbox sends non-processed message to the new one, it has to judge that whether the message is overtime and delete the timeout messages, in order to avoid nonsensical internet transmission.

5 Analysis on performance

5.1 Analysis on reliability

When message sender wants to communicate with another Agent, it must know exactly its name, that is AgentID, and send AgentID with Timeout and Priority to host's Communicator. We can make messages sent to destinations correctly. Timeout and Priority avoid frequent moving and long waitness. This assures that messages can be sent based on reliability.

5.2 Analysis on communication speed

In this model, I separate addressing and message transfer, I use a high performance host as DNR server, and offer some reserve DNR servers, so that efficiency and reliability are maintained. Message is sent based on addressing, and it only has to transfer between sender and receiver's Communicator for one time, so it is efficient.

5.3 Analysis on stability and fault tolerance

In this model, nodes are not controlled by centralized manner, so system operation is not focused on a few communicator servers, but distributed equally to all the hosts attached to this communication. Besides, there are DNRs in Domains to reduce the burden on Domain name servers. So, if DNR or host goes wrong in some conditions, the system can also run regularly to ensure the system's stability and fault tolerance.

6 Conclusion

Addressing and reliable message transferring are two key problems in mobile agents communication. In this essay, I brought forward a new communication mechanism AREM, in this model, I used unique AgentID to name agents and gave a resolver based on it, in order to solve the problem of transparent addressing, besides this, I used message cache and transmission based on mailbox to deal with messages. What's more, I adopted timeout and priority to settle the problem of invalid communication. This mechanism is a best solution to current communication in Internet environments.

References

[1] FENG Xin-yu, LU Jian, CAO Jian-nong, "Design of a Generic Framework for Mobile Agent Communica tion"Journal of Software, Vol.14, No.5,2003，pp.984-990

[2] LI Hui, WANG Ber-zhan, LI Tao, YANG Zhan-hua, "An Improved Active Communication Algorithm of Mobile Agent," The Research on Computer Science , No 11,200, pp.200-203

[3] LI Tao, LI Hui, GU Jian-hua, PAN Hui-fang, "A Mobile Agent Communication Mechanism in an Internet Environment," Micro electronics and Computer , Vol 22 No.7,2005, pp. 42-45

[4] ZHAO Xiu-mei, MA Hong-wei, DU Xiang-hua, LIU Nan, "AN IMPROVED MOBILE AGENT COORDINATION MODEL," Computer Applications and Software, Vol24,No3, Mar 2007, pp.108-110

[5] GENG Zheng, "Research On Communicaton of Mobile Agents," An Hui University, China, April 2007

[6] ZHANG Li-hong, "A Reserch on an Improved Communication Mechanism of Mobile Agent," Lan Zhou University, May 2007

[7] WANG Ji-zeng, MAN Zi-bin, ZHOU Jun, "Informaiton-table-based scheme for mobile Agent communication," Computer Engineering and Design,China, Vol.28 No.11, June 2007, pp. 2566-256

The Automation Tester of Toy Flammability Based on PLC

Xiaoguang Xu[1] Lijuan Yin[2]

Shenzhen Entry-Exit Inspection and Quarantine Bureau, Shenzhen, Guangdong, 518045, China

Email：1 xu.xg@163.com; 2 yinlj411@163.com

Abstract

Based on the requirements of ascertaining the orientation and dimension of toy major axis, calculating the flaming dimension and speed along with the toy major axis in American standard consumer safety specification for toy safety, combined with PLC control technology, high-performance and fine-segmentation drivers as well as step-in motor working principle, it demonstrated the automation tester working principle, construction of mechanism and hardware, programming of software, function and technology characteristics. It is pointed out the tester can settle for testing requirements of American standard consumer safety specification for toy safety and improve the accuracy and efficiency on toy flammability test.

Keywords：Automation Control; PLC; Flammability Testing; Toy Safety Test

1 Introduction

In American standard consumer safety specification for toy safety, toy major axis is defined as: a straight line through the longest dimension of the product connecting the most distant parts or ends of the product. A product can have more than one major axis, but they must be equal in length. For example, in Fig.1, the lines A-A,B-B are major axes, but C-C is not major axis[1]. For toy flammability of plastic toys and soft toys, we often have to spend a great deal of time to ascertain the orientation and dimension of toy major axis. At present, we ascertain the orientation and dimension of toy major axis as well as the flaming dimension by eyes and Measurement ruler. As the shape of toy has been changed after burning, we can not accurately measure the flaming dimension by eyes, artificial factors will affect the test results. The test will be inefficient, inaccurate and poor repeatability, so it is hard to determine the result of toy flammability testing whether passes or not.

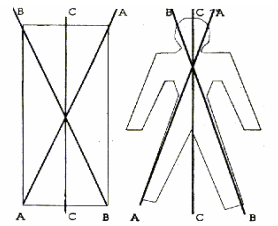


Figure1 Major Axis of Toy

With the development of computer-controlled technology, PLC (programmable logic controller) has characteristics of modular structure, high anti-jamming I/O processing components, flexible hardware configuration, expansible and stability, which provides a stable platform in different application. It has been widely used in the field of automation control device [2][3].

The automation tester of toy flammability uses the advanced PLC control technology, high-performance and fine-segmentation driver control technology. The tester which automatically measure the toy major axis, record the flaming time and calculate the flaming dimension can meet with the related clause about flammability testing requirement of American standard toy safety ASTM F963-03. The tester which has high precision, high efficiency and good stability provides guarantee for the test of toy safety.

2 Working process

The automation tester of toy flammability is platform configuration. The tester is composed of four

parts which are human-computer interaction part, control part, drive part and sustain orientation part. It's shown in Figure2.

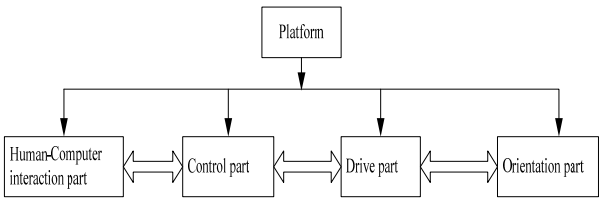


Figure2 Compose Structure

Human-Computer interaction part: Use the F920 operation panel of Japan's Mitsubishi; use the operational button to set the various operational functions. Test parameters and test results are displayed on the screen. The operations are simple and convenient. The results are displayed timely and correctly.

Control part: Controller with high precision, running quickly and reliable stability characteristic uses the FX1S-10MT PLC of Japan's Mitsubishi. Its primary function is to accept input signals, determine and process data in according with signals, export output control signals to the driver of the tester. It also has calculation function.

Drive part: The tester has three step-in motor which uses advanced JQF-MD808 step-in motor drive of the American company WJT. They accept pulse output signals, control revolving angle and orientation, and then finish various drive tasks in the course of the test.

Sustain orientation part: Mechanical parts are formed with base, knighthead, orientation rod and calibration mark. The automation tester of toy flammability mechanical structure is shown in Figure2.

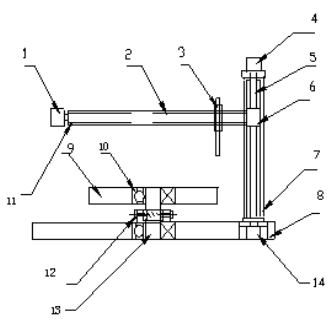


Figure2 Mechanical Structure

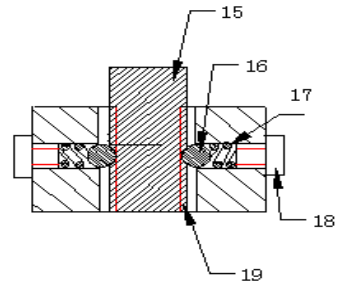


Figure3 Rotate Table Mechanical Structure

1.Motor1 2.Screw pole 3.Calibration mark 4.Motor2 5.Screw pole 6. Active block 7. Knighthead 8.Base 9. Fine-adjusting platform 10.Bearing 11.Orientation rod 12. Bushing 13.Rotary axis 14.Motor3 15. Rotary axis 16. spherule 17. compress spring 18. blot 19. alveolus of gear

Working process:

Ascertain the orientation of toy major axis: Put the toy sample on the rotary table. Step-in motor drives the knighthead with the orientation rod which is insured the same direction of the orientation of toy major axis. To rotate the fine-adjusting rotary table can set the direction and position of orientation rod, and then ascertain the orientation of toy major axis.

Accurately measure the major axis dimension and the flaming dimension of toy: Use the step-in motor to drive the calibration mark on the screw pole moving from one side of the toy major axis to another. The moving distance is the dimension of toy major axis. Move the orientation rod to home position and inflame the toy. When the toy extinguished, move the orientation rod back to the pre-initial position. At this time, the orientation rod and the orientation of toy major axis are with the same direction. This position is ensured by step-in motor which is controlled by PLC. Move the calibration mark from one side of the toy major axis to another burned edge and record the moving distance, so the flaming dimension in the orientation of toy major axis is the margin between two different moving distances.

The display of test result and printing device: When the toy is inflamed, press time button. Stop the timer immediately after the flame is extinguished. PLC automatically records the flaming time, flaming rate =flaming dimension/flaming time. Use the data to calculate the flaming rate and print the dimension of toy

major axis, flaming dimension, flaming time, and flaming rate and so on.

The rotary table structure is show in Fig.3,the working principle is: when adjust the orientation of the major axis of toy, impact the bolt in the axes, the bolt puss the spring to impact the spherule. the spherule prevent the table rotate optionally. When rotate the tabel on force, the spring is reverse compressed, the spherule loosen,this process ensure the axis of table rotate placidly.when the axis rotate one alveolus of gear, the spherule compress the axis over again.This ensure the table rotate little angle in one time,improve the accuracy of ratate orientation.

3 Working principle

The kernel of the automation tester of toy flammability is the PLC and the three step-in motors driven part. Working principle of the tester is shown in Figure4.

According to flammability testing requirement of American standard toy safety, we can set the rise height of the orientation rod by button on the control panel. PLC accepts all information, sends out startup signal after analyzing and processing corresponding data, and starts the step-in motor. Step-in motor set the rise height to finish manipulation after receive the output signal from PLC. Consider the CCW rotate as input signal for PLC, and then control the step-in motor to bring the orientation rod which ascertain angle on toy's major axis direction. PLC records the angle as norm angle for continuing manipulation automatically. After PLC received this input signal, the step-in motor controls the orientation rod with the calibration mark to measure toy major axis L1. PLC records the data and show on the panel. Ascertain the dimension of toy major axis, press Home button to make the orientation rod back to original position.

After ascertaining the dimension of toy major axis, inflame the toy and press Time button at the same time, PLC accepts this signal as a digital input signal. Stop time counter immediately after the flame is extinguished; PLC accepts this signal as another digital input signal, and output the signal to the faceplate so that it can show the flaming time T. Press the Rotate CCW button, so that PLC can control the step-in motor to drive the orientation rod back to the rotary angle of original record, and that the orientation rod can coincide with the orientation of toy major axis. Press the Back button to make sure the calibration mark reach the edge which had burned and the faceplate will show the retrograde distance L2, then flaming dimension L= major axis L1 – retrograde distance L2, and show the flaming dimension on the faceplate, PLC will calculate the flaming rate according to the flaming dimension and the flaming time. At last press the print button, PLC will output the control signal to the micro printer, automatic print the results.

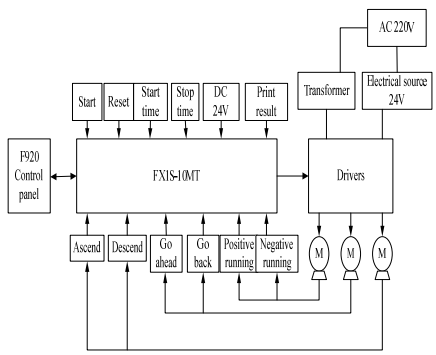


Figure4 Working Principle

At the same time, according to height of different toy, we can set height on the panel to make the orientation rod rise to certain height, which is convenient for the index of calibration mark reach the edge of toy major axis. We also can set the definite angle. First let the orientation rod rotate to a definite angle, then place toy on the rotary table, make the orientation of toy major axis the same direction of the orientation rod. When deviation appears, adjust them consistent by the fine-adjusting function.

4 Function characteristic

High testing precision. FXIS-10MT type of PLC is a 12 digit machine with high precision, fast speed

characteristic which makes tester has excellent testing precision and response time. Because of MC-808MDE high-performance and fine-segmentation step-in motor drive machine using new type bipolar crosscurrent carrier wave drive technology which has 256 times fine-segmentation make the step-in motor achieve a higher speed and torque, fine-segmentation function can provide motor operate higher precision, less shaking and lower noise.

Easy operation, use conveniently. Just need to operate keys on the operate faceplate. Screen display correlative rotating angle and measure dimension clearly, briefness, convenience, precision. The using of the tester can greatly reduce work intensity, increase work efficiency.

High reliability, great stability. Because of the characteristics of PLC with high reliability and great stability, MC-808MDE high-performance and fine-segmentation step-in motor drive machine has advanced over current protection (peak over 10A), over voltage protection (more than 85VDC), over temperature protection (≥ 70 ℃ stop working, ≤ 50 ℃ resume work) and fault protection function which make the tester operate more reliable, more secure, has a good electrical stability and reliability [4].

Rational structure design. The tester is platform structure, the operating faceplate installed in the chassis, the step-in motor which is packed in the knighthead and bearing inside is difficult to contact. At the same time the various parts of the tester use antirust treatment and are processed reasonably, so debugging and maintenance are convenient.

Flexible provisioning setting function. The tester which can preset testing speed and time has a wide range of settings. The default angle is 0-180 °. The range of default rate is 0-600mm / s. The burning time is 0-90 s.

Strong anti-interference. As PLC components of the tester have photo-electricity coupling function to digital signal, it can filter out the wrong action signal. MC-808MDE high-performance and fine-segmentation step-in motor drive machine which has photo-electricity coupling function to input signal, input signal TTL compatibility, differential signal acceptability, good heat dissipation and fine-segmentation function can restrain the vibrant interference, the electromagnetic interference and the environmental interference very well[5-10].

5 Software design of control system

The I/O variables of the automation tester of toy flammability make up of digital input signals, digital output signals and intermediate variables. Digital input signals: running signal, recover signal, start timing signal, stop timing signal, up signal, down signal, forward signal, back signal, clockwise rotate signal and counter-clockwise rotate signal; digital output signals: running control signal, print control signal; intermediate variables: height and angle setting.

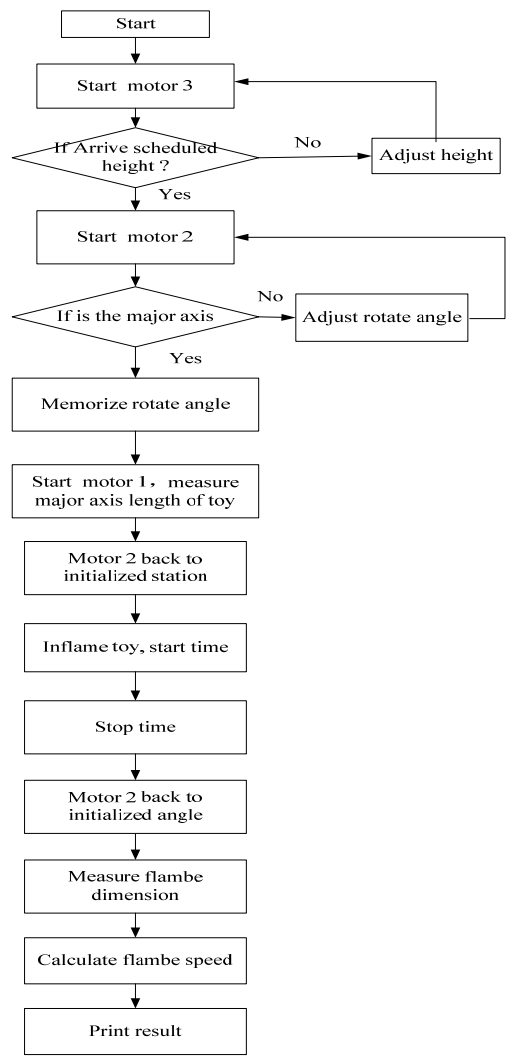


Figure5 Control Program

According to the received of digital input signals and intermediate variable, PLC start and control the tester running. The control program is shown in Fig.5. First ascertain the orientation of toy major axis, second test the dimension of toy major axis. After that adjust the orientation rod back to the pre-initial position, inflame the toy and start the timer. Stop the timer immediately after the flame is extinguished, and then adjust the orientation rod back to the record position, test the flaming dimension. Calculate the flaming rate and print them at last, and then the test is finished.

6 Conclusions

The automation tester of toy flammability which uses the FX1S-10MT PLC controller of Japan's Mitsubishi and advanced JQF-MD808 step-in motor drive of the American company WJT ensure stable operation, fast response and high accuracy. Use this tester can not only reduce the labor intensity of the work, but also improve the test efficiency and accuracy, advance the level of automation in the test work as well. This tester has a broad application prospect in the test of toy safety.

References

[1] Standard Consumer Safety Specification for Toy Safety ASTM F963

[2] Yu Hanqi, Electric control and PLC application technology, Nanjing, Southeast University Pub., 2003

[3] Zhao Yuehua, PLC technology and application, Chengdu, University of Electronic Science and Technology Pub., 1998

[4] Shen Shibing, "Communication between Mitsubishi PLC and PC", Journal of Control & Automation, April 2006, pp.81-83

[5] Shuqing Wang, Advanced control technology, Beijing Chemistry Industry Pub., 2001

[6] Wu Zhongjun and Huang Yonghong, PLC principle and application, Beijing, Machinery Industry Pub., 2005

[7] Qi Rong, New PLC tutorial, Northwestern Polytechnical University Pub., 2000

[8] Liao Changchu, S7-300/400 PLC application [M], Beijing, Machinery Industry Pub., 2005

[9] Yu Qingquang, PLC principle and system design, Beijing, Tsinghua University Pub., 2004

[10] Chen Boshi, Electric power automation system, Beijing, Machinery Industry Pub., 1991

A Flexible Authorization Delegation Method in Multi-domain Environments Employing RBAC Policies

Junguo Liao Feng Yang Huifu Zhang Gengming Zhu Bin Zhu

School of Computer Science & Technology, Hunan University of Science & Technology, Xiangtan, 411201, China

Email：liaojunguo@gmail.com

Abstract

In recent years, many researchers have noted that it will helpful to extend RBAC model to dynamic coalition environments. In this paper, we address the authorization issue in multi-domain environments where RBAC policies are employed. A flexible authorization delegation model is proposed, which combines RBAC model with delegation. The proposed model has powerful expressiveness, which supports these features as follows: multi-types of delegation, fine-grained delegation, temporal constraint, control on depth of delegation. In addition, this paper discusses other issues related with the proposed model, such as certificate storage, compliance checking and certificate revocation.

Keywords：Authorization; RBAC; Delegation; Compliance checking

1 Introduction

Access control is an important security issue in large organization such as commercial companies, hospitals, government organizations, and colleges. Role-based access control [1] has received considerable attention as an established alternative to traditional discretionary and mandatory access control for large organization. With the development of computer and network technology, it is certain that multiple organizations work together to achieve a common goal. In such a situation, the entities must cooperate to share the subset of their protected resources that is necessary to the coalition, while protecting the resources that they don't want to share. The growth of network-based services on the Internet promises to make this paradigm pervasive. In dynamic coalition environments, access control presents a number of challenges: the transitivity of authorization, the unpredictability of some entities' identity, and so on. Traditional role-based access control cannot address these issues. Delegation of authorization is an efficient approach to address these issues. Delegation means that a person gives all or part of his authority to somebody. In this paper, we propose a new flexible authorization delegation model, which provides secure interaction among multi-domains that utilize RBAC. The proposed model supports multi-types of delegation, role and permission level delegation, integer control on depth of delegation, temporal constraints, and multi-option revocation.

The rest of this paper is organized as follows. In section 2, we introduce the previous related works. Section 3 defines some terms in the model proposed in section 4. Section 4 presents a new authorization delegation model in dynamic coalition environments employing RBAC policies. Section 5 discusses the issue on delegation revocation. Conclusions and future work are presented in section 6.

2 Related works

Because RBAC model is convenient to implement security policies according to the structure of an organization, it is widely used in many large organizations. RBDM0 [2,3] is the first attempt to model delegation involving user-to-user based on roles. It is a simple one-step role-based delegation model with

total delegation. Revocation in RBDM0 is done either by an expiration mechanism or by any member of the same role as the grantor. Some extensions are discussed: grant-dependent revocation, delegatable and non-delegatable permissions, and two-step delegation, as well as delegation in hierarchical roles. RDM2000 [4,5] is an extension of RBDM0, which supports regular role delegation in role hierarchy and multi-step delegation. It uses can_delegate condition to restrict the scope of delegation. The unit of delegation in RBDM0 or RDM2000 is "role".

PBDM [6] supports flexible user-to-user and role-to-role delegation with role and permission level. A delegator can delegate his/her entire or partial permissions to others by using it. PBDM supports flexible delegation by separating delegation role from regular role and delegatable role and by separating temporal permissions delegated from other roles and its original delegatable permissions. RBDM0 and RDM2000 can be interpreted as special case of PBDM. But PBDM does not support constraints.

HyungHyo Lee et al [7] use sub-role hierarchies to support various delegation types and restricted inheritance. Dong-Gue Park et al [8] propose a delegation model using characteristics of permissions, in which security administrator can easily perform partial delegation, permission level delegation and restricted inheritance. It divides a role into sub-roles according to characteristic of permissions assigned to the role and considers delegation and inheritance simultaneously. Jacques Wainer et al [9] proposes a fine-grained, controllable, user-to-user delegation, which provide a rich set of controls regarding further delegation of permissions, generic constraints that further control delegation, and an innovative model for revocations. A fine-grained role-based delegation in presence of the hybrid role hierarchy is proposed [10], which is more expressive than delegation model in presence of the general hierarchy type.

These delegation models mentioned above are suitable to a single administration domain. Trust management (TM) is an approach to access control in multi-domain environments where users that are not in the same security domain need to share resources. Several TM systems have been proposed in recent years, e.g., PolicyMaker [11], KeyNote [12], SPKI/SDSI [13], and RT framework [14]. However, these TM systems can not support RBAC efficiently.

We propose a new authorization delegation model that combines the advantages of RBAC and TM. It is suitable for authorization in dynamic coalition environments where each security domain utilizes RBAC.

3 Preliminary Terminology

In this section, we define some terms in the authorization delegation model proposed in next section.

Administrative Domain A collection of hosts and routers, and the interconnecting network(s), managed by a single administrative authority. Administrative domain sometimes is also called security domain. In dynamic coalition environments, an organization often is an administrative domain.

Entity Each entity is mapped to a public/private key pair so that he (or she) can be uniquely identified. An entity represents a user or an administrator in an administrator domain.

Permission A permission means an ability to do some action or to access certain resource.

Role A role represents a set of permissions, which is often corresponding to a function or a job in an organization.

Privilege A privilege is either a permission or a role.

Delegation An entity gives some permissions to an object (e. g. another entity or a role). The entity is called delegator, and the object is called delegatee.

Delegation Chain Privilege can be passed from entity to entity in a transitive fashion. An entity, who has been granted privilege pr, may be able to further delegate privilege pr to others. The transitive passing of privilege from entity to entity can be imaged as a delegation chain.

Certificate Certificate is the representation of

delegation, which contains all information about delegation, such as the identities of delegator and delegatee, the delegated privilege, the expired datetime, and the depth of delegation. The certificate is signed by the delegator.

Expired Datetime Expired datetime means that a certificate is valid only before the expired datetime in the certificate.

Depth of Delegation Depth of delegation represents the number of steps in a chain where the privilege in the certificate can be further delegated.

4 Authorization delegation model

4.1 Syntax and Semantics

In order to formally describe our proposed authorization delegation model, let *D.p* and *D.r* denote a permission *p* and a role *r* in security domain *D*, respectively. Let *D.u* and *D.Adm* denote a user *u* and an administrator *Adm* in security domain *D*, respectively. A delegation can be represented as follows.

<delegator, delegatee, privilege, expireddatetime, depth>

delegator represents the grantor of a delegation, who is an entity such as a user or an administrator in a security domain.

delegatee represents the receipt of a delegation, which is a role or a user in a security domain.

privilege represents a permission or a role in a security domain.

expireddatetime represents the period of validity of the delegation.

depth represents the number of steps in a delegation chain where *privilege* can be further delegated.

The semantics of the delegation *<delegator, delegatee, privilege, expireddatetime, depth>* is that *delegator* gives *privilege* to *delegatee*, the delegation is valid before *expireddatetime*, and *privilege* can be further delegated within *depth* steps.

For example, the administrator *Adm* of security domain *A* wish to delegate the permission p_1 from security domain *A* to the user *Bob* of security domain *B* before *01/31/2008/12:00:00*, and the user *Bob* of security domain *B* can further delegate the permission p_1 from security domain *A* within *5* steps in delegation chain. The delegation for the example is *<A.Adm, B.Bob, A.p1, 01/31/2008/12:00:00, 5>*, and the delegation is signed by *A.Adm* using his/her private key.

Definition 1. The strong relation $\geq$ among privileges is defined as follows:

If $D.x = D.y$, then $D.x \geq D.y$.

If $D.r_1$ and $D.r_2$ are two roles from same security domain, and *imply*($D.r_1$, $D.r_2$), where *imply* is the RBAC relation between two roles, which means $D.r_1$ is senior to $D.r_2$, then $D.r_1 \geq D.r_2$.

If $D.p$ and $D.r$ are a permission and a role respectively, and $D.p$ is a permission member of $D.r$, then $D.r \geq D.p$.

Definition 2. If one of the following conditions is true, a delegation *<delegator, delegatee, privilege, expireddatetime, depth>* is legal.

delegator is an administrator of security domain D, and privilege is any permission or role from security domain D.

delegator is a user member of a role, and privilege is the role or any permission belonged to the role.

Exist another delegation <delegator', delegatee', privilege', expireddatetime', depth'> is legal, and depth' $\geq$ depth, and privilege' $\geq$ privilege, and expireddatetime is before expireddatetime', and delegator is delegatee' when delegatee' is an entity or delegator is an administrator or a user member of delegatee' when delegatee' is a role.

4.2 Certificate Storage and Collection

In our proposed authorization delegation model, delegation is represented as certificate. Because the model is applied to dynamic coalition environments, how to store and collect certificates in distributed environments is a very important issue. We address the issue in the section.

Most prior work that addresses the problem [11,12] assumes that all potentially relevant credentials are

available in one central storage. There are some exceptions. QCM [16] and SD3 [17] are two trust management systems that consider distributed storage of credentials. A limitation of the approach in QCM and SD3 is assuming that issuers initially store all the credentials, which may be impractical for some applications. This limitation was addressed by Li et al [15], who presented goal-directed credential chain discovery algorithms that support a more flexible distributed storage scheme in which credentials may be stored by their issuer, their recipient (also called their "subject"), or both. The algorithms dynamically search for relevant credentials from remote servers to build a proof of authorization.

In our proposed model, each security domain has a server to store certificates and determine whether the certificates submitted by an entity prove that his request is authorized. A delegation certificate will be stored in the server of the security domain which the receipt of the certificate belongs to.

For example, there are some delegation certificates as follows: C_1=<A.Adm, B.Bob, A.p_1, 01/31/2008/12:00:00, 5>, C_2=<A.Adm, B.r_2, A.r_1, 12/15/2007/12:00:00, 3>, C_3=<B.Bob, D.Jack, A.p_1, 12/15/2007/12:00:00, 3>, C_4=<D.Jack, E.Smith, A.p_1 12/10/2007/12:00:00, 1>. So, the certificates C_1 and C_2 will be stored in the server of security domain B, the certificates C_3 and C_4 will be stored in the servers of security domain D and E respectively.

When an entity requests to access some resource in other security domain, he/she must retrieve and collect all potentially relevant certificates from local and remote servers to prove that he/she is authorized to access the resource. The approach of retrieving and collecting relevant certificates is as the backward search algorithm in distributed credential chain discovery [15].

4.3 Compliance Checking

Compliance checking is to address the problem of determining whether collected certificates prove that an entity's request is authorized. In our proposed model the approach of retrieving and collecting relevant certificates is the backward search algorithm, so compliance checking is the problem as follows: given an entity D.u and a set of certificates C, solve all permissions that belong to D.u, denoted as SP(D.u), and determine whether the requested permission P belong to SP(D.u), denoted as $P \overset{?}{\in} SP(D.u)$.

Define 3. If an entity D.u is assigned to a role D.r through RBAC, then D.r is called as affiliated role of D.u. If a role E.r' is granted to an entity D.u through delegation, then E.r' is called as delegated role of D.u. If a permission E.p is granted to an entity D.u through delegation, then E.p is called as delegated permission of D.u.

Compliance checking is done by the server of requested security domain. The process of compliance checking is as follows:

Check the validity of each certificate in C by verifying the signature of each certificate and checking the expired datetime and revocation of each certificate. Delete the invalid certificates in C

Search all certificate chains by backward search algorithm as in [15], meanwhile check the legality of each certificate, and then solve affiliated roles, delegated roles and delegated permissions of requester.

Judge whether the requested permission belongs to the set of affiliated roles, delegated roles and delegated permissions of requester. If true, then the requested permission is authorized; otherwise the requested permission is not authorized.

Example 1. D.John wishes to access a certain resource in security domain A. The request requires that D.John should have the permission A.p2. D.John retrieves and collects relevant delegation certificates C = {C1, C2, C3, C4, C5, C6, C7}, where C1=<A.Adm, B.r1, A.p1, 01/31/2008/12:00:00, 4>, C2=<A.Ailce, C.r1, A.r1, 10/31/2007/12:00:00, 2>, C3=<B.Bob, C.r2, A.p1, 01/31/2008/12:00:00, 3>, C4=<C.Jack, C.Smith, A.p1, 01/15/2008/12:00:00,2>, C5=< C.Smith, D.John, A.p1, 12/31/2007/00:00:00, 0>, C6=< C.Adm, D.John, A.r1, 10/01/2007/12:00:00, 0>, C7=< C.Adm, D.John, C.r3, 01/31/2008/12:00:00, 0>. B.Bob and C.Jack are user members of B.r1 and C.r2 respectively in RBAC. Does C prove that D.John has the permission A.p2?

Suppose each certificate in C is valid and legal, there are three certificate chains: C7, C6←C2, C5←C4←C3←C1. So C.r3 and A.r1 are delegated roles of D.John, and A.p1 is delegated permission of D.John. If A.p2 is a permission member of A.r1 in RBAC, then C can prove that D.John has the permission A.p2; otherwise C cannot prove that D.John has the permission A.p2.

5 Revocation

Revocation is process by which a delegation certificate is removed and retracted when the certificate is also in period of validity. However, since delegation certificates may be chained, a revocation can produce side effects and other consequence. This section will examine the details related to revocation.

In the following cases, revocation is valid.

The administrator of security domain D can revoke any delegation certificates issued by any entity in security domain D.

Any delegator can revoke any delegation certificates issued by him/her.

If Cn←…←C2←C1 is a legal and valid certificate chain, then the delegator of Ci can revoke any certificate Cj where $j \geq i$.

The information about revocation must be signed by the entity that issues revocation. The server of each security domain D maintains a certificate revocation list, which contains all information about certificates revoked by entities in security domain D.

In example 1, suppose C3 is revoked by B.Adm who is the administrator of security domain B, the information about this revocation is signed by B.Adm and keeped in the server of security domain B. Because C3 is revoked, the certificate chain C5←C4←C3←C1 in example 1 is valid. So A.p1 is not delegated permission of D.John.

6 Conclusion and future work

In recent years, researchers have noted that it will helpful to extend RBAC model to dynamic coalition environments. In this paper, we have proposed a flexible authorized delegation model, which combines RBAC model with delegation. Not only does our proposed model support many types of delegation such as user-to-user delegation, permission-to-role delegation, role-to-role delegation, permission-to-user delegation, and role-to-user delegation, but it also supports fine-grained delegation with role and permission level. The proposed model also provides control on depth of delegation, temporal constraints and multi-option revocation. In this paper, we discuss the scheme to store and collect delegation certificates. Compliance checking is done through the backward search algorithm.

The future work includes the study of constraints in delegation. Also, separating assignment-right of role (or permission) from use-right of role (or permission) is a future research issue.

Acknowledgments

This paper is supported by National Natural Science Foundation of China (No. 50775070), and by Hunan Provincial Natural Science Foundation of China (No.07JJ6104), and by Scientific Research Fund of Hunan Provincial Education Department (No. 07C272).

References

[1] R. S. Sandhu, E. J. Coyne et al. Role-Based Access Control Models. IEEE Computer, Vol.29, No.2, pp.38~47

[2] E. Barka and R. Sandhu. Framework for Role-Based Delegation Models. In Proc. of 16th Annual Computer Security Application Conference (ACSAC 2000), December 2000, pp.168~176

[3] E. Barka and R. Sandhu. A Role-Based Delegation Model and Some Extensions. In Proc. of 23rd National Information Systems Security Conference (NISSC 2000), December 2000

[4] L. Zhang, G. Ahn, and B. Chu. A Rule-Based Framework for Role-Based Delegation. In Proc. of 6th ACM Symposium on Access Control Models and Technologies (SACMAT 2001), May 2001, pp.153~162

[5] L. Zhang, G. Ahn, and B. Chu. A Rule-Based Framework for Role-Based Delegation and Revocation. ACM Transactions on Information and System Security, Vol.6, No.3, 2003, pp.404~441

[6] X. Zhang, S. Oh, and R. Sandhu. PBDM: A Flexible Delegation Model in RBAC. In Proc. of 8th ACM Symposium on Access Control Models and Technologies (SACMAT 2003), June 2003, pp.149~157

[7] H. Lee, Y. Lee, and B. Noh. A new Role-Based delegation Model Using Sub-role Hierarchies. International Symposium on Computer and Information Science (ISCIS 2003), November 2003. LNCS 2869, pp.811~818

[8] D. Park and Y. Lee. A Flexible Role-Based Delegation Model Using Characteristics of Permissions. The 16th International Conference on Database and Expert Systems Applications, August 2005. LNCS 3588, pp.310~323

[9] J. Wainer and A. Kumar. A Fine-grained, Controllable, User-to-User Delegation Method in RBAC. In Proc. of 10th ACM Symposium on Access Control Models and Technologies (SACMAT 2005), June 2005, pp.59~66

[10] James B. D. Joshi and Elisa Bertino. Fine-grained Role-based Delegation in Presence of the Hybrid Role Hierarchy. In Proceedings of 11th ACM Symposium on Access Control Models and Technologies (SACMAT 2006), June 2006, pp.81~90

[11] M. Blaze, J. Feigenbaum, and J. Lacy. Decentralized trust management. In Proceedings of the 1996 IEEE Symposium on Security and Privacy, May 1996, pp.164~173

[12] M. Blaze, J. Feigenbaum, J. Ioannidis, and et al. The KeyNote Trust-Management Version 2. RFC 2704: http://www.faqs.org/rfcs/rfc2704.html, 1999

[13] D. Clarke, J. E. Elien, C. Ellison, and et al. Certificate Chain Discovery in SPKI/SDSI. Journal of Computer Security, Vol.9, No.4, 2001, pp. 285~322

[14] N. Li, J. C. Mitchell, and W. H. Winsborough. Design of a Role-based Trust-management Framework. In Proceedings of the 2002 IEEE Symposium on Security and Privacy, May 2002, pp.114~130

[15] N. Li, W. H. Winsborough, J. C. Mitchell. Distributed Credential Chain Discovery in Trust Management. In Proceedings of the 8th ACM Conference on Computer and Communication Security, November 2001, pp.156~165

[16] C. A. Gunter and T. Jim. Policy-directed Certificate retrieval. Software: Practice and Experience, September 2000, No.30, pp.1609~1640

[17] T. Jim. SD3: A Trust Management System with Certified Evaluation. In Proc. of the 2001 IEEE Symposium on Security and Privacy, May 2002, pp.106~115

The Intelligent Tester of Flammability Based on Kinetic Control Technology

Suping Gao[1] Lijuan Yin[2] Xiaoguang Xu[2]

1 Automation Technology Institute, Shenzhen Polytechnic University, Shenzhen, Guangdong, 518000, China

Email：gaosup@oa.szpt.net

2 Shenzhen Entry-Exit Inspection and Quarantine Bureau, Shenzhen, Guangdong, 518045, China

Email：yinlj411@163.com

Abstract

This paper presents a intelligent flammability tester based on the PLC control technology and working principle of step-in motor, it demonstrates the working principle, hardware structure, software programming, functional characteristics of the developed intelligent tester of flammability in detail, expatiates expressly the step-in motor control method and its application and points out the tester can meet the flammability test requirements of various types of toys of Europe standard consumer safety specification for toy safety and international standard of toy safety. It refers that the tester realizes the test process with automatism, intelligence and information.

Keywords：Flammability; Toy safety test; Step-in motor; PLC; EM253

1 Introduction

All countries in the world focus on Toys safety all the time, Europe standard of toy safety EN71, America standard of toy safety ASTM F963 and International standard of toy safety ISO8124 have strict requirement on toy test of mechanical and physical properties and flammability, and nearly increase new requirement each year.

China is one of the biggest toys export countries, the toy must pass through a series corresponding standard test before import and export, and the flammability test is one of the most important testes. At present, the research in the tester of the toys' flammability test is far inferior to other products such as home electrical appliances and mechanical and electrical engineering appliances at home and abroad. Now the test methods in China are nearly in entire by artificial way, the test will be inefficient, inaccurate and poor repeatability. It had some problems such as hard to fix the toys, adjust the ignite position, measure the height of the toy and the flaming dimension exactly, record the flaming time, record the flaming spread time and make sure the igniting time and so on.

With the development of the computer control technology, PLC (programmable logic controller) now is the wider application of the control device. It has many advantages as modular structure, high-speed processing speed, accurate computing, variety of control, network technology and so on. Step-in motor which has the characters of low rotor inertia, high positioning accuracy, no cumulative error and simply control now is one of the major implementation components and widely applies in all kinds of control systems and electromechanical integration equipments. The tester which used advanced PLC control technology and step-in motor will solve the problems about the flammability test of toys.

2 Structure Design

Intelligent flammability tester designed as cabinet form configuration and model structure. It make up of eight parts which are ventilating cabinet, test device,

control system, ignition device, HCI(Human Computer Interface), output peripherals, communication interface and long-distance monitoring control system, as shown in Figure1.

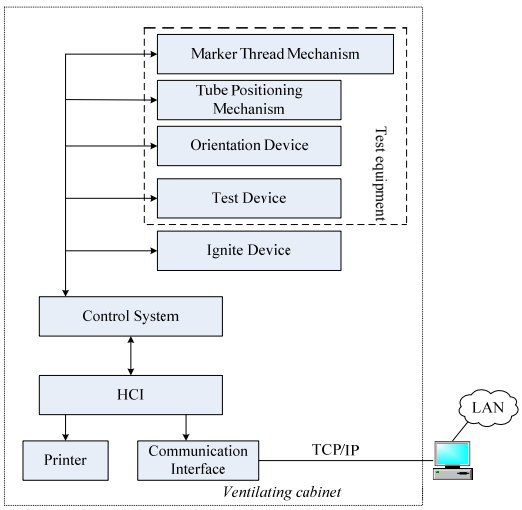


Figure1 Tester Structure

Ventilating cabinet: the cabinet of stainless steel with the centrifugal-type exhaust fan; it is constituted by the sealed test chamber, glass observation window and the control chamber.

Tester: In the test chamber, it is constituted by "orientation device" which fixes and positions the test sample, "test device" which measures the length and the flaming length of the sample, the "tube positioning mechanism" which position the flame tube, "tube mechanism" which position the flame tube in horizon and "marker thread mechanism" which position the marker thread. These mechanisms make up of different guide screws, mechanical slide, guide track, electro-optic limit switch and timer switch, and also setup the temperature and humidity sensor.

Control system: it makes up of the PLC controller, step-in motor, driving control device and power source. The PLC controller includes a CPU224XP of Siemens S7200 series, two extended bit-control module EM253 and a DC power of 24V. CPU224XP contains 14 digital inputs, 10 digital outputs (DC output transistor circuit type), 2 analog inputs and 1 analog output. The bit-control module EM253 which connects with the CPU224XP by the extended cable achieves the function of kinetic control. Step-in motor, which uses the 86BYG250BN of Helishi Motor Technology Co. Ltd in Beijing, has step angle 1.8°. It also has the characters of positioning accuracy, large range of speed, high resolution, low-speed and running smoothly, low power consumption and so on. With the function of offline and staggered phase protection, driving control device uses the corresponding SH-20806N-DA drive which offers the choice of whole step,2,4,8,16,32,64 dividable modules. The output phase current can be set by the corresponding code switch in the drive[2].

Ignition device: including acicular flame tube, regulating valve and the soft transportation hose of butane and gas.

HCI (Human Computer Interface): MT508TE4 touch screen of WEINVIEW Company, TFT LCD of 8 colors, resolution of 640×480, 256 colors, CPU of 200MHz, specially designed for PLC control application.

Communications Interface: including one printing port, 2 RS-232 ports and Ethernet port.

Peripherals: uses the micro stylus printer.

Control system controls test device and ignition device, touch screen connects PLC with serial port and the printer with parallel port. The system connects the long-distance watching PC or managing net through Ethernet

3 Working Principle

3.1 Step-in control principle

Step-in motor which is controlled by the pulse signal is an executive mechanism. It changes the pulse signal into corresponding angular displacement or line displacement. Being controlled by pulse signal, the rotor angular displacement and speed are directly proportional to the number of input pulse and the frequency of pulse. The numbers of pulse control angular displacement to achieve positioning exactly; the frequency of pulse controls the speed and acceleration of the motor to adjust the speed of the tester. It is changed by the electrified order to change the motor direction[1][9-10].

Using the PLC to control step-in motor, the pulse equivalent, the high limit frequency and the max number of the pulse in the system should be calculated with the functions as follow. The frequency when the PLC high-speed pulse needed is determined by the frequency of the pulse. The width of bit in PLC is determined by the number of the pulse. Step-in motor can be controlled with the function of PLC's high speed pulse and the function of kinetic control.

3.2 Work principle

The system of the tester is a step-in motor control system that based on PLC, touch screen and the drive, as shown in Fig.2. PLC is the controller of the system, step-in motor is the executing device, receiving PLC control signals which are controlled by the drive, connecting with touch screen and signal detection device to form a closed-loop control system.

After a toy is fixed to the test object mechanism, choose the toys type through the operation interface on the touch screen. The tester sets the benchmarks according to the type of the toy and then sends the pulse signal when the PLC in the control system processes the signal according to the positioning benchmarks and HCI information. The drive receives these signals and makes them match and amplify to control the step-in motor. Step-in motor as the executing device drives load (every mechanism) to run automatically; PLC in accordance with the running state of mechanism will record the output pulse signal automatically, process these signals, calculate the value of the operation accurately, and get the test toys' parameters of the length of toys, flaming length, flaming height positioning and so on, through various mathematical model of different mechanism. And get the precise flaming time in accordance with the signal of the timing switch, automatically calculate the test results that including toys effective length, extending the length, height positioning of flame, flaming time, flaming length, flaming ratio or flaming efficiency, the environment temperature and humidity, the test date and time. These results can be shown in man-machine interface, printer, network and so on. The tester has four-axis to drive, forms a closed-loop test and controls process, as shown in Figure2.

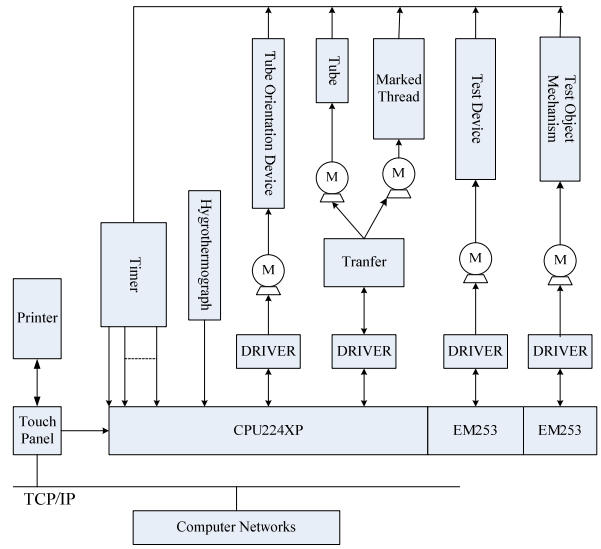


Figure2 Working Principle

In the control system, the high-speed pulse signal out from the CPU224XP controls 3 step-in motors directly, every one bit-control module EM253 controls one step-in motor. Since the drive chooses 32 dividable functions and the phase current is 4A, this tester tests high speed and high accuracy with the 4-Axis.Setting limit switch for every test device not only holds the test range, but also make the test safe. One drive controls running of the 'tube' and the 'marker thread'. Flame tube can achieve three-dimensional positioning with the 'tube positioning mechanism' and the 'tube' mechanism. The tube's positioning rang from 0° to 180°, positioning radius is 600mm, response time is 0.001s, and the source power is 220V AC[5-8].

4 Function Design and Realization

（1）Many functions: This tester is developed for no electric toys which fit the flammability requirements from 4.2 to 4.5 in the European Standard EN71-2. It can finish the flammability test of 5 types of toys as Soft-filled toys≤520mm, Soft-filled toys＞520mm, toys to be worn on the head such as beards, moustaches, wigs and so on (≥50mm), toys to be worn on the head such as

beards, moustaches, wigs and so on (＜50mm), flowing elements of toys to be worn on the head hoods, head-dresses and so on, and toys intended to be entered. In the test, you can press button to choose the right type of toys by the toy character, and then the tester will finish the test[3][4].

（2）Design of fixing toys: fixing toys is the key chain of the test due to the varied toys. The tester can fix up all kinds of toys with a three-dimensional plank and a 'U' plank, using the slide point between the track and the slide block to increase fixing-point. It operates conveniently and makes the sample balance easily so it works stability. This tester, which solves the problem that different toys need different fixing method, can fix toys rang from 80mm to 520mm and toys more than 520mm.

（3）Design of positioning method: in the test, find the benchmarks position according to the different toys first, and set this position as the benchmarks position of the whole tester including the test object position, test device position, three-dimensional position of the flame tube and the position of test marker thread. The positioning accuracy comes to ±0.5%

（4）Test process automation: the test article, load devices, PLC, step-in motor, driving control device, limit switch, sensor and the touch-screen form into a closed-loop test. Base on the accurate benchmarks position and the toys types, the mathematical model of the test process is created and the test works automatically.

（5）Intelligent processing: combustible parameters can be set by the keys on the Man-machine interface. The tester shows the test process and the test results, prompts attentions to the operation and warns the wrong operation and the error messages. The test results can be remembered, stapled and quired.

（6）High accuracy, high speed and strong stability: since the PLC processing fast, the step-in controlling response rapidly, step-in motor and drive control device dividable multiples setting and their matching signals reasonably, so the test will be done with high speed, high accuracy, strong stability and good repeatability.

（7）Design of Human-Computer Interface: It is easy to use and simple to operate. Operating interface mainly includes main screen, toys stapled choosing screen, stapled toys testing screen, stapled toys results screen, records screen and debugging screen. The changing relations of these screens are shown in Fig.3. It sets protect function for the wrong operations and also prompting messages menu or 'ENTER' operation for the important step and notes with different test of toys. All the operations which all be shown in Chinese in the test will be done on the touch screen directly. It works simply, conveniently and intuitively.

（8）Records save and query function: the tester saves 10 effective recent results which can be queried directly by the operation screen on the touch screen. If the tester connected to the Ethernet or LAN, the results in the hard disk or CD-ROM can be saved and queried although after a long time. Every test result is saved in the file with the name of the date and time when doing the test. The file can be opened by the notepad and the office software like Word, Excel and so on.

（9）Print function: records printing function for every tested result.

（10）Ethernet function: The tester is a network device, with the Ethernet interface, complied with the TCP / IP principle, with the communication rate of 100 Mbps, as long as the IP address of the correct settings. It communicates with LAN, so as to share resources and provide protection for the realization for information and digital of the inspection.

（11）Long-distance monitoring function: The tester connects to the Ethernet LAN. As long as the PC or long-distance monitoring computer install the monitoring software, the computer can monitor the tester, operate or watch the test process, set or modify the parameters, query the test results.

5 Software Design

PLC sends out the received import signals that had been processed to control the motors. The main control program is shown in Fig.4. Firstly, PLC handles the received signals which are the type of toys, test order

and so on, and then sends out pulse controlling signals, the drive which receives these signals makes them match and amplify to control the step-in motor running; Secondly, determine the benchmarks position, this position is the benchmarks position of the whole flammability tests, this mark is benchmark to determine the length of tested toy, the flaming height and the flaming length; and then the tester will run with the appropriate position and test subroutine, controls four-axis mechanism to find flaming position; After confirming the flaming position, ignition and flaming with the ignition device; PLC control the time of flaming according to the type of toys, when the time runs out, flame tube will be taken out from flaming position; After the toy flames adequately, the system records the flaming time automatically, measures the flaming length and the temperature and humidity of the test environment. Finally, according to these test data, PLC processes them to get the test results and show them in the form of display, save record, print and network.

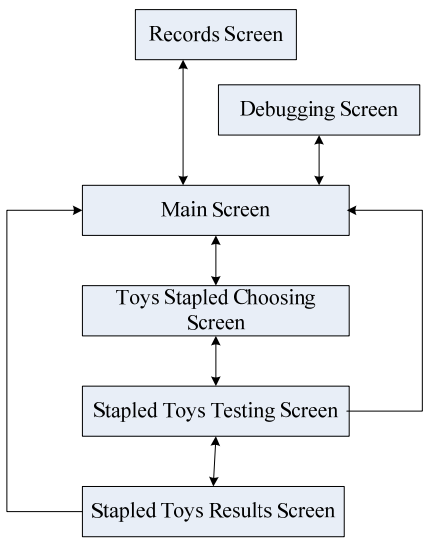


Figure3 Relations of Operation Screen

PLC completes the whole test running different positioning subroutine, test subroutine and data processing subroutine, these subroutine perform needed function according to toy types and the determined datum mark.

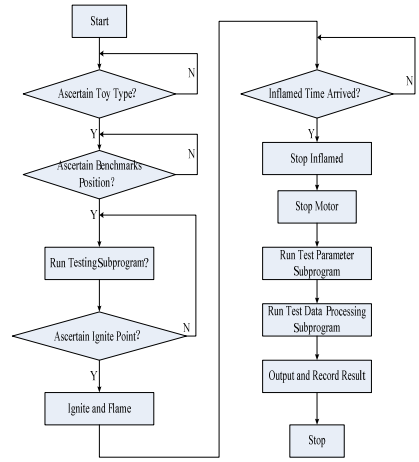


Figure4 Control System Program

6 Conclusions

With the development of computer controlling technology, PLC gets more and more strong functions and controls the motor position, speed and acceleration by EM253. Using this technology, the tester not only satisfy the requirements of the flammability in toy safety standard, but also has high accuracy, fast responding, good repeatability and simply operation, so as to make the whole test with automatism, intelligence and information.

References

[1] Jinqiu Zhang, PLC principle and application, Beijing Machinery Industry, 2003

[2] Shuqing Wang, Advanced control technology, Beijing Chemistry Industry, 2001

[3] BS EN71-2:2003 the European Standard

[4] International Standard ISO 8124-2

[5] Zhuang Lijuan, sewage manage control system base on PLC, Journal of Control & Automation, January 2005, pp.24-25

[6] Chen Zaiping and Zhao Xiangbin, PLC technology and application system design, Beijing: Machinery Industry Pub., January 2003

[7] Gong Shuzhen，PLC principle and application, Beijing: People's post & telecommunications publishing house, 2002

[8] Fang Chengyuan, Electric control principle and design, Beijing: Machinery Industry Pub., 2000

[9] Wang Yonghua, Modernistic electric and PLC technology, Beijing: Beihang University Pub., September 2002

[10] Tang Yifan, Electric and PLC technology, Beijing: Machinery Industry Pub., 2004

An Resource States Detecting Algorithm For Manufacturing Grid

Huifu Zhang Hong Wen Anhua Chen Deshun Liu Wenhui Xiao Ran Chen

School of Computer Science and Technology, Hunan University of Science and Technology
Xiangtan, Hunan 411201, China
Email: hfzhang@hnust.edu.cn

Abstract

Manufacturing grid serves as an integrated platform for manufacturing resources, and the good knowledge of and efficient control over the state of manufacturing resources are critical in that the resource exceptions can be timely noticed and removed so as to ensure reliable service. Given the features of manufacturing grid, the writer, adopting the strategy of active detecting in the mode of PULL, propose a new mode of states detecting algorithm of manufacturing grids. This detecting algorithm is capable of meeting various demands of all kinds of manufacturing resources, supporting the strategy of priority of local resources, and demanding no clock synchronization. With experimental evaluation, the detecting algorithm shows good stability and correctness.

Keywords: Manufacturing Grid; Resource Sharing; Failure Detector; Heartbeat.

1 Introduction

The individuation, diversification and shorter life cycle of future products advances the requirements of fully sharing manufacturing resources. It is an efficient approach for continuable development of manufacturing by taking full advantage of network technique to recombine the manufacturing resources. Grid technology supplies an efficient method for resources sharing[1,2]. Domestic scholars have obtained much achievement on this aspect[3,4,5].

The grid is a large-scale distributed system with complicated construction, excessive and wide-spread heterogeneous nodes. So the failure detector is a general component of constructing the grid condition[6,7]. Chndra and Toueg have firstly presented that failure detector is an available approach to augment asynchronous system computation model[8]. At present, failure detector is widely applied to grid computation[9], cluster management, communication protocol and other related fields[10,11]. Failure detection is also the general reliable safeguard technology for manufacturing grid system, detecting shared manufacturing resources in time and changing the detecting quality dynamically according to the grid, the local resources demands and system conditions.

Heartbeat strategy is the most common implementation of failure detector. There are PUSH and PULL models of heartbeat in terms of implementation mode[12]. Among them, PULL model is active and initiates only when needed, which is much more adaptive for grid computation. So, the detection strategy based on PULL model is adopted in manufacturing grid.

As for evaluating the QoS of failure detector accurately, Chen and others have proposed a qualitative set of QoS metrics[13]: Detection time (TD), Mistake recurrence time (TMR), Mistake duration (TM), these indexes can guarantee the integrality and accuracy of the failure detector.

2 The Basic Conception & Characte -ristics Of Manufacturing Resources State

Manufacturing Grid(MG) is an integrated supporting environment both for the share and integration of resources in enterprise and social and for

the cooperating operation and management of the enterprises. Based on the grid and relative advanced computer and information technologies, MG shields the heterogeneousness and the regional distribution of resources by the way of encapsulating and integrating of the design, manufacture, management, information, technology, intelligence and software resources separated in different enterprises and social groups[3].

As an integrated platform of manufacturing resources, the Grid encapsulates the resources into grid service. These resources may be temporarily invalidated as a result of network fault, system detection, etc. we define a manufacturing grid system in which the number of manufacturing resources is n, then the resources set is $\sum = \{R_1, R_2, ..., R_n\}$. To narrate much easier, we assume that there is an imaginary global clock. where the value field is natural number set, T is the time set. Then, Failure Detection (FD) can be defined as follows[8]: Supposed $p \in \sum$, $t \in T$, the function $FD_{p(t)}$: $\sum \times \mathrm{T} \to 2^{\sum}$; $q \in \sum$, $q \in FD_{p(t)}$, then it means that q that is the Failure Detection of p is invalidated at the time t. The output result of the function FD is the objects that is found to be disabled by detecting, i.e.: $Failed = \bigcup_{t \in T} FD_{p(t)}$.

As far as a manufacturing resource q in manufacturing grid is concerned, there are two basic conditions: UP and DOWN. UP means that q is able to provide sharing service; DOWN means that q is not able to provide sharing service.

Compared with other distributed conditions, the difficulties that exist in the state detection of manufacturing grid resources are as follows:

1. Diversity of Manufacturing Resources:

Manufacturing resources of the product whole life cycle in the manufacturing grid include many types of resources, such as hardware equipment, software, human resource, and so on. They have a complicated logic relation, so different methods are needed when we name, define, organize, or visit these resources.

2. Tremendous whole Life Cycle Manufacturing Resources:

Manufacturing grid system will gradually mature and tend towards alignment or consociation, becoming the global scale of the manufacturing enterprise. Organization and management of so many resources involve the management factors such as geography position, topological structure, resources type, relation between resources, user need etc..

3. Local Control Priority of Detected Resources:

Manufacturing grid links lots of manufacturing resources together to share. While the resources that join the manufacturing grid are unified to schedule by the grid, it gives priority to the local resource to control the manufacturing resource. That's to say, unified scheduling of the manufacturing grid should submit to local scheduling of the resource, which inevitably make the resource management in the manufacturing grid more complicated.

4. Dynamism of Manufacturing Resources:

The network is a changing environment. The information that delivers through the network will delay at certain scope, and even jam for a while. The strategy of resource local management is uncertain, with components leaving or joining all the time, increasing the dynamic requirements of the resources management.

As for the manufacturing grid system, in addition to the problems cited above, how to resume the resource should also be considered. Different from other distributed systems, dispatching strategies of resources sharing in the manufacturing gird are mainly from the owners of local resources, namely with the resources joining or leaving freely and dynamically. In addition to state detection of resources, the detection service of manufacturing resources is able to identify the unable sharing of resources because the owners of the resources withdraw them from the manufacturing grid. Therefore when the detection service of the resources assures that a certain resource at present is under invalid circumstance, it is also necessary to continue to detect the resources with another detection strategy so that the resources timely resume resource sharing. The above circumstance is also applicable to the network fault, which can be resumed in a certain time.

In addition to reporting that the resource is available or not, the detection service of the resource

should also record instantaneous service state of the resource, which mainly includes network delay, service load of the resource, and so on. Through these status messages, the manufacturing grid system can dispatch the resources more effectively.

3 State Detection Service Of Manufacturing Grid Resources

The state detection service of each node sends heartbeat messages to the services of other nodes termly; the opposite service will report the current work state of the node with heartbeat message reply after receiving the heartbeat request. The time delay of the received heartbeat messages is counted and analyzed; the expectation and variance of the time delay of the resource service are calculated and stored as its network state using for the performance reference of the resource scheduling. In the meantime, the time of the next heartbeat message according with the confirmation request is also calculated based on the time delay condition. As for the situation that heartbeat messages haven't been received because of overtime, another detection will be carried out in a very short time. If the long-distance node is shut off or busy in a period of time, the detection service will adopt corresponding strategy and lower the heartbeat speed accordingly.

In such failure detection model, failure detection classes are defined as follows[14]:

Strong completeness[8]. Eventually every process that crashes is permanently suspected by every correct process.

Eventual strong accuracy[8]. There is a time after which correct processes are not suspected by any correct process.

We assume that the QoS requirements are expressed using the primary metrics. More precisely, a set of QoS requirements is a tuple (T_D^U, T_{MR}^L, T_M^U) of positive numbers, where T_D^U is an upper bound on the detection time, T_{MR}^L is a lower bound on the average mistake recurrence time, and T_M^U is an upper bound on the average mistake duration. In other words, the QoS requirements is denoted by the following equation (1):

$$T_D \le T_D^U, T_{MR} \ge T_{MR}^L, T_M \le T_M^U \tag{1}$$

The delay of heartbeat messages i.e. Δt is a random event, which distribute in the range (0, ∞), E(Δt) and D(Δt) are respectively used to denote distribution expectation and variance of the time delay, thus $E(\Delta t) = \sum_{i=1}^{\infty} \Delta t_i \cdot p_i$, $D(\Delta t) = \sum_{i=1}^{\infty} [\Delta t_i - E(\Delta t)]^2 p_i$, where pi is the probability of Δti, i.e. pi =P{Δt=Δti}, i=1,2, ⋯. In order to calculate distribution expectation and variance of the time delay, a sliding window WT having fixed size is needed so as to save the detection time record of w query messages after receiving the time reply, set up the samples of E(Δt) and D(Δt). The accuracy of the system computation is opposite to the cost of system detection. In order to calculate the more precise mean of the time delay, it is necessary to increase the size of the sliding window; to decrease the system cost, it is indispensable to reduce the size of the sliding window. We use the following equation to calculate the mean of the time delay, thus which can not only improve the computation precision, but also reduce detection cost in the course of the massive resource detection.

$$E(\Delta t_{i+1}) = \frac{i \cdot E(\Delta t_i) + \Delta t_{i+1}}{i+1} \tag{2}$$

Under the situation that the variance precision can't be acquired , we adopt the following equation:

$$D(\Delta t_{i+1}) = \frac{i \cdot D(\Delta t_i) + [\Delta t_{i+1} - E(\Delta t_{i+1})]^2}{i+1} \tag{3}$$

The process how Correct resource p uses the detection service to detect resource p is expressed as follows: upon detecting, resource q firstly sends detection strategy request to detected resource p; resource p can return either the QoS requirements of local detection or special value to tell resource q to adopt the default globally QoS parameter. The detection service of resource q is responsible for sending periodicallyΔheartbeat to resource p as detection strategy requests; resource p returns its status information to resource q after receiving the detection strategy requests so as to denote that what kind of resource status resource is (busy, free, under

examination, etc.). After resource q received the reply, it saves the detection time of the message after a time delay, and calculates E(Δti)、D(Δti) and ρ .If the reply can't send back because of network fault, the system record it as one overtime, and use the timeout as the time delay for calculation. PS and PD are set up according to the overtime number Cd and the requirements of QoS; if the overtime number reaches PS , resource q is suspected, and adopted the suspect detection strategy; if the overtime number reaches PD , resource q is assured to be out of work, and adopted the state detection strategy. The whole algorithm of the state detector is shown as the following Figure 1.

```
Process q:
Initialization:
1.      call p.getLocalDetectStratagem
2.      setΔTbusy; ΔTdown; ΔTsuspect, PS, PD;
3. do:
4.      sleep(Δheartbeat);
5.      p.status = Call p.getHeartBeatStatus();
6.      if timed out Cd ++ else Cd=0;
7.      calculate and save current E(Δti)、D(Δti);
8.      calculate ρ i, γ;
9.      if Cd > PD p.status set DOWN;
10.       else if Cd > PS p.status set SUSPECTED;
11.       else p.status set UP
12.       calculate Δheartbeat;
13. loop
Process p:
14. upon receiving getHeartBeatStatus from q do
15.     return current status to q;
```

Figure 1 The algorithm of the heartbeat detection

Obviously, Δheartbeat is the important parameter that influences the mistake detection time. In this algorithm, Δheartbeat can be calculated according to the following equations :

$$\rho = \frac{D(\Delta t_i)^2}{(\Delta t_i - E(\Delta t_i))^2 + D(\Delta t_i)^2} \tag{4}$$

$$\gamma = \mu_{busy}\Delta T_{busy} + \mu_{down}\Delta T_{down} + \mu_{suspect}\Delta T_{suspect} \tag{5}$$

$$\Delta_{heartbeat} = \begin{cases} \rho(\Delta T + \gamma) & \Delta t_i < E(\Delta t_i) \\ \Delta T + \gamma & \Delta t_i \geq E(\Delta t_i) \end{cases} \tag{6}$$

Where

Δti——time interval between two detections;

E(Δti)——average reply time of detection i;

D(Δti)——variance of the average reply time of detection i;

ΔTbusy——time interval of state detection when the resource is busy by local or globally strategy;

ΔTdown——time interval of state detection when the resource crashes by local or globally strategy;

ΔTsuspect——time interval of state detection when the resource is suspected by local or globally strategy;

μbusy——its value is 1 when the resource is busy, otherwise it is 0;

μsuspect——its value is 1 when the resource is suspected, otherwise it is 0;

μbusy——its value is 1 when the resource crashes, otherwise it is 0;

γ and ρ are middle computational variables.

Seen from equations 4, 5 and 6, the heartbeat time is related with the network average delay and the current time delay. The longer the current time delay is, the faster the next heartbeat will be, shown as equation (4). The heartbeat speed is also related with the state of the resources, such as the heartbeat speed in each states and heartbeat strategies of the local or globally setup, shown as equation (5). During each heartbeat query work of resource state is carried out, if the number of time delay is exceed the valve value of the detection, the detected resource is determined to break down.

Suppose a right resource q uses the detection to check the resource p. The p will make no response to any service after it crashes, that is to say, q will receive no message after that. The counter for continuous errors of the algorithm increases by degrees, the resource was set crashes when the counter exceeding the sated value, and the crashes detection strategy for resources is adopted. When the resource state is normal, each request of the detection service can be returned correctly; and corresponding detection strategy is adjusted when the resource state changes.

For the detection service, each resource that crashes may be finally suspected by every correct detection service. After a certain time t, each right resource can return its own state to the detector, so as not to be mistakenly suspected. According to the classification of state detection services by Chandra and Toueg, the detection service is equal to a ◇P one.

4 Implementation Of State Detection Service And Test Results Analysis

We have carried out three-day testing through the detection service to test the resources located at other places. The detection results meet the state reference requirements in the course of scheduling the manufacturing resources, which is worth to be the reference of resource scheduling.

The detection result of the resources for a period of time is showed as Figure 2, from which we can see that, when the resource state changes, the detector can report its actual conditions within the scope of QoS requirements set by the strategy.

From the Figure 2, we can also see that when the resources break down because of the network communication jam, the measures adopted by the detection service, which can speed the detection hearbeat and identify the status of the resources in a short time.

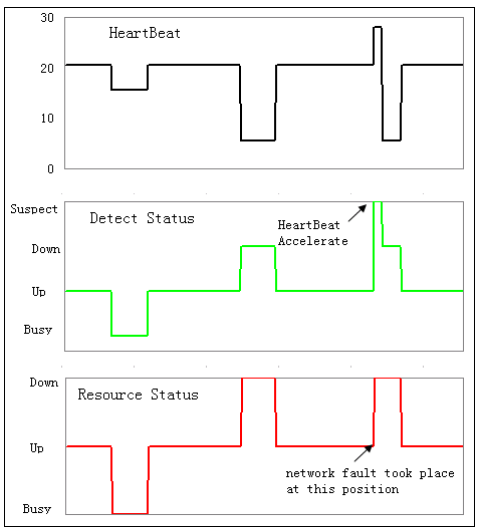


Figure 2 The detection result of actual detection service of a certain design resource at a certain time

5 Conclusion

Resource state detection is an essential service function for better resources sharing in manufacturing grid system. Based on heartbeat message mode, this paper offers the state detection service algorithm of manufacturing grid, and designs the state detection general component. This service makes the ◇P failure detector a reality. It gives priority to the local detection strategy of manufacturing resources to adjust detecting quality or default detection strategy, and adapt to the changing network conditions of the manufacturing grid. This detector works on the background of the manufacturing grid, reports timely the state of shared resources as for references of cooperative dispatch. According to the results from long-term testing operation on the manufacturing grid platform, this service has comparatively good effect on resource detection, satisfies the resource state requirements of manufacturing grid system.

As a general component of the manufacturing grid, the stability and simultaneity of detecting ability of large capacity resources is vital to the detector. As for our future work, we will test the performance and long-time operation stability in a large-capacity manufacturing grid system.

Acknowledgments

This work was supported by 973 Project under grant no: 2003CB317001 and Project 50775070 supported by NSFC, by Scientific Research Fund of Hunan Provincial Education Department under grant no: 07C272.

References

[1] Foster I, Kesselman C, Nick J, Tuecke S. Grid services for distributed system integration. IEEE Computer, 2002. 35(6): 37~46

[2] Czajkowski, K.; Fitzgerald, S.; Foster, I.; Kesselman, C.. Grid information services for distributed resource sharing. Proceedings. 10th IEEE International Symposium on High Performance Distributed Computing, 2001.8. 181 - 194

[3] Yushun Fan, Dazhe Zhao, Liqin Zhang, Shuangxi Huang, Bo Liu. Manufacturing Grid: Needs, Concept, and Architecture. Grid and Cooperative Computing: Second International Workshop, 2003.12. 653-656

[4] Zhanbei Shi, Tao Yu, Lilan Liu. MG-QoS: QoS-Based

Resource Discovery in Manufacturing Grid. Grid and Cooperative Computing: Second International Workshop, 2003.12. 500-506

[5] Lilan Liu, Tao Yu, Zhanbei Shi, Minglun Fang.. Resource Management and Scheduling in Manufacturing Grid. Grid and Cooperative Computing: Second International Workshop, 2003.12. 137-140

[6] Jin H, Zou DQ, Chen HH, Sun JH, Wu S. Fault-Tolerant grid architecture and practice. Journal of Computer Science and Technology, 2003,18(4): 423-433

[7] Horita Y, Taura K, Chikayama T. A scalable and efficient self-organizing failure detector for grid applications. In: Katz DS, ed. Proc. of the 6th IEEE/ACM Int'l Workshop on Grid Computing. Washington: IEEE CS Press, 2005. 202-210

[8] Chandra TD, Toueg S. Unreliable failure detectors for reliable distributed systems. Journal of the ACM, 1996,43(2): 225-267

[9] P. Stelling, I. Foster, C. Kesselman, C. Lee, and G. von Laszewski. A fault detection service for wide area distributed computations. In Proc. of the 7th IEEE Symp. on High Performance Distributed Computing, 1998.7. 268-278

[10] Dong J, Zuo DC, Liu HW, Yang XZ. An adaptive failure detector for grid based on QoS. Journal of Software, 2006,17(11): 2362-2372

[11] Chen NJ, Wei J, Yang B, Huang T. Adaptive failure detection in Web application server. Journal of Software,2005,16(11): 1929-1938

[12] Hayashibara N, Cherif A.. Failure detectors for large-scale distributed systems. Kikuno T, ed. Proc. of the 21st IEEE Symp. on Reliable Distributed Systems (SRDS 2002). Washington: IEEE Computer Society, 2002. 404-409

[13] Chen W, Toueg S, Aguilera MK. On the quality of service of failure detectors. IEEE Trans. on Computer, 2002. 51(5): 561-580

[14] Gupta I, Chandra TD, Goldszmidt GS. On scalable and efficient distributed failure detectors. Kshemkalyani A, Shavit N, eds. Proc.of the 20th Symp. on Principles of Distributed Computing (PODC 2001). New York: ACM Press, 2001. 170-179

Appication of Embeded System in Sharing Manufacturing Resources

Huifu Zhang Ran Chen Xiaohui Xie Wenhui Xiao

School of Computer Science and Technology, Hunan University of Science and Technology
Xiangtan, Hunan 411201, China
Email: hfzhang@hnust.edu.cn

Abstract

Embedded software and systems are increasingly becoming a key technological component of manufacturing systems. Competition demands that manufacturing be more connected. All of these make the manufacturing resources share a focus of advanced manufacturing research. In this paper, the characteristic of embedded system and the architecture of embedded system grid are introduced; the interface method between embedded system and grid and their corresponding interface are discussed. The control model of embedded system resources under the gird environment, the method of overall sharing for embedded system and its equipment, the embedded system grid technology of sharing and controlling embedded resources are also presented.

Keywords: Manufacturing Resource; Embedded System; WSDL; GARA

1 Introduction

With the advent of system level integration (SLI) and system-on-chip (SOC), the center of gravity of the computer industry is moving from personal computing into embedded computing. An embedded system is a special-purpose computer system, which is completely encapsulated by the device it controls. An embedded system has specific requirements and performs pre-defined tasks, unlike a general-purpose personal computer. The embedded system has played an important role in such fields as industrial robot, medical equipment, telephone system, satellite, and flight system, et al. At present, the key to realize intellectual control and network control of manufacturing equipments is to embed the embedded system into manufacturing equipment and extend the control of manufacturing equipments to the long-range network through the network function of the embedded system.

With the development of network technology in recent years, the demand to share, manage and control non-Internet equipment information is formed gradually. Grid is a new developing technology built on Internet[1,2]. Through high-speed sharing network that connects extensively geographical distributed isomeric resources, it is used to solve individual problem that usually needs a lot of CPU or memory to cooperate with to dispose and visit. In the formative latticed network of calculating of high performance, each resource is a grid node. Grid resources include computers, workstations, cluster systems, storage equipment, databases, et al. of different operating systems, even some special scientific instrument including sensor equipment, daily mobile phones and such intellectual equipment as PDA. In this paper, gird interface technology, gird resource management and Characteristics of embedded system are studied; a method of sharing and controlling embedded resources under the environment of grid is proposed.

2 Technological Characteristics of Embedded System

The embedded system usually consists of two parts including hardware and operating system that constitute running environment of software. Running environment

and application occasion of the embedded system determine some characteristics, which are different from other operating systems.

Most embedded operating systems have adopted the little kernel structure; the kernel only offers the basic function, such as task scheduler, communication and synchronism among tasks, memory management, clock of task et al.. Other application groupware, for instance: network function, file system, GUI system, etc. that work under user state, work in terms of system process or function call. Therefore the systems all can be cut down; users can select the corresponding groupware according to their own needs.

At present, most embedded operating systems adopt the grabbing dispatching method based on PRI to solve the task of different PRI, and adopt time slice cycle dispatching method to solve the task of the same PRI.

Some embedded systems have relatively high expectations for time; we call it the real-time system. There are two kinds of real-time systems: hard real-time system and soft real-time system. The soft real-time system does not limit a certain task to be finished within a certain time and only require every task to run well faster. The hard real-time system is strict with system response time, once the system response time can't be satisfied, may cause the system crash or deadly mistake. Generally, the hard real-time system has wide application in industrial control.

It can be seen that the developers of the embedded system must participate in systematic memory management. On starting to compile the kernel, developers must ensure the system that how much memory this developing board has; while developing the application program, developers must consider distribution situation of memory and the size of the space that application program need to operate. In addition, because of adopting real memory manage tactics, user program, the kernel and other user programs being on an address space, developers should guarantee not to infringe the address space of other programs at the time of program development in order to make the procedure unlikely to destroy the normal work of the system, or cause other program to run abnormally. Therefore, the developers of the embedded system should be extremely careful of the operation of some memory in the software.

Due to management mechanism of memory of the embedded system, the embedded operating system adopts the form of static chaining to user program. Of the embedded system, application program and operating system kernel code create a binary scale image file to run by compiling and linking.

3 Interface of Embedded System Grid

The embedded system grid provides a single integrated system that receives, processes, displays, maintains, and assesses the controlled devices information. Enables users to plan, direct, and control devices under the user's operational control.

To achieve sharing embedded system resource over grid, it is in the first place to release embedded resource to grid by means of grid interface. According to interface and grid service definition that OGSA offers[3,4], the general grid service structure sketch is shown in Figure 1. In the sketch, that while OGSA defines a variety of behaviors and associated interfaces, all but one of these interfaces (GridService) are optional[5]. To realize Web Serve that OGSI offers, a general mechanism is needed to enable the applicant of service to inquest and upgrade the state data of service instances and receive corresponding notice when these data change. Presently, service Data is used to express these data that embody state of serve instance [6]. In order to set up grid service joint, the key problem lies in defining one or more essential interfaces for various kinds of resource; each interface is able to realize a certain operation and describe essential attribute of resource and service data. Adopt the description which carry on the net service joint of WSDL (Web Service Description Language) [7] is adopted to describe grid service joint. In the embedded system grid, the following elements can be defined with WSDL:

-Name, name of the embedded system equipment resource

-Types, type of grid service of embedded resource;

-Properties, to understand the detail of the joint when using and configuring grid service joint;

-Port Type, ports that each service interface supports, namely, interface for gird service joints and all abstract set of supported operation;

-Operation, Concrete operation that each interface supports

-Message, the data structure that this service joint supports;

-Port, individual port of serve and visit in which protocol / data form and Web visit address are assembled

-Binding, binding for concrete protocol of specific port and data form criterion;

-Service, the visit set of relevant services.

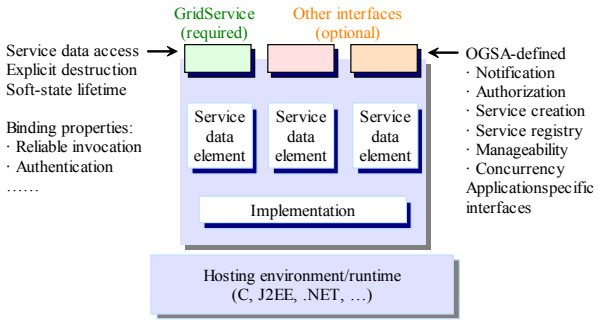


Figure. 1 OGSA Grid service

The service consists of data elements and various required and optional interfaces, with potential instantiation via different implementations, possibly in different hosting environments.

The course for service arrangement is also the course of releasing services, at this moment, according to the service information including IP address; port; and disposition position et al, grid will allot a Grid Service Handle (GSH) which is unique in the earth.

Having been encapsulated, embedded service must be registered in the center for registration service so as to be a real grid service for users. The registration service as the middleware of the grid is very important in the whole OGSA framework; after being registered in the above registration center, all grid services could be added to classified service warehouse and be assigned a unique GSH in the world so as to realize the localization for service in grid platform. Meanwhile, we can release the grid service to registration center for Universal Description, Discovery, and Integration (UDDI), UDDI enables us to release and search for the transactions of business partners and their grid services.

4 Sharing Of Embedded System Resource

Resource management devices of the gird can manage scattered various resources in order to enable multiple applicants for resources to share and use the same resource in the grid and enable a resource applicant can use multiple resources in the grid at the same time or successively according to the needs of transactions, instead of paying out extra work. Resource management offers the simple interfaces, which can visit resources to users; as definite details of resources using are hidden, what users see is an abstract resource. Resource management device harmonizes resource sharing, not only supporting multiple applicants to use the same resource, but also supporting an applicant to use multiple resources. Resource management device will replace the applicant to use resources and establish the safe mechanism of using grid resources.

Grid resources have autonomy nature, control of resource management device over resources does not mean to deprive resource owners of control power, but act according to the resource owner's will. The control of resource management device over resources is only limited to the control that resource owners has authorized. So the control power of resources is still in resource owner's hands, management of resource management device is only the share function authorized by resource owner. Among the administrative system for resources, according to the difference of information flow route that three entities including resource applicants, resource intermediaries and provider of resource, the resource administrative system has three kinds of forms: straight line, broken line and triangle[8].

Type of straight-line. The applicant for resources puts forward the claim to the resource intermediary

person, the resource intermediary person looks for suitable resources and urges resources to work and provides service for users. The result that resources serve user is still returned to the applicant for resources by the resource intermediary person. The advantage of the straight-line shape is that the resource management device is responsible for satisfying the demands only if users put forward their claims. The users pay attention to service result and simple interface.

Type of broken line. The resource applicant puts forward resource request to the resource intermediary person, the resource intermediary person searches for suitable resources and offers resource identification and user interface of resources for users. Users organize the news and data, impel the resource to work, and obtain the service of resource offering according to returned information. In this case, user side needs to understand resource interface.

Type of triangle. The applicant for resources puts forward the claim to the resource intermediary person, the resource intermediary person looks for suitable resources and urges resources to work for users, and tell the resource to use what form to return service results to which address. Resources begin to work; once the service ends, service results are returned to applicants according to news that the intermediary person offers. In this case, if users send one request, they can obtain the service that resources offer and does not need to understand the interface of resources using.

Among three different structures, the function of resource management device is different too. In the straight line shape, resource management device is a tie between user and resource, namely, not only being responsible for matching user resources but also being responsible for the switching of the mutual information of users and resources; this kind resource management device with complicated functions is suitable for simple user interface. However, in the shape of broken line, resource management device is only responsible for the match between user resources, not caring about the real trade between the user and resource; this kind resource management device with simple functions is suitable when both users and resources need to consult. The triangle also has the shortcomings of the above two while having the advantage of two; this kind resource management device with complicated functions is suitable when the interfaces are simple and the output results are valued.

The most realistic problem that grid technology needs to solve is to harmonize resource-sharing in virtual organizations that are changing dynamically[9]. Under the distributed environment, relative independence of every entity leads to the dynamic attribute which can be reflected from the following such as dynamic and variable user quantity, dynamic and variable resource quantity, dynamic use of resource over grid et al. The grid manages rational distribution and dispatching of various mass resources among trans-organizations and management fields of the grid. The functions of management must support adaptability, malleability and expansibility of resources, allow mutual operations for systems with different management tactics while keeping resource autonomy of the website, distribute resources in coordination, have good performances such as fault tolerance and stability.

To grid computing system, the most basic problem is to obtain systematic structure, status information and resource status information in real time, carry on resource distribution to gird application by using the information. Some services mix together and become Resource Allocation Manager (GRMA)[10] in GT3 Globus, GRAM offers a simple interface for the long-range system for users, is responsible for resource request for long-range application, dispatching for long-range task, management for long-range task dispatching, et al., is responsible for analysis and disposal of Resource Specification Language (RSL) information, GRAM is the center task execution under the grid environment. Globus Architecture for Reservation and Allocation (GARA) expands the resource management of Golobus, introduces resource target and reserve mechanism, realize QoS[11] of application.

Metacomputing Directory Service (MDS) is the information service centre under grid computing environment and fulfills mainly discovery, registration

and inquiry of information under the grid environment, offers a true real-time dynamic reflection of grid computing environment[12]. MDS is based on LDAP, the information that MDS needs to deal with is mainly various description of resources including resources of the data, calculation resource et al, services and other entries under grid computing environment. The information provider offers the information of MDS mainly, including the key information providers, general information providers and self-defining information providers; the key information providers provide the key information and operation state of grid resource. The information provider registers the agreement through the soft-state registry agreement. The users or the s senior services search for or book the information through inquiry agreement. After identifying and analyzing information inquiry request safely, on the basis of request information type and its buffer memory situation, the information provider is invoked. The information that the information provider returns is gathered and filtered in order to get rid of the information that the customer does not need and return the result to the person who gives information inquiry including the users and the senior services.

The resource management model structure of the embedded system grid is shown as Figure 2. Create Reservation operate establishes resource reserve, the operation interacts with local resource management so as to insure that the quantity and quality of resource can be used at the beginning of request, and can be used within duration hoped. If resources can't get the assurance, Create Reservation will fail; if Create Reservation succeeds, returns reservation handle, which can supervise and control state of reservation. Cooperative reservation agency is responsible for discovering resource collection of end-to-end QoS, which is able to meet application request; cooperative reservation agency does not distribute resources, but only reserve resources. According to required QOS, cooperative reservation agency is transferred; reservation handle returned is passed to cooperative reservation agency.

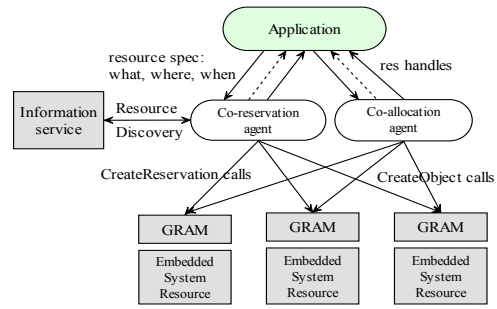


Figure 2 The Embedded System Grid GARA resource management architectures

5 Conclusions

The integration of embedded systems into an efficient, effective, embedded grid system is a challenging task. In this paper, the Characteristics of embedded system and application of grid technology are presented, and the basic attributes of embedded system resources are described in WSDL language, and the method of registering the embedded system resource at grid registration center is released. This paper also studies the main technologies that embedded system resources are put into the grid and become gird nodes and the method to manage embedded system resource by means of resource reserve of Globus and distribution frame. In the end, this paper proposes how to realize grid distribution and control management of embedded system resources and offers a new idea for intellectual equipment resource to be integrated into the grid. As the demand to share various intellectual equipment and large-scale valuable equipment increases constantly, and as grid technology grows up constantly, embedded systems will surely become sharing resources as new nodes of grid.

6 Acknowledgments

This work was supported by 973 Project under grant no: 2003CB317001 and Project 50775070 supported by NSFC, by Scientific Research Fund of Hunan Provincial Education Department under grant no: 07C272.

References

[1] Foster I , Kesselman C , Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal Supercomputer Applications , 2001 ,15 (3): 200～222

[2] Foster I , Kesselman C. The Grid: Blueprint for a New Computing Inf rast ructure. San Fransisco : Morgan Kaufmann , 1999

[3] GWD-I Open Grid Services Architecture. http: //forge.gridforum.org/ projects/ogsa-wg

[4] Foster I, Kesselman C, et al. Grid services for distributed system integration IEEE Computer, 2002,35(6): 37~46

[5] Foster I, Kesselman C, Nick J M. Steven Tuecke The Physiology of the Grid An Open Grid Services Architecture for Dist ributed Systems Integration. www.globus.org/ research/ papers/ogsa.pdf

[6] GWD-R(draft-ggf-ogsi-gridservice-23) Open Grid Services Infrastructure (OGSI) http: //www.ggf.org/ogsi-wg

[7] Web Services Description Language. http: //www.w3.org/ TR/2004 / WD-wsdl20-200408 03

[8] Xu Zhiwei, Feng Baiming, Li Wei.Grid Computing Technology. Electronic Industry Press, 2004, 5

[9] Karl Czajkowski, Steven Fitzgerald, Ian Foster, Carl Kesselman. Grid information services for distributed resource sharing. Proceeding of the 10th IEEE International Symposium on High-Performance Distributed Computing (HPDC-10), IEEE Press, 2001,(8)

[10] GT3 GRAM Architecture. http: //www-unix. globus.org/ developer/gram- architecture.html

[11] Foster I, Kesselman C, Lee C, et al. A distributed resource management architecture that supports advance reservations and co-allocation. Workshop on Quality of service, 1999

[12] Fitzgerald S, Foster I, Kesselman C, et al. A directory service for configuring high-performance distributed computing. 6th International Symposium on high performance distributed computing (HPDC'97), 1997

A Modeling Method of Manufacturing Resource Sharing Based on Knowledge Grid

Hong Wen Huifu Zhang Bo Gong Deshun Liu Anhua Chen

Hunan University of Science and Technology, Xiangtan, Hunan, 411201, China
Email: {wenhong74,huifuzhang}@yahoo.com.cn

Abstract

Manufacturing resources sharing based on grid had been investigated by many researchers. However, such a resources sharing has neither semantic support nor standard resource express model. This problem results in a lot of difficulties for virtual enterprise in manipulating manufacturing resources effectively. Knowledge grid proposes a Resource Space Model (RSM). By using RSM, manipulator can specify, share and manage versatile web resources in a universal resource view. Based on RSM, this paper presents a method to construct a manufacturing resource space model (MRSM), and then gives an applied example. At end of this paper, we design a framework for the research of manufacturing resources sharing.

Keywords: Manufacturing Resource Sharing; Virtual Enterprise; Knowledge Grid; RSM; MRSM

1 Introduction

Grid can be regarded as the next generation Internet [1,2]. Researcher had proposed a lot of application grids, such as data grid [3], DOE science grid [4] and earth system grid [5] and so on. Manufacturing grid (MG) is an integrated supporting environment which can be used to share and integrate manufacturing resources in enterprise and social. By using Internet, grid technology and other advanced computer and information technologies, MG can effectively organize all kinds of resources separated in different regions, enterprises, organizations, and individuals to support the design, manufacture and services of products. With the help of services provided by MG, users can obtain various manufacturing services conveniently as obtaining information from Internet [6].

The key problem of MG is how to provide a secure and high efficiency scheme to meet the customers' resource request in a transparent way. Nowadays, the researches of MG mostly focus on algorithms of resource deployment, search, management and scheduling. These researches rarely refer to the key problems of grid such as uniform resource analysis and classification, RSM design and operation etc. In a word, MG lacks the support of basic theory and methodology. Knowledge grid (KG) [7,8] proposed a set of concepts, rules and schemas of RSM which can be used to design MG resource space model. And then we can also utilize the resource using mechanism (RUM) of KG to implement uniform manufacturing resource distribution, management and search. The main aim of this paper is to propose a RSM based manufacturing resource sharing framework for MG. This framework includes a manufacturing resource space model and a manufacturing resource sharing architecture based on RSM and RUM cooperative mechanism, which provides a useful theory for manufacturing resource sharing research.

2 The Architecture and Application of MG

We can depict MG as a hierarchical structure which can be divided into three layers. As shown in Figure 1, the functions of each layer can be described as follows:

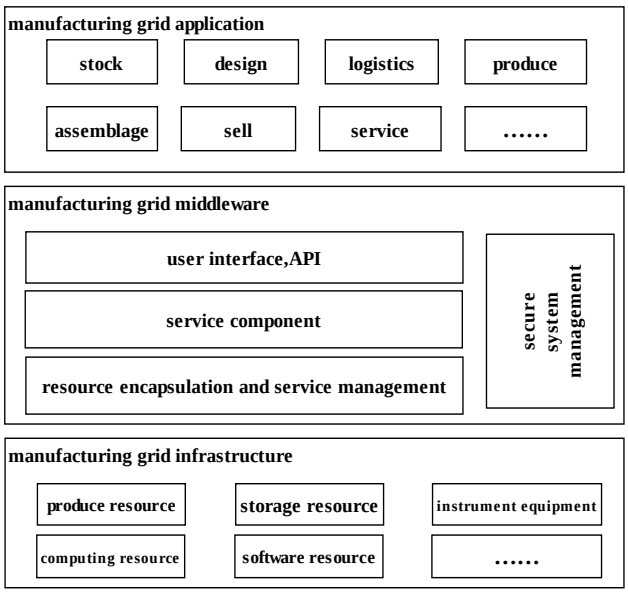


Figure 1 Architecture of manufacturing grid

1. MG infrastructure layer: it is the basic structure of MG, which supports all kinds of heterogeneous resources of all life-cycle products. All of the resources must accord with the request of MG interface to connect to the MG.

2. MG middleware layer: this layer is to manage MG resources and provide user's interface and encapsulate resources.

3. MG application layer: it ought to provide all the applications of whole lifecycle product, such as material stocking, products assemblage, logistics management and service after sell etc. All of the applications ought to use a uniform interface and must make the resources which integrated on MG been used transparently.

Virtual enterprise (VE) that using grid technology is essentially different from traditional VE. The components of traditional VE must be enterprises, but the components of VE which using grid technology are not only enterprises or departments of enterprises, but also resource nodes such as computers, storages etc. These resource nodes may belong to an enterprise, or a research institution or an individual [9].

Figure 2 present VE I and VE II based on grid and their components---enterprise A, E and F(they are all manufacturing enterprises and possess manufacturing resources), enterprise B(possess computing resources), enterprise C(possess storage resources) and enterprise D(possess instrument equipment resources).

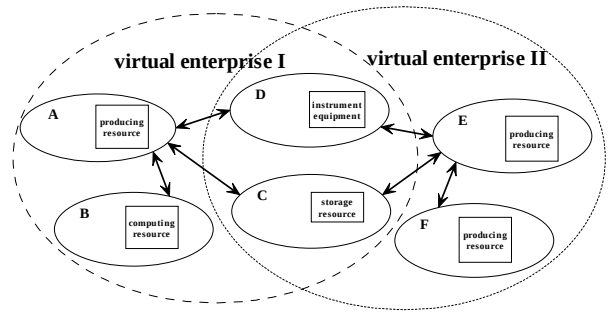


Figure 2 An example of grid virtual enterprise

The resources of enterprises A~E are all encapsulated into the nodes of MG, so the MG components may access the resources nodes each other. In other words, these nodes of MG are the enterprises' resources, which include computing resources (high performance computer, PC and all kinds of software etc.), manufacturing resources (numerical control machine tool, production line control machine etc), storage resources (storage equipment and database etc) and instrument equipments (testing instruments, simulation instruments and test equipment etc)Enterprise A is a aircraft manufacturing enterprise, because the architecture of aircraft is very complex and the design procedure is strict, so A log on B and use its high performance computer to tackle the raw data. The raw data and the calculated data stored in storage resources owned by C enterprises, if necessary, A can read or store data through high-speed network; In addition, since the simulation equipment the task want to use is very expensive and owned by D enterprise, so A enterprise issue requests of sharing equipment resources to enterprise D. In this situation, enterprise A, B, C and D constitute a virtual enterprise I, all of the entities in VE I complete the task with the manner of collaborative design and sharing resources. For the same reason, the high-grade car manufacturing enterprises E also need to resort to the resources of enterprise D, C and F, then, enterprise E, D, C, F form a virtual enterprise II.

It is shown from the above analysis that the

resource sharing of virtual enterprises is no longer the simple document exchange. It includes remote accessing, controlling and using of computing resources, storage resources, manufacturing resources and equipment devices. The RSM of knowledge grid can help to create a uniform, standard MG RSM and using RUM mechanism of KG can help to operate the fundamental resources of MG high efficiently[10,11,12].

3 Manufacturing Resource Space Model Design

Manufacturing resource space model design will follow four steps: resource analysis, top-down resources partition, two-dimensional space design and join [10].

3.1 Manufacturing resource analysis

In manufacturing resources analysis, how to construct the ontology of manufacturing resources dictionary (RD) is a key problem. The target of ontology is to capture knowledge of relevant domain, then provide a common understanding and identify the common vocabulary for this domain knowledge, in addition, give a clear definition of these terms and their interrelationship from different levels.

Manufacturing resources are infinite, so an exhaustive approach is not appropriate to construct RD, and then constructing the ontology of RD is vitally important. In essence, Analysis of resources is ontology analysis. Ontology analysis is to identify the structure of knowledge, for a given domain, its ontology constitutes any core knowledge representation system. Therefore, the first step of design an effective resource expressing systems and RD is to make a good analysis for ontology in a given domain, otherwise will lead to an inconsistent RD.

After finish the job of building resources dictionary, we can use XML query language, or other query language (SQL) to manage resources dictionary. The RD of manufacturing resources is similar to the RD of relational database, but it has two different characteristics to the relational database RD: in the manufacturing resource space model, First, RD is used to construct multi-dimensional model of space resources; Second, the resources of RD are determined uniformly, minimized the redundant between resources [10].

3.2 Manufacturing resource partition

Manufacturing resources is almost infinite, then, the partition of resources can be focused on the partition of the resources ontology. Top-down resource partition is a direct and effective solution. Each designer may have his own method of resource partition, so a uniform viewpoint on resource partition is needed. The first step to reach a common viewpoint is to reach a universal top-level resource partition agreement [10]. For manufacturing resources (MR), it is a key point to reasonably classify the manufacturing resources under the universal level. Figure 3 shows an example of how to partition manufacturing resource.

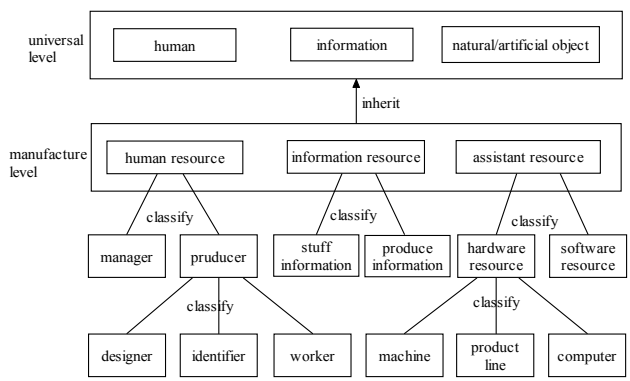


Figure 3 An example of manufacturing resources partition

The top-level resources are universal level resources, which can be classified as three independent categories: human, information and natural/artificial object. The manufacture level is under the universal level, it inherit all properties of universal level and has its own categories, each category can be refined top-down until the category is small enough to serve for domain applications. By using this partition solution based on RSM, we can take full advantages of a series of basic theory, rules, and the methods of RSM to manage and operate manufacturing resources.

3.3 Design 2D manufacturing resource spaces

In section 3.2, we have introduced a solution of top-down manufacturing resource partition, its purpose is to construct a multi-dimensional manufacturing resource space model (MRSM), and then use the principles, rules and methods of RSM to unify and standardize this MRSM, so, we can use the RUM of knowledge grid to manage, operate and edit the resources in MRSM, which make it possible to share the manufacturing resources.

Compare with high-dimensional spaces, people can better manage two-dimensional spaces. So for the MRSM shown in Figure 3, the first step we can do is to design a set of two-dimensional resource spaces then consider joining them into an entire resource space. The design process will keep to the following principles [10]: determine the number of manufacturing space resources, determine the axes name, determine the first-level coordinate names, determine coordinate hierarchies, Check independency between coordinates, Check orthogonal relationship between axes.

As show in Figure 4, following the rules above, we design two 2D spaces for the MRSM.

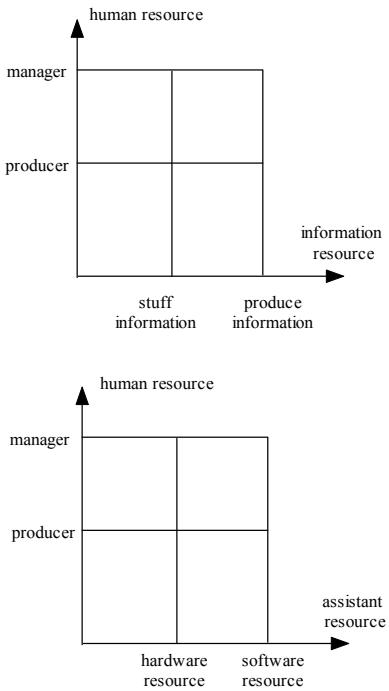


Figure 4 Construct 2D manufacturing resources space

3.4 Join between manufacturing resources spaces

The target of 2D space design is to construct a universal resource view, so we ought to join the two 2D resource spaces into a 3D space. For a good MRSM, it ought possess three characteristics [10]: 1)in the same axis, there are not two coordinates have the same name; 2)any two coordinates are independent; 3)any two axes are orthogonal with each other. In order to construct a complete MRSM, we need to join the two 2D models shown in Figure 4 into a 3D MRSM as shown in Figure 5

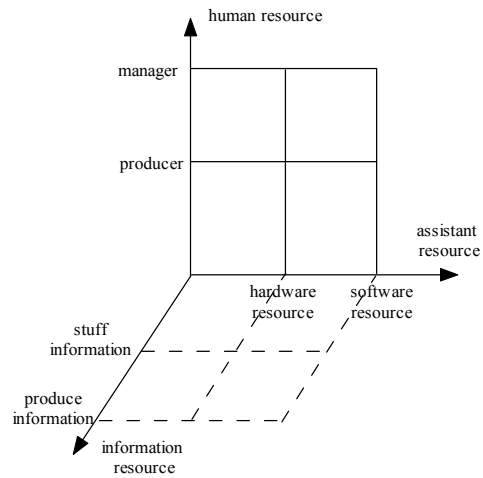


Figure 5 A 3D space joined by two 2D spaces

The join operation uses the human resource axis as the common axis, generate a new two-dimensional space: (information resource, assistant resource). But the new space will no longer meet the request of the third characteristic above. So we must to reinterpret the coordinates to make the two axes in new space orthogonal each other. For this new space, we will reinterpret coordinates of information resource axis based on the assistant resource axis, for example, we can reinterpret the (stuff information, produce information) as (state information, function information) in information resource axis, in this way, the coordinates of information resource axis will be the function of two sub-spaces: (human resource, information resource) and (information resource, assistant resource), i.e.: information resource ((human resource, information resource))=(stuff information, produce information) and

information resource((information resource, assistant resource))=(state information, function information).

Thus, we complete a manufacturing resource space model by using the concepts, rules and methods of knowledge grid RSM. The new model is comparatively simple, unable to describe the true manufacturing resources, which is used to elaborate a scheme of how to construct a MRSM. There are two other aspects of model constructing: one is design strategy, include two kinds: using two-dimensional reference model (RS=(category, level)) and using abstraction and analogy strategy; the other is the design tools, include independent checking tool and orthogonal checking tool.

4 A Manufacturing Resource Sharing Framework

The resource space grid has another important mechanism: RUM. RUM provides not only a resource browser based on a built-in resource operation language (ROL) for end user, but also a ROL-based programming environment. RSM and RUM cooperative mechanism can be used to construct the resources discovering, search and remote accessing. This will be a key point to effectively solve the problem of manufacturing cooperation and resources sharing in heterogeneous and distributed environment. Meanwhile, this cooperative mechanism will focus on the information dynamic sharing of relevant design resources, various processes and knowledge resources of product whole lifecycle.

In this manufacturing resources sharing framework, resources information is the basis, knowledge is the kernel of effectively managing and operating resource, services is the products provided to the users and developers, it is also the purpose of constructing the cooperative architecture.

Now we construct a manufacturing resource sharing framework, shown in Figure 6, it merge the MRSM and MRUM, under the cooperative of MRSM and MRUM, the resources in the whole manufacturing grid can be browsed, operated and shared conveniently. The whole system include four parts: the bottom layer is resources entity layer, include all kinds of resources entities; the second is the manufacturing resources management platform based on knowledge grid, include four sub-layers: knowledge storing, knowledge expressing, knowledge management and reasoning sub-layers, its function is to organize and manage all the manufacturing resources; the third layer is interface layer, provide programming interface and system authorization etc; the topmost layer is resource operation layer, users can operate various resources conveniently through the resources browser, meanwhile, the developer can use ROL to program through the API.

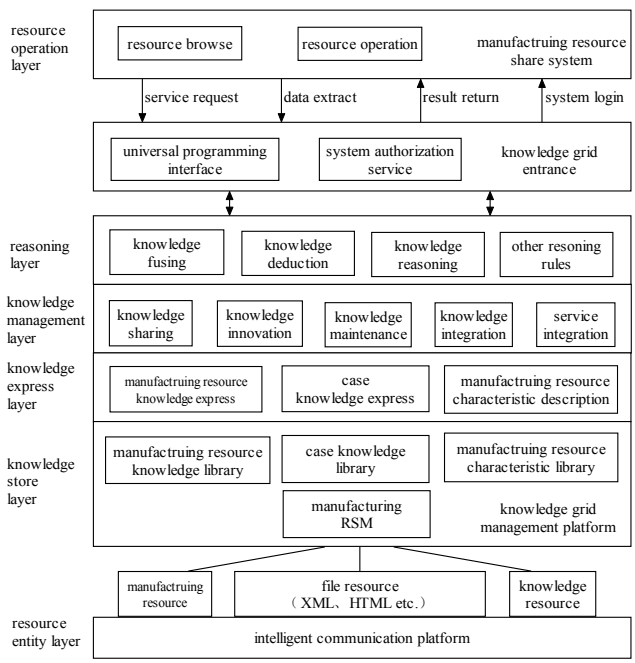


Figure 6 Manufacturing resources sharing framework

5 Conclusion

With the rapid development of Internet and other relevant technologies such as grid technology, manufacturing grid becomes a hot application research now. However, how to make computers understanding the semantics of manufacturing resources and make manufacturing resources management, operation and search more efficient is difficult. This paper proposes a set of methods based on knowledge grid to construct

MRSM and then verify the methods through constructing a universal MRSM. We also construct a manufacturing resource sharing framework based on RSM and RUM cooperative mechanism, which provide a useful theory for manufacturing resource sharing research.

6 Acknowledgments

This work was supported by 973 Project under grant no: 2003CB317001 and Project 50775070 supported by NSFC, by Scientific Research Fund of Hunan Provincial Education Department under grant no: 07C272 and 07C273 and by China High Tech. plan, 863/CIMS, under grant no: 2006AA04Z152.

References

[1] I. Foster and C. Kesselman, “The Grid: Blueprint for a New Computing Infrastructure”, Morgan Kaufmann, San Fransisco, CA, 1999

[2] I. Foster, "Internet Computing and the Emerging Grid" NatureWeb Matters, 7 December 2000

[3] http: //eu-datagrid.web.cern.ch/eu-datagrid/

[4] http: //doesciencegrid.org/

[5] http: //www.earthsystemgrid.org/

[6] FAN Yushun, ZHAO Dazhe and ZHANG Liqing,et al, “Manufacturing grid: needs, concept and architecture” , GCC2003, LNCS3032, Springer-Verlag, Berlin Heidelberg, 2004, PP.653-656

[7] H. Zhuge, “The Knowledge Grid and Its Methodology”, 1st International Conference on Semantics, Knowledge and Grid, Nov 27-29, 2005, Beijing, China

[8] H. Zhuge, “Semantic Grid: Scientific Issues, Infrastructure, and Methodology”, Communications of the ACM. 48 (4) (2005), PP.117-119

[9] Li Chen, Hong Deng, Qianni Deng and Zhenyu Wu, “A Research on the Framework of Grid Manufacturing. Grid and Cooperative Computing”, Second International Workshop, GCC 2003, 2003.12, PP.19-25

[10] H.Zhuge, “Resource Space Model, Its Design Method and Applications”, Journal of Systems and Software, 72 (1) (2004), PP.71-81

[11] H.Zhuge, “Resource Space Grid: Model, Method and Platform”, Concurrency and Computation: Practice and Experience, 2004, 16,PP.1385-1413

[12] H.Zhuge, E.Yao, Y.Xing, and J.Liu, “Extended Normal Form Theory of Resource Space Model”, Future Generation Computer Systems, 21 (1) (2005), PP.189-198

The Study on a Method of De-Normalization and Synthesis

Jing Lin

Department of Computer and Information Engineering, Wuhan Polytechnic University, Wuhan, 430023,China
Email: arlovegirls@sina.com

Abstract

This paper presents a method of converting relational database into XML documents , that is de-normalization and synthesis. Both the principle and implementation of the method are discussed. This method involves three steps. Firstly, we revert the normal relation schema to original unnormal one without the functional dependency, transfer functional dependency, multivalued dependency and join dependency. These unnormal relation tables are called connection tables. Then the connection tables are changed to XML documents and XML tree. In the end, all those XML documents are merged. According to the principle, the method is implemented in the case of bank loan. In this example, the data demonstrate the validity and feasibility of the method.

Keywords: Relation Schema; De-normalization; XML; XML DOM; Connection Table

1 Introduction

Based on the traditional mapping schema between XML document and relation database, this paper improves the conception of de-normalization and XML DOM synthesis in literature[1] , presents a new method to convert relational database into XML, and implements it in the case of bank loan. In the end, we compare the new method with the old one and conclude the merit of the new one.

2 Principle of the Method

2.1 Definition

The method is a de-normalization of the relational database and a synthesis of the XML DOM. The de-normalization is a reversion of the normal relation schema to the original redundant one without the functional dependency, transfer functional dependency, multivalued dependency and join dependency. These original unnormal tables are called connection tables, which are to be converted into the XML DOM. Finally all XML DOM are merged to a synthetical one.

2.2 Workflow

By the method, the steps of mapping a relational database into XML documents are as follows:

(1) The relation schema is de-normalized to connection tables and the result is saved.

(2) The connection tables are converted into XML documents.

(3) The relation schema of the connection tables is converted into XML tree.

After step1, the dependence relations among tables are reverted to the originality. The rules of mapping the dependence relations into XML documents in step2 and step3 are as follows: In the connection tables, the functional dependency is mapped into the relation between the element and son element; the transfer functional dependency is mapped into the relation among the element, son element and grandson element; the multivalued dependency is mapped into the relation among the element and multi-son elements; the join dependency is mapped into the group element. According to these rules, the dependence relations among tables are converted into XML DOM and XML tree.

(4) According to the chosen root element and the chosen main tree, all XML trees are merged to the synthetic XML tree.

(5) According to XML tree, we fill all the XML DOM data produced in step2 into the attributes of the synthetic XML document.

3 Implementation of the method

In this part, we make an experiment to demonstrate the validity of the method. In this experiment, the method is applied to bank loan. And the programming language is Java2 and the RDBMS is Microsoft Access.

3.1 Relation schema in the case of bank loan

There are seven relation tables in this bank loan instance. They are:

1. Customer-Security(Customer, Security)
2. Customer-Branch(Customer, Branch)
3. Security-Loan(Security, Loan, Maturity_date)
4. Loan-Index(Loan, Index)
5. Loan-Customer(Loan, Customer)
6. Index-Rate(Index, Rate)
7. Customer-Account(Customer, Account)

In each relation schema above, the primary key is underlined. And the relation schema R is classified as follows:

PR1: All prime attributes of relation R don't contain the prime attributes of other relations.

PR2: The prime attributes of relation R contain the prime attributes of other relations.

SR: Relation R has more than one attribute in primary key. All prime attributes of relation R are composed of the prime attributes of other relations.

Each field of the relation schema is classified as follows:

KAP: The attribute is not only one of prime attributes of relation R, but also one of prime attributes of other relations.

KAG: The attribute is one of prime attributes of relation R, but not one of prime attributes of other relations.

FKA: The attribute is not only one of nonprime attributes of relation R, but also the foreign key.

NKA: The attribute is one of nonprime attributes of relation R, but not the foreign key.

3.2 Implementation process

According principles of the method, the steps of this instance are as follow:

(1) The relation schema is de-normalized to connection tables.

Dependence relations among tables are found out via the classification of the relations and fields, then the normal relations are de-normalized to connection tables.

There is a join dependency*(Customer, Security, Loan) among the relations of SR type in Table1. So the three SR relations are de-normalized to connection table R1(Customer, Security, Loan, maturity_date).

There is a multivalued dependency (Customer－>－>Account,Customer －>－>Branch) among the relations of PR2 type in Table2. So the two PR2 relations are de-normalized to connection table R2(Customer, Account, Branch).

There is a transfer functional dependency (Loan－>Index－>Rate) among the relations of PR1 type in Table3. So the two PR1 relations are de-normalized to connection table R3(Loan, Index, Rate).

(2) The connection tables are converted into XML documents and XML tree.

When converting the connection tables into XML DOM, we need not save each XML DOM into .XML file because XML DOM are merged directly in memory. In order to test whether these three XML DOM in memory are correct or not, the .XML files are created temporarily.

i) The connection table R1(Customer, Security, Loan, maturity_date)which comes from join dependency is converted into XML DOM1.

Since XML DOM1 is the main tree in the following operation, BANK is chosen as the root element for the XML document. The relevant XML tree is showed in Figure 1.

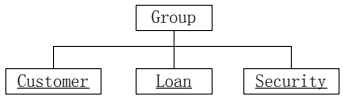


Figure 1 Tree from join dependency

ii)The connection table R2(Customer, Account,Branch) which comes from multivalued dependency is converted into XML DOM2.The relevant XML tree is showed in Figure 2.

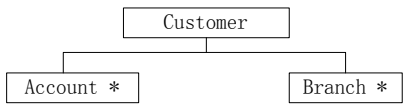


Figure 2 Tree from multivalued dependency

iii)The connection table R3(Loan,Index,Rate) which comes from transfer functional dependency is converted into XML DOM3. The XML tree is showed in Figure 3.

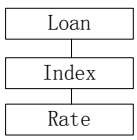


Figure 3 Tree from transfer functional dependency

(3) After a main tree is chosen, three XML DOM trees are merged to the synthetic XML tree.

Among all the XML DOM, a main tree is chosen for synthesis. Here XML DOM1 is chosen as the main tree. For each node in XML DOM1, we should search for the node with same element name and element value in the other two documents. Then the three XML DOM are merged to a synthetic XML DOM. And the synthetic XML DOM is saved into .XML file. In detail, firstly the main tree and XML DOM3 are merged to get intermediate synthetic XML DOM. Then, as the main tree, the intermediate synthetic XML DOM is merged with XML DOM2 to get the final synthetic XML DOM which is to be written into .XML file. The file is showed in Figure 5 and the relevant XML tree is showed in Figure 4.

In order to merge the XML DOM, we programme a recursive function which is called mergeDOM. It is:

Void mergeDOM(Element mainDOMRoot, Element beenmegedRoot)

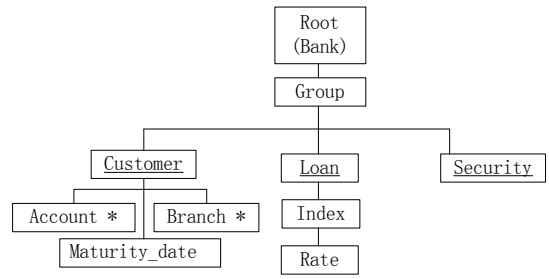


Figure 4 The synthetic XML tree in Loan instance

Figure 5 The final synthetic XML DOM

In this function, the parameter mainDOMRoot is the root element in the main tree and the parameter beenmegedRoot is the root element in the search XML tree.

3.3 Performance analysis of the algorithm

In order to compare the old method with the new one, we record the runtime of this program. They are DOM generate time, Write file time and DOM merge time. In the DOM generate time, the connect table is converted into the XML DOM. In the Write file time, the XML DOM is saved into .XML file. In the DOM merge time, all XML DOM are merged.

In the experiment, we use three sets of data to test the three runtime. 200 items data, 500 items data and 1000 items data are separately inserted to each relation table. Each set of data runs 4 times to get the average runtime. The data are as follows in table 4.

Table 1 Relation schema of SR type

Relation Name	Relation Type	Primary Key	KAP	KAG	FKA	NKA
Customer-Security	SR	Customer, Security	Customer, Security			
Security-Loan	SR	Security, Loan	Security, Loan			Maturity_date
Loan-Customer	SR	Loan, Customer	Loan, Customer			

Table 2 Relation schema of PR2 type

Relation Name	Relation Type	Primary Key	KAP	KAG	FKA	NKA
Customer-Branch	PR2	Customer, Branch	Customer, Branch			
Customer-Account	PR2	Customer,Account	Customer,Account			

Table 3 Relation schema of PR1 type

Relation Name	Relation Type	Primary Key	KAP	KAG	FKA	NKA
Loan-Index	PR1	Loan	Customer, Branch		Index	
Index-Rate	PR1	Index	Customer,Account			Rate

Table 4 The data of algorithm in this paper

	200 data set	500 data set	1000 data set
DOM generate time (s)	1.0317	1.5445	2.015
DOM merge time (s)	1.412	10.014	44.193
Write file time (s)	0.0325	0.09525	0.17075
Total (s)	2.4762	11.65375	46.37875

The data used in the original algorithm in literature[1] are as follows in table 5.

Table 5 The data of algorithm in literature [1]

	200 data set	500 data set	1000 data set
DOM generate time (s)	1.23675	1.6225	2.020
DOM merge time (s)	3.2145	19.67825	78.643
Write file time (s)	0.3255	0.676	1.3545
Total (s)	4.76775	21.97675	82.0175

From the two tables above, the conclusion is that the DOM generate time and the Write file time is less. And there is not a big difference between this paper and literature[1], which is showed in the following Figure 6/ Figure 7.

The merge time from XML DOM to the synthetic XML DOM is the most. So it is also the key of the method, which is showed in Figure 8.

From the pictures we can see there are obvious advantages of the algorithm in this paper. Not only the DOM merge time reduces 50.43%, but also the slope from 200 data set to 500 data set and from 500 data set to 1000 data set decreases. Therefore the algorithm in this paper is much better than that in the literature [1].

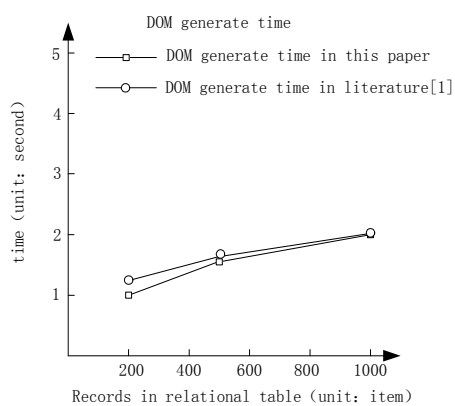


Figure 6 The comparison of XML DOM generate time

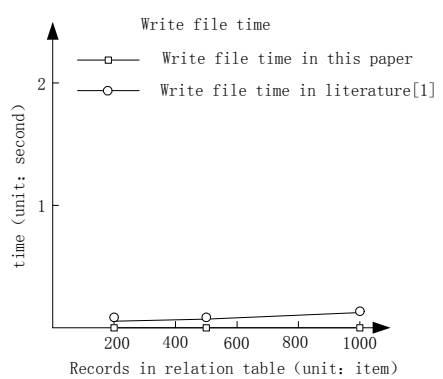


Figure 7 The comparison of Write file time

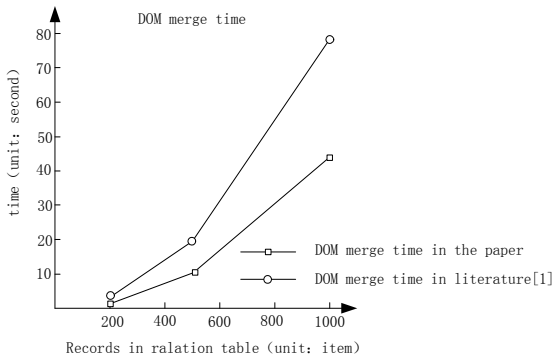


Figure 8 The comparison of XML DOM merge time

4 Conclusion

Compared to literature [1] , the method of de-normalization and synthesis is improved in this paper. Different implementation ways and the relevant algorithms are developed, which makes the method of de-normalization and synthesis simpler and more efficient. The advantages of the method are as follows:

(1) In course of converting connection tables into XML DOM, the data in relation tables are saved by the node values of the elements, and the names of node are addressed as the field names in the relation table. Compared to the method saving data by element attribute value and element value in literature[1], the implementation in this paper makes the node type in the XML tree consistent and simple, so the program is easier to design.

(2) In order to save the data by element attribute value and element value as discussed in literature[1], different types of node are created according to different status during the conversion. For example, such types as Document, Element, Attr and Text are all built. So when all XML DOM are merged into the synthetic XML DOM, each type of node is estimated repeatedly. According to different types of node, different operations are carried, so it takes more time. While in the paper, only type of node, i.e. Element type is used in the program. Thus so much time is saved when program is proceeding.

(3) The algorithms are different. In literature[1], the algorithm involves repeated structures and simple functions. While recursive algorithm is used in this paper.

References

[1] J.Font, H.K.Wong,Z.Cheng, “Converting relational database into XML documents with DOM”, Elsevier Science B.V, No.2, 2003

[2] Sudhanshu Sipani, Kunal Verma, John A.Miller, Boanerges Aleman-Meza,“Designing a high-performance database engine for the ‘Db4XML’ native XML database system”, Elservier, No.3,2004

[3] *Chiyoung Soe, Sang-Won Lee, Hyoung-Joo Kim. “An efficient inverted index technique for XML documents using RDBMS”. Elsevier Science B.V, No.6,2002*

[4] *Xu Guangmei,Cheng Gengguo,Wu Aihua, “Design and Implementation of XML-Based Database Middleware”, Computer Engineering and Design,Vol.25,No.2,2004*

[5] *Wang shixian, “Research on Mapping XML to Relational Database”,Computer and digital Engineering, No.6,2005*

[6] Abraham Silberschatz, Henry F. Korth, S. Sudarshan, “Database System Concepts”, Beijing: China Machine Press, 2006

[7] Ann Navapro, Chuck White, Linda Burman,“Mastering XML”, Beijing: Publishing House of Electronics Industry, 2000

[8] Natanya Pitts, “XML Black Book”, Beijing: China Machine Press, 2002

[9] Fabio Arciniegas, “XML Developer’s Guide”, Bejing: Qinghua University Press,2003

[10] Charles F.Goldfarb, Paul Prescod, “The XML Handbook”, Bejing: Qinghua University Press,1999

Identification for Stephania Tetrandra S. Moore and Stephania Cepharantha Hayata by Wavelet Transform and BP Neural Network*

Changjiang Zhang[1] Min Hu[1] Cungui Cheng[2]

1 College of Mathematics, Physics and Information Engineering, Zhejiang Normal University
Jinhua, Zhejiang , 321004, China
Email: zcj74922@zjnu.cn

2 College of Chemistry and Life Science, Zhejiang Normal University , Jinhua, Zhejiang , 321004, China
Email: ccg@zjnu.cn

Abstract

Horizontal attenuation total reflection–Fourier transform infrared spectroscopy (HATR-FTIR) is used to measure the FTIR of *Stephania tetrandra* S. Moore and *Stephania cepharantha* Hayata. Because they belong to the same family and the same genus Chinese traditional medicinal materials, their chemical components are very similar. In order to extrude the difference between them, continuous wavelet transform (CWT) is used to decompose their FTIRs. Three main scales are selected as the feature extracting space in the CWT domain. According the distribution of FTIR of theirs, three feature regions are determined at every spectra band at selected three scales in the CWT domain. Thus nine feature parameters form the feature vector. The feature vector is input to the BP neural network (BPNN) to train so as to accurately classify the *Stephania tetrandra* S. Moore and *Stephania cepharantha* Hayata. 128 couples of FTIR are used to train and test the proposed method, where 78 couples of data are used as training samples and 50 couples of data are used as testing samples. Experimental results show that the accurate recognition rate between *Stephania tetrandra* S. Moore and *Stephania cepharantha* Hayata is respectively 99.6% and 99.8% by using the proposed method.

Keywords: Fourier transform infrared spectroscopy; Continuous wavelet transform; BP neural network; Stephania tetrandra S. Moore; Stephania cepharantha Hayata

1 Introduction

The Fourier transform infrared spectroscopy method is a very common analysis tool with high sensitivity, resolution and fast speed, which has been widely used in the identification of Chinese traditional medicinal materials. Because the traditional Chinese medicinal material is a kind of compound, the directly measured infrared spectrum is the superposition of all the infrared spectrums. Therefore if general analysis method is used, which will greatly depend on the experience. Wavelet transform is a powerful mathematical tool for signal processing in recent years.

The wavelet transform is being used in chemistry and its related domain in recent years [1-8]. Ehrentreich, F. pointed out that the wavelet transform has been established with the Fourier transform as a data-processing method in analytical chemistry [1]. Liu, Y. et al. formulated fast wavelet-based adaptive collocation method for heat and mass transfer problems involving a steep moving profile of the dependent variable [2]. Shao, Limin et al. introduced the wavelet transform its applications in respect of photoacoustic

* The paper is funded by the China Zhejiang Province Natural Science Grant (Y506203).

spectroscopy, EXAFS spectrum, NMR analysis, and Raman spectrum [3]. Iyengar, S.S. et al. presented rigorous analysis of the primitive Gaussian basis sets used in the electronic structure theory [4]. This leads to fundamental connections between Gaussian basis functions and the wavelet theory of multiresolution analysis. This result will be invaluable in the use of atom-centered Gaussian functions for ab initio molecular dynamics studies using Born-Oppenheimer and atom-centered density matrix propagation. Ying, Y.B. et al. showed a typical example (apple NIR spectra) how wavelet transforms could be used in order to extract quantitative information [5]. The sugar content of intact apple was measured by NIRS and analyzed by wavelet transform. The results show that the spectra treated with wavelet transform indicate more effectively the relationship with sugar content in intact apple. Chen, Jun et al. described a wavelet-based method for analysis of images obtained in heterogeneous polymerization [6].

Some researchers also used continuous wavelet transform to analyze the signal of chemistry [7]. For example, Lei Nie et al. proposed a novel method of calculating approximate derivative of signals in analytical chemistry by using the continuous wavelet transform (CWT) [7]. Additionally, the approximate second derivative evaluated via the CWT method can be used to determine the peak potentials of the overlapping square wave voltammogram (SWV) of Cd (II) and In (III).

In recent years, some researchers also combine the wavelet transform with other some intelligent technique to analyze the signal of chemistry [8-10]. For example, Khayamian, T. et al. developed a wavelet neural network (WNN) model in quantitative structure property relationship (QSPR) for predicting solubility of 25 anthraquinone dyes in supercritical carbon dioxide over a wide range of pressures (70-770bar) and temperatures (291-423K) [8]. Analysis of synovial fluid by infrared (IR) clinical chemistry requires expert interpretation and is susceptible to subjective error. The application of automated pattern recognition (APR) may enhance the utility of IR analysis. Jie Cui et al. described an automated pattern recognition (APR) method based on the fuzzy C-means cluster adaptive wavelet (FCMC-AW) algorithm [9]. Piotrowski, P.L. et al. developed a computational approach for performing efficient and reasonably accurate toxicity evaluation and prediction [10]. The approach is based on computational neural networks linked to modern computational chemistry and wavelet methods.

The advantages of the continuous wavelet transform (CWT) in the singularity detection of a signal are obvious. Compared with the discrete wavelet transform, it can detect the faint signal changes, which cannot be well implemented by discrete wavelet transform [11]. Artificial neural network can learn and train the information samples so that it will have similar memories of the human brain, identification capabilities and the implementation of various information processing functions. It has good self-learning, adaptive, associative memory, parallel processing and nonlinear conversion capabilities, which avoids complicated mathematical derivation. Even in the sample of the defect and parameter drift circumstances, the output can guarantee to be stable, thus it facilitates the theoretical analysis [11].

Stephania tetrandra S. Moore is a kind of commonly used Chinese traditional medicinal material, which can cure dropsy and ache. Because of much clinical usage, as well as proprietary Chinese traditional medicine preparations and chemical composition of extraction, and so on, this results in shortage of medicine, prices rising and chaos of varieties. Recently we discover that some Stephania cepharantha Hayata's roots have been mixed into *Stephania tetrandra* S. Moore. This paper uses HATR - FTIR method determines samples of *Stephania tetrandra* S. Moore and Stephania cepharantha Hayata directly. As *Stephania tetrandra* S. Moore and Stephania cepharantha Hayata belongs to sibling species, which contain the similar chemical composition, thus FTIR spectrums of theirs are also similar. If we use only FTIR spectra to identify them, this will be very difficult.

Therefore, based on our previous work [12-13], we use the CWT to extract the features of both of the FTIR spectra. Then we use BP neural network to efficiently identify them. Experimental results show that this obtains a good result.

2 Apparatus and Materials

2.1 Apparatus

A Nicolet (Madison, WI, USA) NEXUS 670 TTIR Spectrometer, equipped with a temperature-stabilised deuterated tryglycine sulphate (DGTS) detector, a single-bounce horizontal attenuation total reflection (HATR) accessory, spectral range 4000-650 cm^{-1}, resolution 2 cm^{-1}, the cumulative number of scan 64 times.

2.2 Materials

Stephania tetrandra S. Moore and Stephania cepharantha Hayata are derived from Jinhua, zhejiang province, China, in July, 2007. All samples were deposited at the Department of Botany of Zhejiang Normal University in China. All samples have been grounded to fine powder in agate mortars to 200 mesh respectively.

2.3 Spectral Measurements

All spectra were recorded as 64 scans with 2 cm^{-1} resolution. The FTIR spectrum background was recorded before collecting the sample's FTIR spectrum. Reference spectra were recorded using a blank HATR germanium wafer. Single beam spectra were obtained for all the samples and ratioed against the background spectra of air to present the spectra in absorbance units. The powder sample was put on germanium wafer, and then impacted using pressure tower. FTIR were collected according to the instrument test requirement. After each experiment the HATR germanium wafer was thoroughly washed with distilled water and dried with nitrogen, and its spectra were examined to ensure that no residue from the previous experiment was retained on the germanium wafer surface. The powder samples cover the whole area of the HATR element that contributes to the spectral measurement. All spectra were automatic baseline corrected. All experiments were repeated three times and the averaged spectra used for further analysis.

2.4 Data analysis

FTIR of all the samples can be obtained by determination. According to the absorbance value characteristic of absorption peak, we can make the principal component analysis to the data, which are obtained by data copy in different wave bands. Then Matlab software is used to make wavelet transform to further analyze the data. Using Morlet wavelet, which has a good detection capability of the signal singularity, as the analysis wavelet, one-dimensional CWT is done to the FTIR spectra of samples under different scales. Then the difference of FTIR spectra of the samples in various scales is compared. We choose three representative scales to extract features of Stephania tetrandra S. Moore and Stephania cepharantha Hayata, then use BP neural network to identify them. In the experiment we make one-dimensional CWT to the FTIR spectra of the samples (They are decomposed into 17 levels). We choose three scales (13,9 and 5) as the scales to extract the feature vector.

3 Results and Analysis

3.1 Principal Component Analysis

Principal component analysis (PCA) is made to the FTIR spectra of the samples. PCA load matrix CL can be obtained by principal component scores matrix C, eigenvalue λ and variance matrix S. Principal component load reflects the correlation between the principal component and the original FTIR variables. PCA case scores are used to draw 2-Dimensional graph, the result is shown in Figure 1. Figure 1 shows the impact of various indexes, that is the main component

load. Horizontal coordinate shows the wave number values of various indexes, and longitudinal coordinate is the impact factor values (PCA case scores).

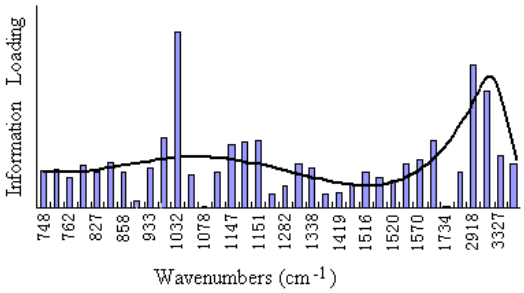


Figure 1 PCA case scores by FTIR of *Stephania tetrandra* S. Moore and *Stephania cepharantha* Hayata samples

Figure 1 shows that the biggest factor affecting the sample is the stretching vibration absorption band about 1050 cm^{-1} in the C-O bond of cellulose molecules. It also can be seen from Principal component analysis map the information load distribution, where larger information load region is in 1800 ～ 750cm^{-1}. Characteristic in the area is not obvious because this region includes fingerprint area, which contains abundant molecular structure information, and in the high-wave-number absorption region is mainly hydroxyl and amino stretching vibration absorption. Therefore, this paper we extract wavelet features in the region of 1800～750cm^{-1} and analyze them using artificial neural network.

3.2 Feature Extraction of FTIR in CWT Domain

Proper wavelet basis function and decomposing level number should be determined when using wavelet transformation to analyze data. The suitable Wavelet Base and wavelet scale are determined by the effect of signal decomposition in different scales and the characteristics of the FTIR signal in wavelet multi-scale decomposition procedure. Its selecting standard is to extrude some characteristic peaks in the original spectra and select wavelet base, which has good smooth property. Having compared the properties of Haar, Daubechies, Mexicon hat, Meyer, Morlet and Symlets wavelet, we select anisotropic Morlet wavelet as " analysis wavelet " for its relatively concentrated frequency energy, and little frequency superposition. It has good symmetry in time domain and linear phase so that it can be sure the wavelet transform without distortion. In this paper the continuous wavelet transform is done to the FTIR spectra of Stephania tetrandra S. Moore and Stephania cepharantha Hayata's respectively, and the decomposing level number is set as 17. We choose representative three levels (5, 9 and 13) to extract their characteristics because the remaining levels have the similar characteristics.

It is very difficult to extract representative features in the small scale in the CWT domain because the detail signal is too sensitive to tinny changes of the spectrum characteristic peaks to result in some false features. For wavelet coefficients in the CWT domain in adjacent scales change little, we select three scales' detail signals (13, 9 and 5) as characteristic variable to construct the features space. Characteristic variable is defined as the energy of spectrum at scale 13, 9 and 5 in the continuous wavelet domain. In order to effectively extract representative characteristics within three scales of continuous wavelet, the spectra in each scale is divided into three representative regions respectively. Figure 2 is the division diagram of the feature regions:

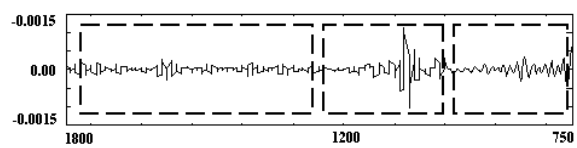


Figure 2 Division of feature region in the CWT domain

From Figure6 we can see that the region of 1800～750cm^{-1} can be divided into three feature regions: 1800 ～ 1300cm^{-1}, 1200 ～ 1100cm^{-1} and 900 ～ 750cm^{-1}. Nine feature regions of three scales in the CWT domain, whose feature values are the spectra energy in the nine feature regions, form the feature vector.

3.3 Classification results

In order to verify the validity of proposed method,

we test our method using the FTIR spectra of 128 pairs of Stephania tetrandra S. Moore and Stephania cepharantha Hayata's. Where 78 pairs of samples are used to train BP neural network, and the remaining 50 pairs of samples are used to test the performance of neural network. Table 1 shows the training and testing results by BP neural network.

Table 1 Training and testing results by BP neural network

Sample type	Identification rate of Stephania tetrandra S. Moore	Identification rate of Stephania cepharantha Hayata's
Training samples (78pairs)	100%	100%
Testing samples(50pairs)	99.6%	99.8%

From Table 1 we can see that the identification rate with BP neural network to identify the Stephania tetrandra S. Moore and Stephania cepharantha Hayata's is 100%, while testing samples of the identification rate is 99.6% and 99.8% respectively. So the Stephania tetrandra S. Moore and Stephania cepharantha Hayata's can be correctly identified by combining BP neural network with continuous wavelet features.

4 Conclusion

Direct determination of plant samples by FTIR is convenient and fast. The proposed method has a high recognition rate to the Stephania tetrandra S. Moore and Stephania cepharantha Hayata's by combining BP neural network with the continuous wavelet features of FTIR of samples. Compared with the traditional PCA method, the proposed method has following advantages:

1) Using PCA to extract Fourier transform infrared spectra features, it requires the number of samples must be greater than the signal dimension and thus sample again has to be done to reduce the dimension number. There is no restriction on the number of samples by the proposed method.

2) PCA method achieves the features according to statistical characteristics of all the samples, which requires the features of new samples to be similar to the ones of original samples. The proposed method directly extracts the features in the CWT domain. Therefore the proposed method has strong adaptive ability to the new samples.

3) To the same testing data, PCA method and proposed method have similar convergence when they are used to extract the features of signals, however, the proposed method has higher corrective identification rate than PCA method.

References

[1] Ehrentreich F., "Wavelet transform applications in analytical chemistry", Analytical and Bioanalytical Chemistry, Vol. 372, No. 1, 2002, pp. 115~121

[2] Liu Y., Cameron I.T., Bhatia S.K., "A wavelet-based adaptive technique for adsorption problems involving steep gradients", Computers & Chemical Engineering, Vol. 25, No.11, 2001, pp. 1611~1619

[3] Shao Li-min, Lin Xiang-qin, Shao Xue-guang, "A wavelet transform and its application to spectroscopic analysis", Applied Spectroscopy Reviews, Vol. 37, No.4, 2002, pp.429~450

[4] Iyengar S.S., Frisch M.J., "Effect of time-dependent basis functions and their superposition error on atom-centered density matrix propagation (ADMP): connections to wavelet theory of multiresolution analysis", Journal of Chemical Physics, Vol.121, No.11, 2004, pp.5061~5070

[5] Ying Yi-bin, Liu Yan-de, Fu Xia-ping, Lu Hui-shan, "Effect of wavelet transform technique s upon the estimation of sugar content in apple with near-infrared spectroscopy", Proceedings of the SPIE-The International Society for Optical Engineering, Vol.5587, No.1, 2004, pp.29~41

[6] Chen Jun, Wang Xue Z., "A wavelet method for analysis of droplet and particle images for monitoring heterogeneous processes", Chemical Engineering Communications, Vol.192, No.4, 2005, pp.499~515

[7] Nie Lei, Wu Shou-guo, Lin Xiang-qin, Zheng Long-zhen, Rui Lei, "Approximate derivative calculated by using continuous wavelet transform", Journal of Chemical Information and Computer Sciences, Vol.42, No.2, 2002, pp.274~283

[8] Khayamian T., Tabaraki R., Ensafi A.A., "Wavelet neural network modeling in QSPR for prediction of solubility of 25

anthraquinone dyes at different temperatures and pressures in supercritical carbon dioxide", Journal of Molecular Graphics & Modelling, Vol.25, No.1, 2006, pp.46~54

[9] Jie Cui, Loewy J., Kendal E.J., "Automated search for arthritic patterns in infrared spectra of synovial fluid using adaptive wavelets and fuzzy C-Means analysis", IEEE Transactions on Biomedical Engineering, Vol.53, No.5, 2006, pp.800~809

[10] Piotrowski P.L., Sumpter B.G., Malling H.V., Wassom J.S., Lu P.Y., Brothers R.A., Sega G.A., Martin S.A., Parang M., "A toxicity evaluation and predictive system based on neural networks and wavelets", Journal of Chemical Information and Modeling, Vol.47, No.2, 2007, pp.676~685

[11] Yang Fu-sheng, Engineering analysis and application of wavelet transform, Beijing: Science Press, 2003

[12] Zhang Chang-jiang, Cheng Cun-gui, "Identification of semen celosiae and cockscomb flower using continuous wavelet transform with FTIR", Rare Metal Materials and Engineering, Vol.35, 2006, pp.614~616

[13] Zhang Chang-jiang, Li Dan-ting, Liang Jiu-zhen, Cheng Cun-gui, "Identification of semen celosiae and. semen celosiae cristatae using continuous wavelet transform with FTIR", Spectroscopy and Spectral Analysis, Vol.27, 2007, pp.50~53

The Research of the Reflection Mechanism to Framework of Persistence Data Layer*

Yuansheng Lou Zhijian Wang Longda Huang Lulu Yue Hongtao Xu

College of Computer & Information Engineering, Hohai University, Nanjing, Jiangsu , 210098,China
Email:Wise.lou@163.com, Zhjwang@hhu.edu.cn

Abstract

Java reflection mechanism is an important technology for improving system flexibility and expandability, which makes the software has the ability of self-adaptive, it changes itself with the external environmental change so as to realize the dynamic evolution of the program. The concept and mechanism of reflection are introduced in this paper, and present a Framework for Persistence Data Layer Based on O/R Mapping. This framework implement dynamic loading and storing the information of Domain Object by adopted the basis of reflection mechanism, and the concrete realizing method example is also presented.

Keywords: Persistence Data layer Framework; Reflection Mechanism; Domain Object; OR Mapping

1 Introduction

In the development of J2EE technology, database access are generally implemented by coding SQL in business logic layer of the three-layer architecture .Business logic and data access logic doping together to form a certain degree of interdependence, resulting system maintainability and scalability poor[1]. Now, the usually solution is to add an additional persistence data layer to access the database between the business logic layer and data layer. It defines the data access interface, business logic layer access database by accessing this interface. This condition not only shielding off the bottom of the complex details, but also enhancing the development efficiency.

The solutions of J2EE persistence layer always include CMP, JDO and OR Mapping recently[2]. Compared with CMP, OR Mapping has obviously advantage in intellectual investment protection, performance and dynamic query, as well as comparison with the JDO[3], in the aspect of packaging and product standards improving obviously, and specially designed to accord with the practical needs of the development. It has become a mainstream design method of persistence layer[4]. But the most important reasons is that OR mapping solve the problem of "incompatible" phenomenon consist in the object model and relations examples, OR mapping based persistence data layer components make up the difference between these two paradigms to achieve the object persistence, and the main application is currently developed by the method of object-oriented program thinking. Domain object information dynamically loaded is the key issues for the OR mapping based persistence data layer framework, because after the complement of loading the domain object information, the specific business logic operation can be designed and program by Object-Oriented method.

Reflection concept is first proposed by BC Smith, which means the ability of system expression and the ability to change self behavior[5][6]. Reflection also support the introspection and adjustments of the system by a principled method, and the ability to configuration and to re-configure the system under the different circumstance[7]. Java Reflection mechanism implemented by meta layer architecture which contents Meta Level and Base Level[8]. Meta layer compose by the

* Supported by National Natural Fund (107056) and the Key(Key grant) Project of Chinese Ministry of Education(107056).

meta-object, used to describe the class itself rather than their usage. Meta layer Object stipulated by MOP (Meta Object Protocol), the Base Level is used to define application logic, and implemented by meta-object. Java does not belong to dynamic languages, but Java's reflection mechanism to achieve a certain extent the dynamic function. Java1.1 providing java reflection mechanisms by the expansion of the Class[9]. Java's reflection mechanism that allows running program dynamically load a class to generate the object's instance and the method of calling this instance. This feature allows the developer can used reflection of mechanism to bring about the dynamic loading of the changes part in the time of applications design. This decrease of the coupling of the system, and improving software reuse[10].

A lightweight persistence data layer framework based on OR mapping are presented in this paper, and put forward to resolve domain object information dynamically loaded by using java's reflection mechanism.

2 Framework of Persistence Layer Based on or Mapping

2.1 Structure of the Framework

In the view of MVC design pattern, persistence data layer corresponding to the part of Model, it is the most important part of information system design, it is the foundation of the performance of the system and it's migration ability. This paper presents a framework for Persistence layer includes five modules (Figure 1), and they implement the corresponding function by the management of Domain Object Factory, the describing of various modules as follows:

(1) Database services module： Realize the database connection and the operation of bottom layer interface.

(2) Primary key cache module：Realize retrieval and cache the primary key of the table from DataBaseMetaData. If you have already cached in the PKCache mapping table in cache, you can directly fetch primary key column according to table name, otherwise you need to fetch from the database, and update primary key mapping table.

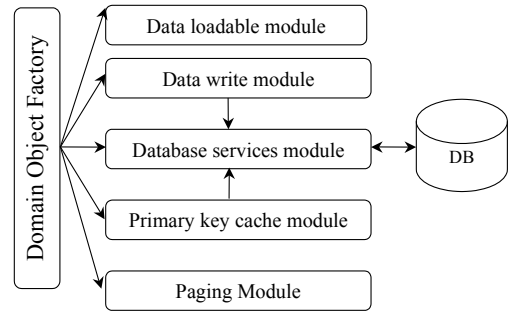


Figure 1 Structure of persistence layer framework

(3) Paging Module：Random get the object of the corresponding page according to the total number of objects and the storage capacity of object per page.

(4) Data loadable module： Construct MetaData of the Object factory class through the mapping file, as well as construct a domain Object list(objectList) through the result set parameters.

(5) Data write module：After get the database table name, it begin to execute insert, update and delete operation of corresponding records accord to the creat, modify and delete method of the object in domain object list and the list to be deleted.

Among the five modules, data loadable modules is the core of Persistence Layer Framework based on OR mapping, its also the main module to implement java Reflection mechanism.

2.2 Framework Working mechanisms of the Framework

In this paper, the presented framework include two core foundation Classes: DomainObjectFactory and AbstractDomainObject, it needs to extend this two core foundation Classes for each of the domain object which waiting for persistent. The specific domain object which extend the class of AbstractDomainObject take the role of describing a record in the result set, and that the DomainObjectFactory of specific domain object responsible for the completion of the management of the domain object by calling other class. (figure 2)

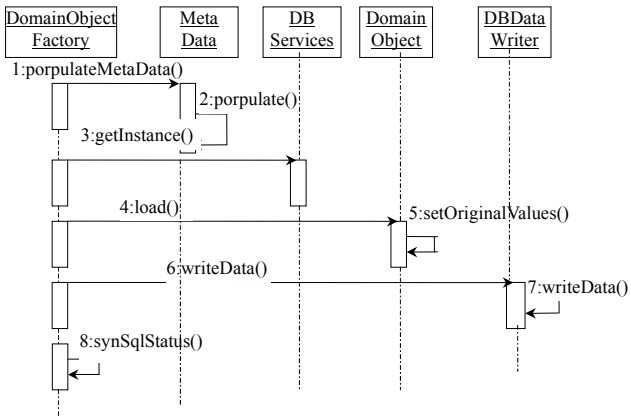


Figure 2 Working Sequence diagram of the Persistence Layer Framework

The working mechanism of the framework can be summarized as the follows：

(1) DomainObjectFactory class by calling the method populate()of the MetaData class, using a column of the database table to the key of the domain object field loading the domain object metadata information to mapping file, including the fields information of object and accordingly getter and setter information.

(2) Establish connectivity with database by creating the instance of DBServices. Initialize the ObjectList of DomainObjectFactory according to the provided ResultSet, each record corresponds to a domain object, and now the original information of each domain object stored in OriginalValues field of the domain object by the form of fields name and the array of fields value.Each DomainObject class has a flag, recorded the action to domain object in the business logic operations, it will be process corresponding database operations according to the corresponding flag when DomainObjectFactory class write data to the database, and also this updated operation only updates the column which the field values of the current domain object different with the original value the originalValues field recorded.

(3) After each successful implementation of a transaction, by calling the method synSqlStatus(), DomainObjectFactory class will synchronous updates domain entities object in objectList,in order to make the conformability between the information of domain object OriginalValues field and the record of current table in the database.

3 The Application of Java Reflection Mechanism in the Framework

3.1 Dynamic loading mechanism of the Persistence layer Framework

Domain object information come from the field which take out the value of the column from a result set(ResultSet) in the database, and put to corresponding field of the domain object, and the different fields of different domain object as a result of the specific write method (write method) is different, so it needs to take into account the write method of different object's specific fields dynamic invocation when design a result set loading in the domain object factory class. Therefore, the framework must be flexible enough to adapt to the whole domain object information dynamically loaded. By using reflective mechanisms and java beans technology, this paper implement the specific domain objects dynamically loaded. Domain object dynamic loading process is shown as Figure 3.

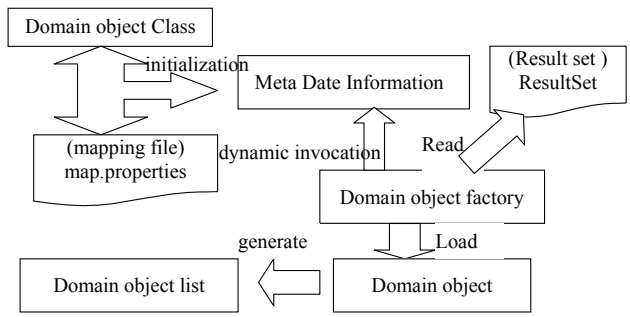


Figure 3 Domain object dynamic loading

(1) The framework load domain object metadata information via a mapping file and the Class of the object, and Stored in metaData field of domain object factory class. Metadata information includes the number of the fields and in storage in the form of an array of field names, field of read/write method, the fields and the fields of class name, and other information. The contents of Mapping file is the mutual of the list for the database to the key of the domain object field, as a resource document format (.Propertty). Because

resources documents are very simple, of course, you can also choose xml documents, JNDI or databases. The concretely format of resource file shows as follow:

TABLENAME.COLUMNNAME=FieldName

Each domain object only needs to define resource file of one mapping, the column name must use the full name, because of a domain object may be correspond to more tables.

(2)Traversing all records according to the given results set ResultSet, and traversing all column to each record. According to the writeMethod array and subscript of metaData dynamic invoke the write method of domain object, Initialize the corresponding fields of the domain object by using the list value of the record, and switch the result set including all records to a list of domain object stored in the domain object factory objectList field.

3.2 Java reflection mechanism implemented in the process of framework dynamic loading

In order to explain the domain object dynamic loading process, Figure4 expressed the relationship between the main class which the loading process concerned, achieved by the following steps:

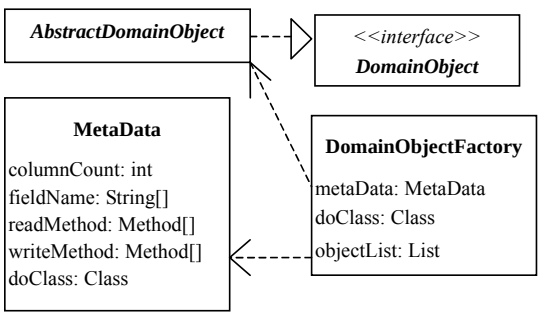


Figure 4 Static structure chat of dynamic load related class

(1) According to the Class of provided domain object and mapping file, domain object factory initialize metadata information of domain object. The initialization of the array of reading and writing method of the MetaData object is very important, because the domain object factory to achieve the domain object information loaded is by calling the read/write function of MetaData object,so it's correct or not directly related to the success of initialization of the domain object information. A function of setting write method according to an array subscript and the field names in MetaData class presented as follow:

protected void populateWriteMethod(int columnIndex, String fieldName){

// write function of to initialize the corresponding fields according to an array subscript and the field names

try{

Method mWrite = (new PropertyDescriptor (fieldName, this.doClass)).getWriteMethod();

//Acquire the field's write method by the use of it's attribute descriptors, which this.doClass express Class object of domain object

this.setWriteMethod(columnIndex, mWrite);

// Assign to the columnIndex section element of Write function array

}catch (IntrospectionException e){}

//ignore Introspection Exception abnormity here;

}

In the basis of this method, you can put data to all write methods array element by traversing all the fields, and complete the initialization of object read/write method array.

(2) Domain object factory traversing each record according to the provided result set(ResultSet), and doing the following operations to each record:

a) DomainObject obj =(DomainObject) doClass. newInstance();

// Generate a domain object, DomainObject

b) //Traversing all column of the record according to column's subscript columnIndex, then initialize each field of the generated DomainObject object;

this.metaData.getWriteMethod(columnIndex).invoke(obj, fieldValue);

// Dynamic invocation write function of domain object corresponding fields, put values to the corresponding domain object field by each column values(fieldValue) of the acquired records.

c) Add domain object obj completed initialization to objectList field of domain object factory.

The initialization of the domain object

implemented based on a result set(ResultSet) achieved now . Through java reflection mechanism to bring about a domain object information dynamically loaded not only increasing the flexibility of the program, and simplifying the code. The main attention of this paper is the java reflection mechanism applied to persistence data layer framework based on OR mapping, so the contents of persistence layer framework about transaction control, concurrency control of Optimistic locking will not given out.

4 Conclusions

Java Reflection mechanism provides a kind multifunctional method of dynamic link program components, it implement domain object metadata and original data dynamic loading through reflection mechanism, this give out a better solution which the persistence layer framework based OR mapping process the real-time dynamic operation. The main disadvantage of reflection mechanism is performance efficiency problems, the methods and properties reflection accessed are more slowly than the directly code. The application of the reflection mechanism in the persistence layer framework, you can decide by your actual situation. In correlative research works, we used this framework as persistence data layer for a embedded lightweight workflow engine, it achieves a better effect for the flexibility, scalability, and the embedded preferable requirements of the engine.

References

[1] YI Yan, ZHOU Cheng, DAI Zhu-ying. Realizing data persistence framework with designpatterns. Computer Engineering and Design. 2005, 26(12): 3365-3367

[2] Xu Changsheng, Dai Chao , Xie li, Research of Data Persistency in J2EE[J], Computer Applications and Software, 2006, 23(4):56-57,75

[3] LIU Ke1,YANG Guangzhong, TANG Ju. Optimizing Persistence Objects Query of Jdo Based on Objects Access Layer. MICROCOMPUIER APPLICATIONS. 2006, 27(3):368-371

[4] TIAN Ke,XIE Shibo,FANG Ma, Solution of J2EE's Data Persistence[J]. Computer Engineering, 2003, 29(22):93-95

[5] B.C. Smith. Massachusetts Institute of Technology [D]. Phd thesis, 1982

[6] HU Hai-yang, MA Xiao-Xing, TAO Xian-Ping, LU Jian. Research and Advance of reflection Middleware. CHINESE JOURNAL OF COMPUTERS. 2005,28(9): 1407-1420

[7] YANG Fu-qing,MET Hong,LU Jian,JIN Zhi. Some Discussion on the Development of Software Technology[J]. Acta Electronica Sinica, 2002, 30(12A):1901-1906

[8] F Kon, F Costa, G Blair, RH Campbell. The case for reflective middleware [J]. Communications of the ACM, 2002,45(6):33-38

[9] Glen McCluskey. Using Java Reflection [EB/OL]. http:// java.sun.com/developer/technicalArticles/ALT/Reflection/

[10] Marry Leisner. Confessions o f a Used Program Salesman: Institutionalizing Software Reuse [J]. ACM SIGSOFT Software Engineering Notes. 1995, 20(5)

Research on the Page Replacement Model in Search Engine Collector

Meiren Zhang[1] Yongfeng Li[1] Yongfeng Li[2]

1 School of Mathematics and Information engineering, Taizhou University, Linhai , Zhejiang 317000, China
Email: tzsimple@163.com

2 School of Computer Science and Techenology, WuHan University of Technology ,Wuhan , Hubei 430070,China
Email: lyf20061031@163.com

Abstract

The method of repeat URL filtering for the existing search engine collector is analyzed, and some shortcomings are pointed out. On the base of virtual memory page replacement algorithm in operation system, a page replacement model is introduced the search engine collector. And then the disk page structure and memory page structure are designed respectively. Finally, a fingerprint search algorithm is given. Through the practical application of the models and technologies in our projects, we find they can solve the speed problem for filtering tens of thousands of URL in the small-capacity memory.

Keywords: search engine; collector; page-replacement; filter; finger mark; URL

1 Introduction

The main function of search engine collector is to collect the resources in the internet to local for further pretreatment. Because they have existed in a large number of redundant links, therefore repeat URL filtering becomes the main question need to be solved in the number of high-class collector. The simplest method for solving the repeat URL filtering is URL string matching, but its efficiency is relatively low in the face of tens of thousands of internet resources. Therefore now to solve this problem, many collectors uses HASH table. That is to store URL strings into HASH table through some HASH algorithm. In essence, the way is to get higher searching rate by losing part of memory. Therefore when the number of the resource searched in the internet become more and more large, Hash table would also become very large. The result is that there is no ordinary computer memory to accommodate such a large HASH table. So in our research, we plan to store Hash table into disk file and create a cache in memory in order to store the last visited URL. In the actual application, however, we find the method can bring read and write disk files irregularly and frequently. It will reduce efficiency and will damage disk. To solve the problem, the paper presents a page replacement solution. And then disk page structure and memory page structure are designed.

2 Page Replacement Model

Since some page links of some website lie in the website, the paper divides URL into two parts for fingerprint calculation. That is host domain and path address. For example, according to the method, this website, which is http://www.163.com/news/Content.asp?id=27, will be divided into " http://www.163.com" and "/news / content.asp? id = 27" , and then be calculated into a 64 bits fingerprint data respectively. So they form a 128 bits fingerprint group together. Because 64 bits can show 64-power of 2 kinds of combinations, it is enough to express the path number in and out host. Therefore

Taizhou Science and Technology plan project(NO: 07322)

the impact on the collision rate can be ignored.

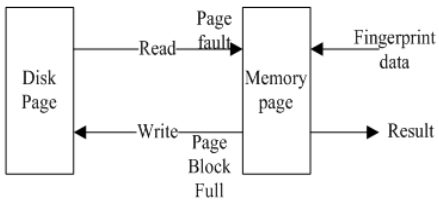


Figure 1 Page Replacement Model

The regions, which are formed by the path segment and data segment of the same host segment, are designed into a page. The page create index through the host segment so that the page data can be accessed frequently for a period of time. As shown in Figure 1, The URL obtained from the website is converted into fingerprint data firstly. Then to it is searched in memory pages. If the result is true, then the fingerprint data will be read a page into memory from the disk, and conversely, will create a new page in memory and write fingerprint data into it. If the stored pages in memory become excessive, the pages of little visitation will be written back to disk.

3 Designing of pages structure

3.1 Page disk structure

As shown in figure 2, disk pages are designed into three parts. They are page table, pages as well as page Block.

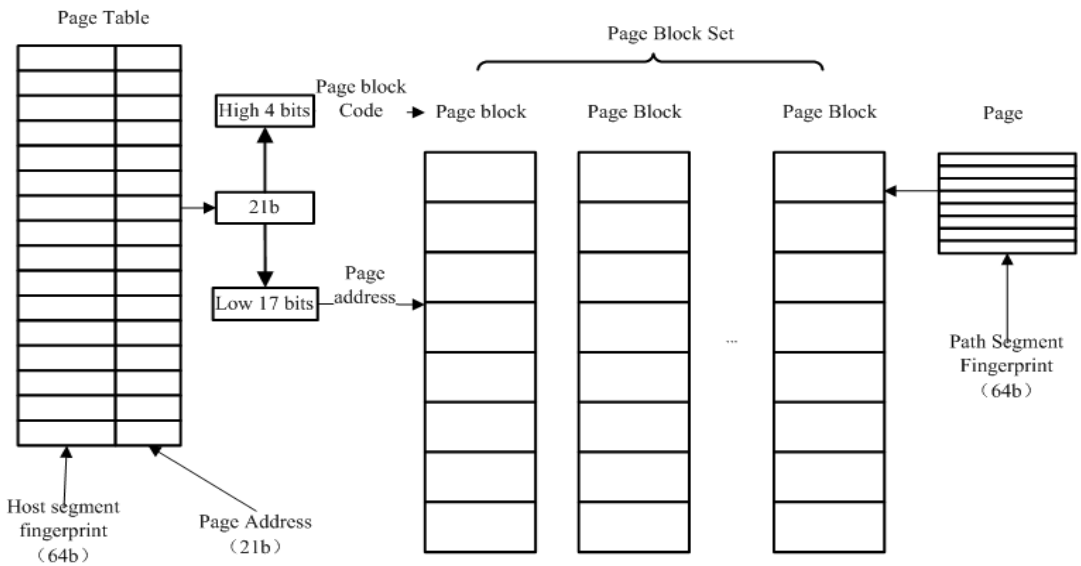


Figure 2 Disk page Structure

As a collection of path segment fingerprints, page is stored in the form of linear table and ordered by sizes of fingerprints. Since the average number of a website page are about 500 to 1000, therefore length of a page is designed into 1 kilobytes, namely 1024 path segment fingerprints. So size of a page is 8 kilobytes. That is, the size of a page = the length of path segment fingerprint data * the length of page =8*1 kilobytes = 8 kilobytes. If the webpage number is beyond 1 kilobyte, the webpage will be shown by multi-page.

Page block is a collection of pages. Each page has an index address. The length of page block may decide the number of pages and the maximal number of hosts. The formula of size of page block as follows: the size of a page block = the length of a page block * the capacity of a page.

At present, the total number of domestic domain names is about 10 millions, so the length of a page block is set 10 megabytes, namely 104857600. Therefore, index address of a page need 21 bits, size of a page block is 80 gigabytes. Such a big document that is not easy to be operated but easy to be destroyed. So a page

block is divided as sixteen pages. The size of each page is set 5 gigabytes, of which the high 4-bit address of page index address explain the page block index, and the low 17-bit address shows page index of the page block.

Page table is a linear table, consists of two domains. One is host segment fingerprint domain, the other is page address domain. Host segment fingerprint domain, 64 bits, stores fingerprint by the way of HASH table. Page address domain stores page index address of corresponding host. Since the maximal host numbers are 10 megabytes, therefore the length of a page table is also set 10 megabytes. The low 21-bit address of host segment fingerprint shows its storage index.

3.2 Memory page structure

In memory, a memory page block is designed for storing the small number of pages in disk page block set. In the meaning time, memory page structure is also redesigned. As shown in figure 3.

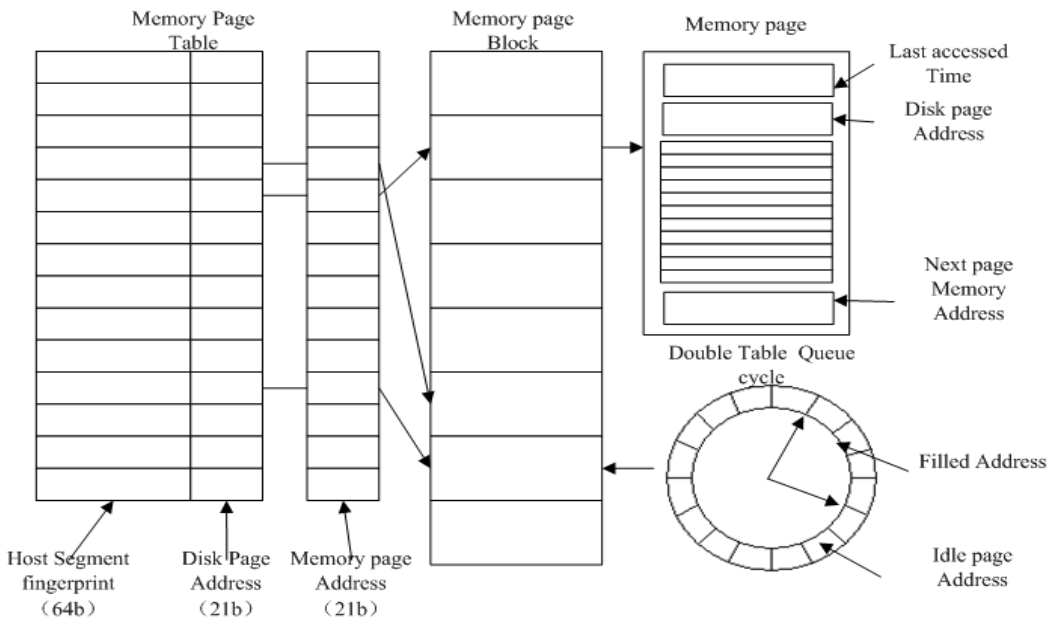


Figure 3 Structure of Memory Page Table, Memory Page Block and Memory Page

To facilitate memory page to be written back to disk, a disk page address domain is designed in memory page. And it is shown by 21-bit address of memory page in disk. In order to the same domain name have much more page, it is necessary to be accessed continuously between pages. In the meaning time, a next memory address domain is designed to store its next page address in memory page block for adding and deleting pages easily in memory page structure. A final timestamps visiting is designed for recording the final time of inserting data in the page in order to clear memory waste pages. A memory page table is designed to store the mirror of disk page table for accelerating page access. In order to search the requirement pages in memory page block easily, a single-byte space (namely memory page address) , whose length is the same as the number of host unit, is distributed to store the memory address of the page to which the host segment points in memory table. If the first page to which the host segment points does not exist, then the value of the single-byte space is 0. If the page has been swapped into memory, then the value is the address of this page in memory page block. A double table pointer queue, whose length is the same as memory page block, is established to explain the full page and the idle page addresses in memory page block. In the beginning, all page block index addresses are stored in ring queue. When storing a page, the page will be placed in the address that is deposited from which the forward table needle point to, then the forward table needle will move

next unit. When releasing a page, the page address is written into the backward table needle unit. And the backward table needle will move previous unit.

4 Fingerprint search algorithm

Through the above designing about disk structure and memory page structure, the URL collected in the internet will be saved the corresponding disk page. Because collector need collect data frequently from the internet by time, the majority of the collected URL has existed in disk, so the repeated URL need be filtered. Namely it is necessary to match the collected URL with the disk existed contents quickly for deciding whether filtering or storing the URL into disk. The searching algorithm is shown in figure 4.

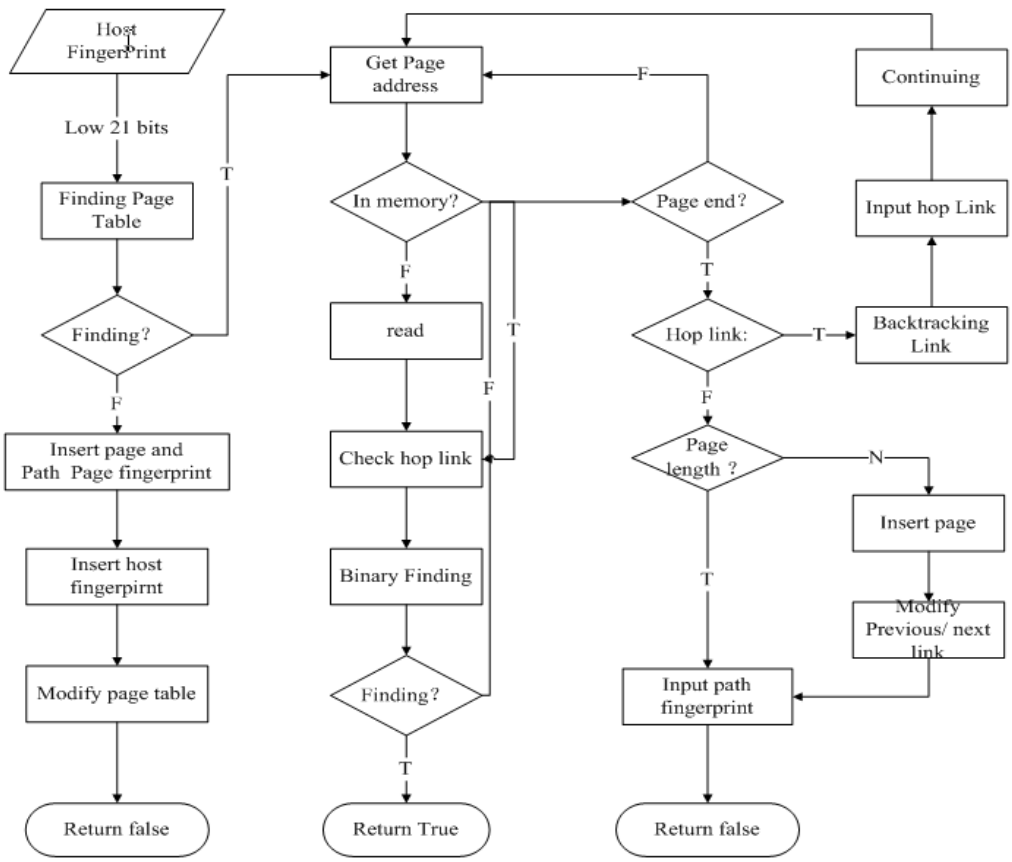


Figure 4 Page Replacement Finding Finger print Algorithm

Firstly, it is necessary to transfer the searched URL into a group of URL fingerprint, and then to get the low 21-bit address of host segment fingerprint and transfer it into index number so as to find the corresponding address in memory page.

If host segment has been not found, then host segment fingerprint is inserted, and page table is emptied temporarily. In the meaning time, a new memory page is created for inserting path fingerprint. The index number of a page, which is pre-inserted in disk page block, will be placed into disk page address domain of the new memory page ， and the disk page index number will be just filled into the corresponding disk page address domain with the host segment fingerprint. Then the new memory page is inserted into the idle region of memory page block and returns memory address. Finally the address is placed the corresponding position of memory page address domain in memory page.

If host segment has been not found, then host

segment finger is inserted, and page table is emptied temporarily. In the meaning time, a new memory page is created for inserting path finger. The index number of a page is pre-inserted in disk page block will be placed into disk page address domain of the new memory page ,and the disk page index number will be just filled into the corresponding disk page address domain with the host segment fingerprint. Then the new memory page is inserted into the idle zone of memory page block and returns memory address. Finally the address is placed the corresponding position of memory page address domain in memory page.

If host segment has been found, then the corresponding memory page address is got out. If the memory page does not exist, then disk page address is got out. The corresponding page, which is set out from disk page block set to fill the corresponding disk page address, will be added the idle region of memory page block. The given memory page address will be added the memory page address domain of host segment in memory page table. If memory page does exist, then path segment fingerprint will be searched in page by binary searching algorithm. If the result is true, then the information that the fingerprint has existed is returned. If the result is false, then it is necessary to judge whether the page is over. If the page is over, then to need find next memory address; If the page does not exist in memory, then to need read disk address of next page, transfer the page from disk, and add it into memory page block. The memory page address will fill memory address value of next page. The binary searching algorithm will be also continuous, so a recursive. Eventually, if the path segment fingerprint can be not found and some page space is available, then the returned result is not to exit the fingerprint.

For helping for finding previous page or memory paper table address, a previous page address domain is added in the memory page structure. But previous page address may point to a page, and may also be a page table. For distinction, the pointed page table address is not by bit. When the domain value is negative, its means is to point to page table, that is the home page of link in the page. The final relation of disk page and memory page is shown in figure 5.

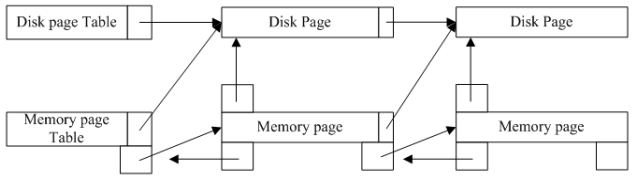


Figure 5 Relations of disk page and memory page

When a page is released, we need judge whether the page is the home page. If it is true, then memory page address of page table will point to next page address the page need to be released, and the previous page address of the page. If it is false, the previous page address of the released page address set as the released page address.

5 Clearing Memory Page And Writing Back Page

When the page number in memory page block reach a higher value. The clearing thread is open. Clearing page uses the least reference used algorithm. When a page is visited each other, the time of the last accessed time is set the current time. More pages will be cleared in a once for reducing the number of clear-up pages.

6 Conclusions

This page utilized the repeat page URL filtering technology to resolve the question that are the high volume webpage and the high speed confliction. What's more important is that it is impossible to do a high level rapid filtration in small-memory computer. The technology can solve the speed problem for filtering billion of URL in the small-capacity memory.

References

[1] Zhong Baorong, Yuan Wenliang, "Research of free page management wary in MMD," Computer Engineering and Design, Vol.28, No.7, 2007, pp.1523-1524

[2] Li Xiaoming, Yan Hongfei, Search Engine Principle, Technology and System, Scientific.Pub. , Bejing, 2005

[3] ZhangChunhong, ”On LRU Algorithm of page-replacement Algorithms,” Journal of Langfang Teachers Colledge, Vol.22,No.4,2006,pp.76-78

[4] Jiang Feihu, Shu Ping, ” Analysis of Page ReferenceSequence's Effecting on LRU Page Replacement,” Computer Technology and Development,Vol.16,NO.5,2006, pp.42-46

[5] Zhang kefei, ”Analyzing embedded real time operating system,” Computer Engineering and Desig*n*, Vol.26, No.8, 2005, pp.20-22

[6] Huang Xianying, Wang Yue, Chen Yuan, ”Memory management strategy in embedded real-time system,” Computer Engineering and Desig*n*, Vol.25, No.10, 2004, pp:1808-1810

[7] Liu Yun, Li Guo, ” Loading of a Real-Time Main Memory Database,” Journal of Softwar*e*, Vol.11, No.6, 2000, pp.829-835

[8] Li Shaohua, Gao Wenyu, ”Survey of Page-ranking Algorithms,” Application Research of Computers, Vol.24, No.6, 2007, pp.4-7

[9] Sheng Buyun, Li Yongfeng, Ding Yufeng, ”Modeling and management of manufacturing resource information in manufacturing grid,” China mechanical engineering, vol.17, No.13,2006, pp.1375~1280

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

ZE: Virtual Environment of Large Scale Worm Tracing*

Wei Shi Qiang Li Jian Kang

College of Computer Science and Technology, JiLin University, Changchun, JiLin 130012, China
Email: loonsw@gmail.com li_qiang@jlu.edu.cn kangjian@jlu.edu.cn

Abstract

Network worms have been a serious security threat on the Internet. Tracing worm propagation path can identify the overall structure of a worm attack's propagation. To detect and defense large scale Internet worms, setting up a convenient and safe experimental environment that capable of running and observing real world worm become an important work, it can be a large scale worm test bed for forensic evidence. We provide a systemic analysis of large-scale worm propagation tracing experiment strategy which is based on virtual machine technology by setting up an experimental environment called zooecium (ZE). First, the framework of ZE is addressed. Then, the design and control of ZE is given. Finally, ZE is analyzed with experiments. Experimental results show that ZE can trigger large-scale worm outbreaks within the controllable scope of human, observe propagation process of the worm, experiment detection and defense techniques, discover worm propagation characteristic such as scanning method and propagation process, real-time collect network traffic and propagation process, investigate network traffic, dynamically throw out the result, launch speculate algorithm for reconstructing out propagation path of the worm. Then actual worm propagation process can be captured and compared with the results using tracing algorithm.

Keywords: Worm; Environment; Tracing

1 Introduction

Worms have been a serious security threat on the Internet. They can spread across the Internet quickly with terrible influence, and Internet worms have been a primary issue faced by malicious code researchers. Currently, research on worm detection and containment continuously improved, tracing the evolution of a worm outbreak (attacking path of worm) is an important research area[1,2,3], it not only reconstruct patient zero (i.e., the initial victim), but also the infection node list in evolution process. The reconstruction result has significance in restraining evolution of worm and forensic evidence.

Large scale network worm tracing research needs a reliable algorithm experimental environment. First, real time tracing algorithm needs to carry out theoretical analysis, and prove the correctness of tracing algorithm under some assumptions and prerequisite conditions. Second, different tracing model with different parameters in the algorithm are established. But theoretical deduce can not reflect the real execution of algorithm. Many researchers use some network simulation platform like ns2 [4] or parallel-ns2 to establish the tracing simulation testing environment, simulate running thousands of nodes in different network topology and bandwidth. But simulation is more applicable to modeling, not real worm spread. Simulation process is too idealistic, not a true reflect of the operating system and demand high performance experimental host. Using physical host for large-scale network worm tracing experiment is also unfeasible. First thousands of physical hosts can not be guaranteed. Second, because of worms destructive, the large number of physical host unable to quickly reuse, management and configuration workload is huge.

* Supported by NSFC (60703023) .Contact author: Qiang Li li_qiang@jlu.edu.cn

In recent years, virtual machine technology's development promoted its application in the field of network security research. Researchers have begun network worm detection and defense experiments using virtual machine technology [5, 6, 7, 8]. One physical host can run a number of virtual machine installed real operating system, and connected to the network. External visitors perceived no internal differences except for a little performance odds. So they can use the virtual machine technology to establish a high realistically, control flexibility, encapsulate and reusable virtual experimental environment. After optimize virtual machine and the installed operating system, the performance requirements of physical host can be reduced. Optimal use of virtual machine technology can simulate thousands of virtual operating system nodes in nearly dozens of physical host, more clearly discover propagation process of network worm in the operating system and network, further observe invaders motivation, tools and methods.

This paper presents a large scale worm propagation experimental environment called zooecium (ZE) for tracing algorithms, which is an isolation environment that can progress related experiments. ZE is based on virtual machine technology, can simulate a large number of hosts and network equipments attend. According to the actual worm, ZE can trigger large-scale worm outbreaks within the controllable scope of human, observe propagation process of the worm, experiment detection and defense techniques, discover worm propagation characteristic such as scanning method and propagation process, real-time collect network traffic and propagation process, investigate network traffic, dynamically throw out the result, launch speculate algorithm for reconstructing out propagation path of the worm. Then actual worm propagation process can be captured and compared with the results using tracing algorithm.

This paper follows as: Part 2 introduces framework of ZE; Part 3 shows main functional design and control of ZE; Part 4 gives a specific example of experiment; Part 5 is concluded.

2 Framework of ZE

To establish environment for large-scale worm tracing has the following main objectives: a), worm experiments can be fully controlled within the scope of human, the start-up and shut down of experimental environment dominated by the experimenter, b), the experimental environment is independent and self networking, communications with the outside world under surveillance, c), experimental process and results can fully be observed, true infection and algorithm results can be compared, d), the experimental environment can be reused, minimized the cost of maintaining.

Figure 1 shows the diagram of the large-scale worm online tracing environment. The whole environment is composed by multiple physical hosts and switch components. Physical hosts connect each other form a LAN through switches.

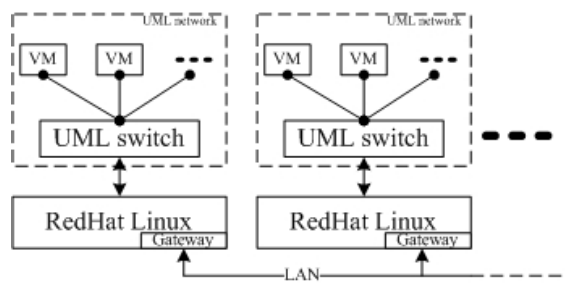


Figure 1 diagram of the large-scale worm online tracking environment

UML[8] is a lightweight virtual machine system on Linux. It can run numerous instances on physical host, with the various versions of Linux operation systems. It can customize operation system of the virtual machine according to the requirement; only need install the necessary system software and system services. Therefore it has a higher performance and occupy fewer resources of the physical host.

Each host installs a UML system in the experimental environment, running advance customized client operating system image, serve as various experimental roles according to the pre-configuration. After environment launched, several virtual machines in a physical host form a virtual local network (VN), and connected via UML virtual switch. Each physical host,

as a gateway of its own local network, connects other VNs on other host. Extending like this, a basic multi-VN experimental environment can be setup.

In ZE, many virtual clients in a physical host have system security holes, and they can be infected by worms. In order to control propagation process of the worm, ZE has some functional modules showing in figure 2. Virtual clients in ZE is using real operating system, and running some services with security holes. Main functions are as follows:

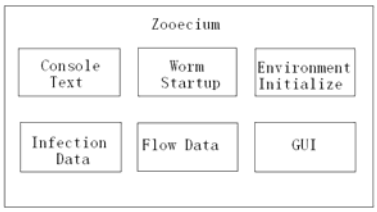


Figure 2 functional modules of ZE

a) virtual environment startup and shutdown

Launch related startup script after physical host LAN is ready, and then start all virtual machines in every physical host. Hosts and virtual clients have to execute initialization script, configure network connections and some other issues.

b) Worm launch

A random virtual client is selected, and executes worm startup script. The script first infected the chosen virtual machine, and then begins to propagate.

c) Infections collect

During the worm propagation, each virtual machine is monitoring its own change of infection characteristic according to the pre-configurations. Once infected, the infection details will be recorded in the host.

d) Network flows collect

After virtual environment startup, every host has the responsibility of monitoring communications between its VN and other VNs of other hosts. Based on predetermined monitoring rules, hosts capture data and transfer to the unified host running tracing algorithm and dynamically throw out the result.

3 Design And Control

According to the propagation characteristics of the worm, ZE accomplish its missions mainly depend on network traffic collection model and various scripts.

3.1 Virtual client system

Virtual client system uses Linux version with security holes. To facilitate the maintenance, all the virtual systems adopt a unified system image file, so the operating system software installed exactly the same. Each virtual client is running different initialization script based on its own id, and completing their respective functions. Using the COW(Copy On Write) technology of UML, all virtual machines only use one unify image file when startup virtual operation system, and create their own differences file to storing data. In this way can avoid each virtual client use a separated virtual image, improve the system efficiency, and reduce the storage space required. When launching the virtual environment, experimenter can specify the number of running virtual machines and their operating roles.

The virtual client startup command is like this:

../linux ubd0=cow$1,../root_fs umid=uml$1 eth0=daemon, unix, /tmp/umlsw con=xterm con1=null con2=null &.

The main image file system create command is like this:

```
stripped=`echo $base|sed 's+\.noarch++g'|sed 's+\.86++g'`
exists="0"
if [ ! -z "$DUPES" ]; then
    exists=`$RPMCMD -qa --root $TMPDIR | grep $stripped | wc -l`
fi
if [ "$exists" -ne "0" ]; then
    echo "skipped: $stripped"
else
    $RPMCMD -Uvh --root $TMPDIR $local
    rm -fr $TMPDIR/var/lib/rpm/__db*
fi
```

3.2 Host network configuration

HOST network configurations separated into two parts. First interconnect all VNs in physical hosts. Then

physical hosts should connect their VN to the LAN. Implementation steps are as follows:

a) Fetch network addresses configuration file, startup virtual switch (uml_switch) to interconnect all virtual clients on a physical host. Then configure address information of virtual switches.

b) Set up packets transmits and ARP resolve of every physical host, ensure that host can communicate with its virtual clients.

c) Use route command to set up routing tables between different VNs, ensure that virtual clients on different hosts can communication each other.

The main commands of configuration script are as follows:

```
uml_switch -hub -tap tap0 -unix /tmp/umlsw -daemon
ifconfig tap0 192.168.$NET.1 netmask 255.255.255.0 up
echo 1 > /proc/sys/net/ipv4/ip_forward
route add -host 192.168.$NET.$i dev tap0
echo 1 > /proc/sys/net/ipv4/conf/tap0/proxy_arp
arp -Ds 192.168.$NET.$i eth0 pub
```

3.3 Virtual client network configuration

a) Launch virtual client; fetch its id according to the corresponding startup command.

The main commands of configuration script are as follows:

```
while [ $i -le 254 ];
do
if [ "$Line" = "ubd0=cow$i,../root_fs umid=uml$i eth0=daemon,,unix,/tmp/umlsw con=xterm con1=null con2=null root=/dev/ubd0" ];
then
        echo $i;
        break;
fi
i=$(($i+1))
done
```

b) Set up its VN address according to the fetched id.

The main commands of configuration script are as follows:

```
while [ $i -le $2 ]; do
    sh small.x86 $i $NET
        rm -rf cow$i
    ID=$i
    segment=$NET
id1=$(($ID/16))
id2=$(($ID%16))
diff=$(($id1-10))
```

c) Set the type of startup service.

3.4 Background flows and data collection

Starting up HTTP service and running lynx command in the active virtual machine, and generating background flows according to the predetermined time cycle. Every host running tcpdump, collect network flows according to some rules and transmit to the designated host.

4 Experiment

Using UML virtual machine technology, we establish an experimental environment include 1000 virtual nodes base on 25 PCs. Virtual clients running Redhat Linux 6.1 operation system with BIND security holes. Physical hosts running Redhat Linux 9.0 operating system. Several virtual clients in a physical host form a VN, virtual clients in different host communicate with each other using gateway in every physical host.

Manually launch a worm propagation break source in one of the twenty LANs, startup Lion worm attack [9], then running tracing algorithm to analyze the final result and true infections. The continuous real time collection network flows include not only worm flows, but also pre-installed normal background flows.

We define the start time of the first attack flow to be the origin of the time axis, is time 0 and each time unit is a second.

The collected flows are shown as the following example:

10.2.19.41 > 10.0.111.244 at 517
10.1.72.94 > 10.4.90.141 at 1380

10.4.253.153 > 10.2.241.139 at 898
10.3.45.35 > 10.0.19.117 at 1191
10.5.79.58 > 10.1.192.70 at 1343
10.4.253.153 >10.3.38.141 at 1228
10.4.90.141 >10.2.19.41 at 1411
10.2.241.139 >10.0.112.191 at 1109

Infection report is shown as the following example:

From 10.1.72.94 to 10.4.90.141 at 1380
From 10.4.253.153 to 10.2.241.139 at 898
From 10.4.253.153 to 10.3.38.141 at 1228
From 10.4.90.141 to 10.2.19.41 at 1411
From 10.2.241.139 to 10.0.112.191 at 1109

Figure 3 and Figure 4 show the worm infected tree dynamically thrown out by the tracking algorithm:

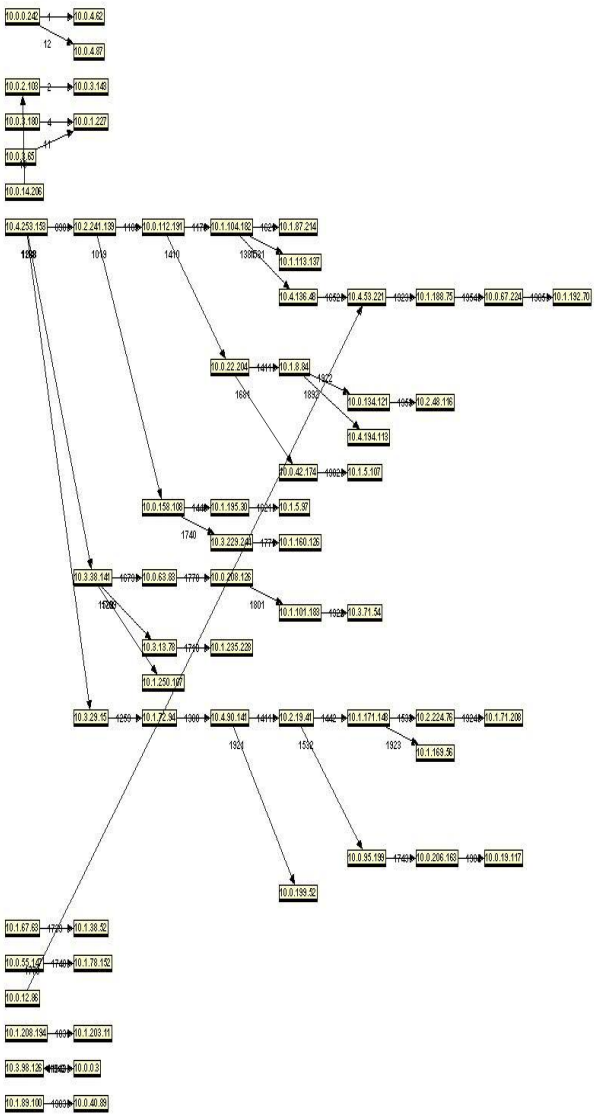


Figure 3 worm infection tree at time 2000

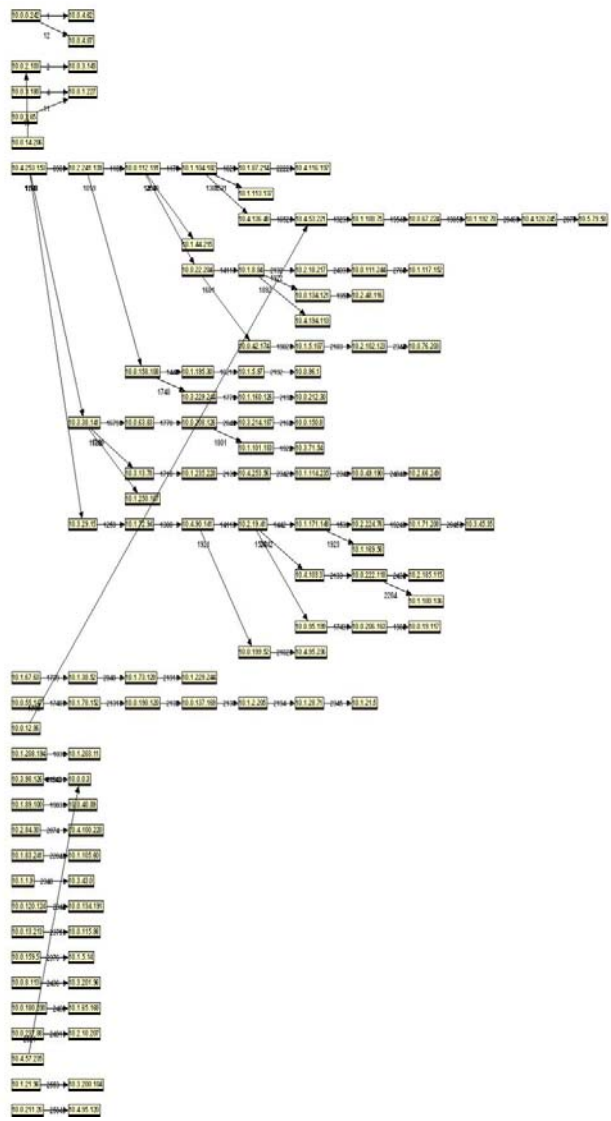


Figure 4 worm infection tree at time 2500

5 Conclusions

ZE, which is based on virtual machine technology, put up high reality and flexibility. ZE is nonproliferation, not destructive, and can effectively use limited resources to simulate thousands of nodes. It is an effective test bed for large scale worm forensic evidence.

Compared with other worm experimental environments[5, 6, 7, 8]. ZE has the following characteristics: a), hosts and virtual clients are controlled by scripts, b), real-time collect the network flows including normal background flows and worm

propagation flows, c), dynamically display tracking results, d), actual worm propagation process is captured and compared with the results using tracing algorithm.

References

[1] D. M. Kienzle and M. C. Elder. Recent worms: a survey and trends. In WORM '03: Proceedings of the 2003 ACM workshop on Rapid Malcode, pages 1–10, New York, NY, USA, 2003. ACM Press

[2] Abu Rajab, M., Monrose, F., and Terzis, A. Worm evolution tracking via timing analysis. In Proceedings of the 2005 ACM Workshop on Rapid Malcode (Fairfax, VA, USA, November 11 - 11, 2005). WORM '05. ACM Press, New York, NY, 52-59

[3] Yinglian Xie, Vyas Sckar, David A.Maltz,Michael K. Reiter, and Hui Zhang. Worm Origin Identification Using Random Moonwalks. In Proceedings of IEEE Symposium on Security and Privacy, pages 242–256, May 2005

[4] The Network Simulator-2, http://www.isi.edu/nsnam/ns/, 2004

[5] X. Jiang, D. Xu, H. J. Wang, and E. H. Spafford, "Virtual Playgrounds for Worm Behavior Investigation", Proceedings of the 8th International Symposium on Recent Advances in Intrusion Detection (RAID 2005), Seattle, WA, September 2005

[6] Michael Vrable, Justin Ma, Jay chen, David Moore, Erik Vandekieft, Alex C. Snoeren, Geoffrey M. Voelker and Stefan Savage, Scalability, Fidelity and Containment in the Potemkin Virtual Honeyfarm, Proceedings of the ACM Symposium on Operating System Principles (SOSP), Brighton, UK, October 2005

[7] Michael Vrable, Justin MaSamuel T. King, Peter M. Chen, Yi-Min Wang, Chad Verbowski, Helen J. Wang, Jacob R. Lorch, "SubVirt: Implementing malware with virtual machines", Proceedings of the 2006 IEEE Symposium on Security and Privacy, May 2006

[8] Symantec Worm Simulator http://enterprisesecurity. symantec.com/content.cfm?articleid=5479

[9] J. Dike. User Mode Linux. http://user-mode-linux.sourceforge. net

[10] Linux Lion Worms. http://www.whitehats.com/library/ worms/lion/, 2001

Adaptive Control System for Ink Key Presetting in Offset Printing Presses

Jinfei Ding[1] Shuangchen Ruan[1]* Ming Fan[2]

Shenzhen Key Laboratory of Laser Engineering, College of Electronic Science and Technology, Shenzhen University, Shenzhen, 518060, China
Email: jfding@szu.edu.cn , scruan@szu.edu.cn *

Technology Quality Assurance Department, Shenzhen Press Group Printing Center Co. LTD, Shennan Ave 6008, Shenzhen, 518009, China
Email: fmszcn@126.com

Abstract

During a make ready in a web offset press it is important to produce as little waste as possible. One way to do faster make ready is to preset the ink keys of the press before it is started. This paper shows how the ink key presetting may be done through an adaptive control system, which predicts ink key presets by using a standard test pattern and the stored printing data as example data, based on the ink coverage distribution of the current print job. Such an adaptive control system for ink key presetting can be used not only for the presses with CIP3/CIP4 control interface but also for the presses without control interface, which feature is shared with a lot of older presses. Use of the developed control systems leads to higher print quality and lower ink and paper waste.

Keywords: Ink key presetting, Adaptive control, Offset printing press, Make ready, Ink coverage distribution

1 Introduction

In today's printing industry with its tough competition and low prices it is important to cut costs wherever it is possible. Along with the introduction of fully digital workflows from customer to printing plate, the development goes toward a higher grade of automation throughout the whole printing process. In the printing-house it is important to reduce the time for make-readies, i.e. the time between one print job stops and the next job starts producing printed sheets that look good should be as short as possible. By reducing the make-ready time the press can produce a larger number of approved sheets per unit of time and the result is a lower cost per printed sheet. When a new print job is started in a web offset press (or in any press which does not use a digital printing method), there is always a certain amount of wasted paper before the printed result looks as it was intended. How much paper that is wasted before the result is acceptable depends in a web offset press mainly on the following three things: register, folding and presetting of the ink keys 0.

Presetting of ink keys is however possible to determine in advance and as soon as a make ready starts 000. In modern presses, which mostly are fixed with sectional duct blade, ink key presetting can be done relatively easy as long as the press are provided with CIP3 interface and corresponding software 000. But the cost for such configuration is usually very high. This is a case especially for the software. Thus in printing company, some presses don't implement CIP3/CIP4 technology yet even if the press has CIP3 interface 0. Furthermore, a majority of presses throughout China are old, which duct blades are unitary and have no CIP3 interface 0. So, it's of great importance that how to implement ink key presetting in old presses without CIP3 data interfaces, or in presses that configured with CIP3 but no corresponding software yet.

In manually operated presses, the pressman will first scan visually the printing plate and estimate the amount of ink needed within each of the sections controlled by the keys of the ink fountain. The disadvantage of this kind of manual presetting is that on the one hand, there is always a great of waste paper and ink, and on the other hand, the adjust time is relatively long 0 . There are other systems wherein an optical scanner is used to scan a printing plate to determine the amount of ink needed within certain narrow sections of the printing plate, and that information is then processed to set automatically the corresponding keys of each fountain. Owing to various reasons, it hasn't been reached to large-scale application 0. Linus Lehnberg put forward an ink key presetting in offset printing using digital images of the plates in his diploma work 0. He found the relationship between the ink coverage distribution and its corresponding ink key opening can be described with a transfer curve. In Brovman's patent, he invented adaptive control system for press presetting, which based on objective data obtained from scanning an image to be printed by means of a light table 0. In Doherty's patent, an inker is preset in accordance with the ink coverage distribution wherein inking operation is simulated and a simulated ink coverage distribution is obtained by driving a steady state error between the printed ink film commands and the ink coverage to zero 0. In this paper, we will show a new adaptive control system for ink key presetting based on the ink coverage distribution of current print job and family of printed jobs as sample together with standard test print job.

2 Basic Printing Theory

2.1 Four-colour printing

The primary colours usually used in four-colour printing are cyan (C), magenta (M), yellow (Y), and black(K), CMYK. A CMY overprint creates black colour. However, black ink (K) is also used in printing. Due to economical reasons, black ink often replaces the CMY overprints. Moreover, black ink is often used to improve the quality of colour pictures.

2.2 Inking system

To get the ink contact with the printing plate, via a rubber blanket, and then be transferred to the paper, printing units are used. Each color has its own unit. As illustrated in Figure 1, a traditional offset printing unit consists of an inking device, a dampening device, a plate cylinder, a rubber blanket cylinder, and a back-pressure cylinder. The width of the cylinders sets the limit of maximum printing width and the circumference maximizes the printing length. Remark that this printing unit prints on one side only. When both sides are printed at once, the inking unit is a little different 0.

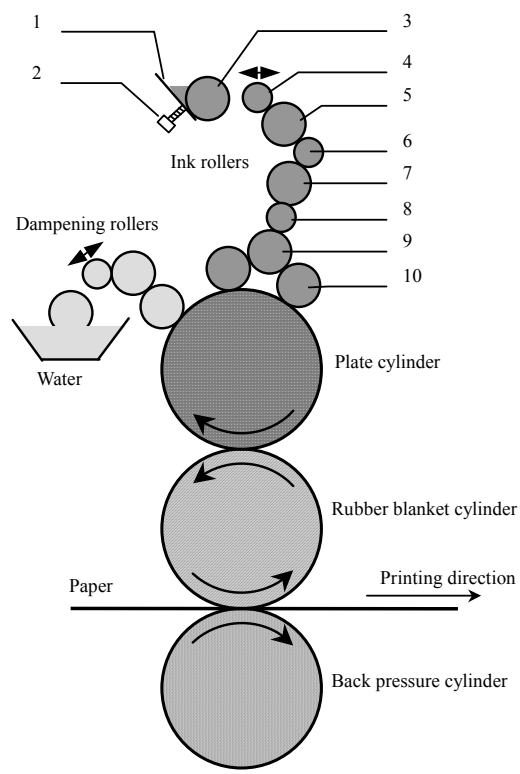


Figure 1 Schematic side view of a typical offset printing unit

1. fountain blade; 2. ink key; 3. ink fountain roller; 4. ink ductor roller; 5,7,9. ink vibrator roller; 6,8. ink distributor roller; 10. ink form roller

In offset printing 0, ink is supplied via an ink train, also referred to as the inker, from an ink fountain to a plate cylinder and then to a blanket cylinder, from which the ink is transferred to the print substrate (e.g. paper),

as shown in Figure 1. The ink train includes an ink fountain roller, also referred to as an ink pickup roller, which picks up the ink at the ink fountain and transfers it to an ink ductor roller. The ductor roller oscillates between the ink fountain roller and an ink vibrator roller and thereby transfers the ink from the ink fountain roller to the vibrator roller. From there, the ink is transferred via distributor rollers to other vibrator rollers, which distribute the ink onto several ink form rollers. The ink form rollers ink the printing plate on the plate cylinder by depositing the ink onto the oleophilic surfaces on the plate. From there, the ink is transferred onto the rubber blanket in accordance with the image to be printed.

The amount of ink that is transferred is most important for the print quality. Too little ink in the ink train leads to faint print and uneven distribution of ink color. Too much ink leads to smearing and blurring of the printed image 0. Different amounts of ink are required in various zones according to the image to be printed. In order to vary the ink feed laterally across the width of the inker, the ink supply leaving the fountain can be adjusted with ink keys, which control the gaps between the blade and the fountain roller, since in a typical ink distribution system wherein ink is placed in a trough formed between fountain roller and the fountain blade. Mechanically, the ink keys control a thin metal blade, the duct blade, which is fitted across the fountain roller. Each ink key controls an equal fractional part of the width across the ink fountain. This fractional part is known as an ink zone. The width of each ink zone is simply the total controllable width divided by the number of ink keys. In modern presses the duct blade is cut up in slices to match the width of ink zones. There also exist duct blades, which are not cut up into slices, in the traditional printing presses. Each ink key thereby defines the ink supply for a respective zone. The number of ink zones and ink keys may vary from six to sixty, or even more 0.

The ink supply is subject to a vast number of variables. To begin with, due to the rheological characteristics of the ink, the relationship between the ink key feed gap and the amount of ink supplied at a given key is not linear. The type of printing ink, dampening agent, and paper, as well as process temperature and plant humidity influence the steady state behavior as well. Further, the various oscillating rollers in the ink train cause a substantial lateral distribution of the ink, so that the amount of ink supplied to a given zone at the rubber blanket is not only dependent on the ink key associated with that zone, but also on adjacent ink keys. In other words, as the ink travels from the ink fountain to the rubber blanket via several laterally oscillating vibrator rollers, a certain amount of ink bleeds from one zone to another. This phenomenon is referred to as the lateral coupling of the inker 0.

3 Ink Key Presetting

Ink key presetting includes two procedures of predicting and setting, i.e., calculating the ink needed in advance and setting the ink keys on presses, in which predicting is the most important. The target of ink key presetting is to get the exact information about the key opening based on digital ink coverage distribution before a print job starts.

Here we developed an adaptive control system in ink key presetting. The entire system flow is illustrated in Figure 2, which is explained in detail in terms of the following three parts.

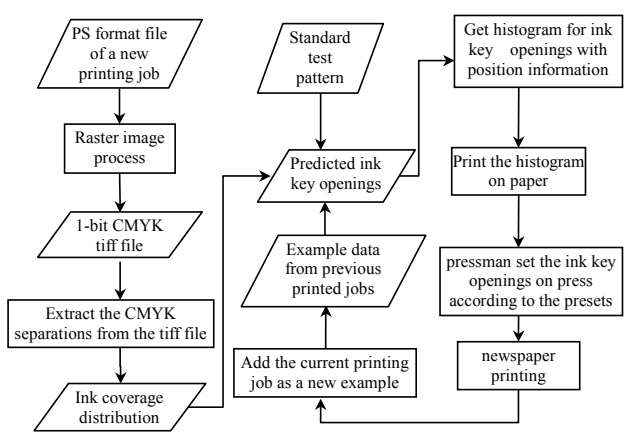


Figure 2 Technical flowchart of the adaptive control system for ink key presetting

3.1 Ink coverage calculation

As mentioned before, in any ink key presetting methods, the ink coverage distribution is the most fundamental and also very important information. The calculation of ink coverage is illustrated in the left part of Figure 2. For a new printing job, firstly a 1-bit CMYK tiff file is obtained from prepress by doing RIP on the PS file. Then calculate the ink coverage distribution through extracting the CMYK separations from the above tiff file.

3.2 Calculate the ink key presets

The offset printing is a typical analog process, which is usually affected by technical characteristics of press and various environmental factors. Hence it is impossible to find an accurate formula for calculating the actual ink quantity mathematically. For example, according to ink coverage distribution, it is a common case that difference between two or more neighbored key opening is very large, which should be adjusted by pressman to get an "OK" effect. Because the duct blade is unitary and continuous, any key opening will affect its adjacent key opening. The large difference between adjacent two or more ink opening will bring on an effect that the highest key opening degree will not obtained the highest coverage distribution and vice versa. Furthermore, Owing to lateral coupling of the inker, it can't be taken for ideality that the released ink quantity from every ink key will directly feed to the corresponding strip of the whole page. That is to say, even if ink quality needed in one ink zone is very small and a very small ink key opening is set, however the actual given ink on the substrate will be amplified as a result of the more ink released from its adjacent ink key.

On the other hand, along with the abrasion and aging of different ink keys, even if they are on the same fountain, their characteristics will be different. To observe the characteristics difference between ink keys on each ink fountain, we design a standard test pattern, in which the ink coverage distributions of CMYK on various ink zones are the same. In order to get the same dot density everywhere in the same ink bar, the actual ink keys' openings are not identical. Figure. 3 shows the difference between the linearly calculated value from ink coverage distribution and the actual ink key opening, in which we can see the different properties of all the ink keys.

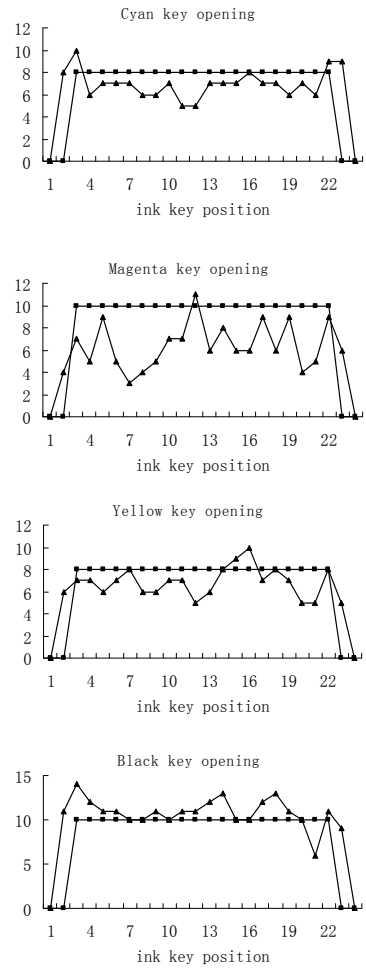


Figure 3 CMYK key opening of standard test pattern (square mark denotes the linearly calculated value from ink coverage distribution, while triangle mark denotes the actual ink key opening.)

In the middle part of Figure 2, it is shown that the ink key presets are decided by the following elements: ink coverage distribution, standard test pattern and Example data from previous printed jobs. Thus, separating a group of jobs which represent an objective relationship between the ink coverage distribution and the key settings for each given ink key is a prerequisite of proper parameter identification. They all are based on large enough examples so that statistical stability can be

reached. Information from a plurality of previously completed jobs, so called exampled data, including data obtained from an objective source and data obtained from subjective source, are analyzed and compared to provide parameters which thereafter are used in ink key presetting in response to subsequently obtained objective data. Specifically, the objective data, such as the amount of coverage, for each of the elements to be controlled, such as keys, is analyzed mathematically.

After a printing job comes into steady status, its ink key opening information as a new example will be added to the example family. From this point of view, the examples always keep fresh and the parameter identification can be adapted to the current status of the press.

3.3 Using the ink key presets in the press

The resulting ink key presets are then used to preset the ink keys of the inker at the state of the actual print job. It is not the same for different presses. Some modern presses that provided with electro-mechanical means for setting the keys from a remote location, and also transducers for indicating each key position at a remote location, for example, on a television screen. In this case, ink key openings can be presets remotely. But for manually operated presses that have no electro-mechanical means for setting the keys, the presetting the ink keys of the inker can be done as the right part of Figure 2 illustrates. A histogram for ink key presets with position information, which is more readable for pressman, is printed on an A4 paper. Pressman set the ink key opening according to the information which the histogram gives.

4 Conclusion

We have developed an adaptive control system in which ink key presets are determined based the ink coverage distribution, employing a standard test pattern and the stored printing data as samples data. By using an adaptive control system for ink key presetting, one can eliminate the inconsistent sampling and subjective color compensation made by the operator and therefore one can expect a more uniform print quality through the production. Consequently, paper and ink waste will be reduced as the print quality variations decrease. Pressman's time is also set free for the benefit of service and maintenance of the printing press equipment.

References

[1] L. Lehnberg, “Ink key presetting in offset printing using digital images of the plates”, Swedish university essays, Linköping University, Sweden, 2002

[2] Y.Y. Guan, “Review of ink presetting technology”, Print today (in Chinese), No.3, pp.64-65, March 2007

[3] M. Fan, H. Li and J. Yang, “Manual ink-presetting on newspaper printing machines”, Printing Field (in Chinese), No.243, pp.66-70, June 2006

[4] C. Chen, “Application of ink control systems in printing”, Printing Field (in Chinese), No. 232, pp. 60-61, July 2005

[5] W. Z. Liu, “Shaping digital workflow for newspaper”, Printing Field (in Chinese), No. 222, pp. 14-15, September 2004

[6] W.Q. Shen and B.Q. Li, “Preparation and setting details of ink preset”, Printing field (in Chinese), No.258, pp.53-56, September 2007

[7] B.Q. Li, “A dilemma between digital proof and ink presetting”, Printing Field (in Chinese), No.251, pp.47-49, February 2007

[8] D.J. Xiao, Offset printing technology (in Chinese), Beijing: Printing industry press, 2001

[9] Y.Z. Brovman, “Adaptive control system for press presetting”, United States Patent 4655135, 1987

[10] N. Doherty, “Ink key presetting system for offset printing machines”, United States Patent 6477954, 2002

Model of Bridge Collaborative Design CAD System

Ming Chen

School of Civil Engineering and Safety, Shanghai Institute of Technology, Shanghai 200235, China
Email: chenmchen@21cn.com

Abstract

The bridge industry has a long tradition of collaborative working between the members of a bridge project team. At the design stage, this has traditionally been based on physical meetings between representatives of the principal design disciplines. These have yielded some success but are hampered by the problems posed by the use of heterogeneous software tools and the lack of effective collaboration tools that are necessary to collapse the time and distance constraints, within which increasingly global design team work. This paper examines some of the issues associated with the use of distributed systems within the bridge industry. Have finished the data representation model of bridge collaborative design at first, then carried on research to system model based on J2EE, the model is used in an instance and its feasibility is validated finally.

Keywords: Bridge Design, Collaborative Design, Data Model, System Design

1 Introduction

The concept of Concurrent Engineering (CE) in the bridge industry gathers the participants of a bridge project as a team within which collaborative design are made. A typical construction project usually involves up to ten or more different professional disciplines, in many cases they are also geographically dispersed[1]. In such a situation, physical meetings are inconvenient, time-consuming and expensive. Various studies have examined collaborative issues within a team environment and these have provided initial frameworks applicable for different objectives[2-4]. However, few appear to provide an efficient framework particularly for bridge collaborative design in bridge project teams. The activities during the design stages involve a lot of negotiation and information/data exchange between these design groups. Of the existing group collaborative framework[5,6], the system is usually Web-based and has the functionality that group participants express their opinions and cast their votes electronically. It is does not deal with any issues of data exchange during design processes. Data models for bridge collaborative design system are considerably more complex than those found in commercial applications, due to a number of reasons: (a) the deep hierarchical structure, (b) the multi-represe-ntational data aggregates, (c) the correlation across data representations, and (d) the connections across time. Moreover, the characteristics of the design process are quite unique, namely the iterative, exploratory, and collaborative nature of the design activities.

In spite of the importance of CAD software, research on CAD data models started only in the early 80s and is still not well-understood by the majority of the researchers in the database community . Firstly, research on this area attempted to extend the relational model and, later, started investing resources on object-oriented models. However, most of the literature still focuses on the issues of version modeling and propagation change. Few contributions present models for collaborative design environments. Furthermore, database authors usually investigate object-sharing mechanisms or general aspects of heterogeneous systems, without considering the functional characteristics of the engineering design process and the complex requirements of handling distributed engineering data. On the other hand, researcher of collaborative CAD systems mostly concentrate on PDM

(Product Data Management) systems based on web technology to provide groupware facility, without being in opposition to obsolete data models based on design knowledge.

This article is intended to provide an data representation model for handling data exchange and reason during the process of collaborative design. The specific objectives of the data representation model implementation are:

(1) To link geographically distributed members of a bridge project team via the data representation model for a collaborative bridge process;

(2) To demonstrate the applicability of the data representation to a range of design scenarios.

2 Related Technology

The idea of collaborative among groups using telecommunication is not new and dates back to the 1950s. Back then, T. K van and E. K van gave us an idea of how architects might use the fax technology to serve design communication in the future. From then on many researchers have done a great deal of work in collaborative design system. In this paper, two kinds of technology will be used in system design.

2.1 System architecture

System architecture is the software organization and construction of collaborative design system, which decides which system characteristics to use, which could provide the greatest convenience and flexibility for collaborative partners. We summarize three kinds of system mode for collaborative design system.

(1) Integrated mode: Integrated mode is an integrated collaborative system, which works as a sharable server and thin client for all users. The mode uses an integrated data model and a central management mechanism. The distributed users register with the server and operate the system remotely. Recently, extensive research and development works have been carried out to develop prototype system and methodologies for collaborative system based on integrated mode. Kalay and Khemlani proposed an integrated model to support distributed collaborative design of buildings in 1998, which comprises a semantically-rich, object-oriented database and forms the basis for shared design decisions. However, such an integrated system does not seem able to meet the complex design requirements needed in collaborative environment, such as heavy burden of server and network, Can't fully use existing resources etc.

(2) Distributed mode: this is a fully distributed collaborative system, which works as a thin server for all clients. Xue and Xu introduce a approach for Web-based collaborative concurrent design. In the approach, system, product libraries, and product database which distributed at different locations are linked through the Web. DPME is a distributed process management environment for collaborative building design. The most obvious feature of distributed mode is its flexibility, but without a central server many model interpreters are required between different domain systems.

(3) Integrated-distributed mode: in this mode, geographically distributed designers usually have their own domain system along with a sharable workspace. Lam et al. make an effort toward an Internet-based environment for distributed collaborative performance-based building design and evaluation. Sriram and Logcher proposed a integrated-distributed system in their DICE program. Prasad et al. also presented a integrated and distributed design environment for a collaborative work group. This system mode can be concluded as the open system and has some features i.e., heterogeneous platforms, system flexibility, and system stability. Our approach is based on this mode, using the latest software component technology and agent technology to design an open collaborative system for building design.

2.2 Software component

OO analysis and programming has been an important methodology for the last two decades. Though

the OO approach brought about a major revolution from traditional software development, the promise of large scales code reuse did not become reality. Recently, a new approach, the component-oriented approach, is becoming the focus in software industry.

To support component-oriented development, a number of standards and development tools are available today. Among such tools, J2EE (Java 2 Platform Enterprise Edition) may be the most popular, which defines the standard for developing component-based multi-tier enterprise applications. J2EE simplifies building enterprise applications that are portable, scalable, and that integrate easily with legacy applications and data. J2EE is also a platform for building and using web services.

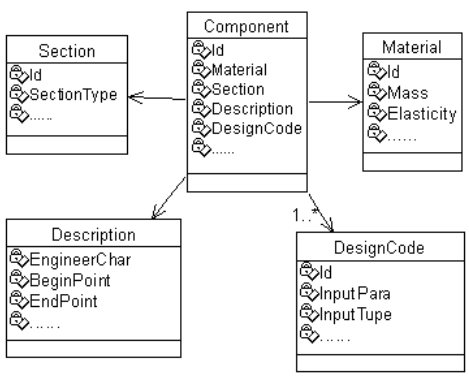


Figure 1 Component model

3 Data Representation Model

Data model is a collection of conceptual tools for describing data, data relationships, data semantics, and constituency constraints. In general, Data model such as relational model involved data value and structure model, rather than the engineering meaning in existing engineering design system. Relational model the most popular data model is a kind of data model for representing simply data. However engineering data is very complex especially in CSCW applications. Engineering data is not single value but combination of value, type, related knowledge and timestamp.

Definition 1. Component: Component has four kinds of content: material, section, description and design code(in Fig1). Component is defined as $E =< M, S, D, C >$. $M = \{concrete, steel, ...\}$ represents the value set of material used in bridge construction.

$S = \{rectangle, circle, ...\}$ represents the value set of section used in bridge component; D represents the value set of description for component; C is a set of design code, where each $c_i \in C$ is a restriction which supported by collaborative system.

Definition 2. Structure/Sub-structure: Structure/Sub-struct- ure is composed of a set of component (in Fig2), which can be described as $S =< E, \theta, F >$, where E is a component set(of size $n = |E|$ and $n \geq 2$).

$\theta = \{//, \perp, \wedge, \vee, -\cdots\cdots\}$ is an arithmetic operators set.

F is an ordered function set, which realizes the evolution from component to Structure/Sub-structure.

4 Architecture of the Bridge Collabo-Rative System

There are two methods to design collaborative system: develop a new system and integrate the existing system. The former method is the more common and can be viewed as a detailing process, where each module must be redesigned. The latter method can be viewed as "integrate" approach to design, where the existing system is integrated. The latter method is selected to design building collaborative system, since our aim is to solve the integration of existing system. The general architecture building collaborative design system includes three layers: user layers, system layers and storage layers.

User layer includes architecture design group, structure design group and other groups. Each group has specific post-process system. The key technology of user layer is how to integrate those existing system.

System layer is the kernel of building collaborative design system, which includes User Interface Agents (UIA), Architecture Design System, Structure Design System, Other Design System, and Storage Interface

Agents (SIA). Among these, Architecture Design System and Structure Design System integrate some application subsystems such as AutoCAD, Sap 2000, and Ansys and so on. UIA and SIA are key technologies of system layer

Storage layer includes Engineering DB, Knowledge Base and System DB. Engineering DB is a distributed system, which stores all data that generated by collaborative system or design groups. Knowledge Base stores all design knowledge, including code, case and experience, which used in collaborative design. System DB stores control information such as user authority, module information, DLL and so on. Design groups can configure System DB according to project characteristic, design stage and personal requirement.

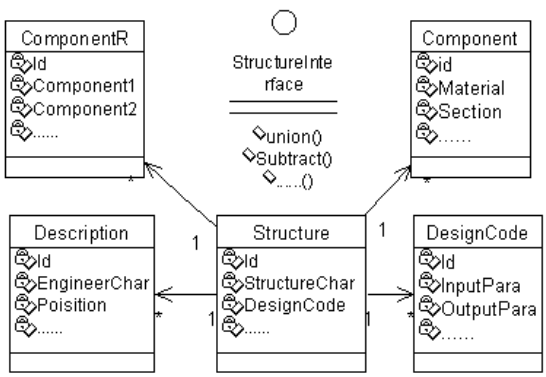


Figure 2 Structure/Sub-structure model

5 Experiment

The prototype system is aimed at collaborative design within a multi-disciplinary bridge project team. It is assumed that the participants in a bridge design activity/task are geographically distributed, as is often the case in real life. For real-life design process on any construction issue, there is usually a leader, such as the project manager, who manages the whole design process. This team leader controls the application forms in the system. The team members based in different geographical locations visit the project web site where the web forms are displayed.

An example of the prototype system follows. A whole bridge is composed of sub-structure and the sub-structure is composed of component. Through gradation division the design project is divided into relation among simple substructure. In Figure 3 a hierarchy model of cable-stayed bridge is depicted where all structures, sub structures and components remain encapsulated.

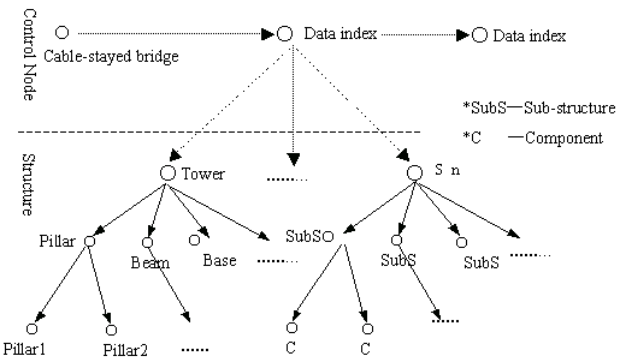


Figure 3 Cable-stayed bridge hierarchy model

6 Conclusition

The collaborative bridge design system divides the bridge design environment into three layers: user layer, system layer and storage layer. The architecture demonstrates clear improvements over existing system in terms of supporting collaborative design, multiple views and customizability. Any given project can be classified as pertaining to more than one design domain and can accumulate component of interest to different designers, shared across multiple views. Although the work is a significant step toward the goal stated at the beginning of this paper, it still falls short from achieving it, such as geometry reasoning, User interface and so on

Achnowledgements

The authors are grateful to the national Natural Science Foundation of China(90715030),Shanghai Natural Science Foundation (06zr14079) and Shanghai Municipal Education Commission foundation(06OZ030) for providing research funding.

References

[1] Chim, Mei Y.,Anumba, Chimay J.,Carrillo, Patricia M. Internet-based collaborative decision-making system for construction. Advances in Engineering Software, Vol35(6),2004,pp357-371

[2] Barbosa, C.A.M.,Feijó, Bruno,Dreux, Marcelo,Melo, Rubens, Scheer, Sérgio. Distributed object model for collaborative CAD environments based on design history. Advances in Engineering Software,Vol34(10), 2003, pp 621-631

[3] Anumba, C. J., Ugwu, O. O., Newnham, L., Thorpe, A. Collaborative design of structures using intelligent agents. Automation in Construction,2002,Vol11(1),pp89-103

[4] Renda, M. Elena; Straccia, Umberto. A personalized collaborative Digital Library environment: a model and an application. Information Processing and Management, vol41(1),2005,pp5-21

[5] Rodriguez, Karina; Al-Ashaab, Ahmed. Knowledge web-based system architecture for collaborative product development. Computers in Industry, vol56(1),2005, pp 125-140

[6] Medlin, Christopher J.; Aurifeille, Jacques-Marie; Quester, Pascale G. A collaborative interest model of relational coordination and empirical results. Journal of Business Research, vol58(2),2005,pp214-222

A Design of Modified PID Regulator for Soccer Robot

Zaixin Liu Jinge Wang Qiang Wang Junfu Zhang Zhongfan Xiang

School of Mechanical Engineering and Automation, Xihua University, Chengdu, Sichuan, China
Email: zhanxinliu@tom.com

Abstract

The critical proportion degree expansion method is used to select PID parameters by analyzing the effects of PID regulator's parameters on the Control Performance. The PID control algorithm was improved by using the integral discretion arithmetic and eliminating the changing impacts of fixed value. And this model is taken as controlled object to do some simulated researches on PID control arithmetic. The simulated curve of distinguished model and experimental curve are almost the same. The application results show that the control algorithm has good control performance, easy to tune the parameters etc. It is fairly appropriate for the engineering application.

Keywords: Digital regulator，PID control，Integral discretion Arithmetic

1 Introduction

The PID control is widely used in industrial control because of its arithmetic not only has a simple structure but also has a better adaptability and good robustness, and its function has been increased a lot than that of the traditional simulated regulator[1] [2] [3]. In the area of designing soccer robot's path, the aim of the path planning is to project the optimal path for the obstacle avoidance[4], as the path should notonly satisfy the beginning position and moving direction, but also meet the target position and moving direction, the control of soccer robot's electromotor has put forward strict requirements[5]. In China, the control of soccer robot's electromotor has used the PID arithmetic at large. And PID control can still be widely applied in the digitalized computer time, as it has the advantages as follows: matured technology, easily being mastered, no need for mathematical model and good effects on controlling. In addition, such kind of control is easy and convenient to use. When a better steady precision is needed, the PID control is applied. And for the larger inertial system, PID control is available. And this kind method of control can tune the parameters (such as: proportion range, integral time, differential time, etc). So this kind of control law has been widely used in the computer control system. How to find out the best adjusting parameter is the most important problem in the PID control[6].

2 The Option of Control Arithmetic

The incremental digital PID regulator has been used in the control system of soccer robot[7], which the output of digital regulator is the increment of the controlling $\Delta u(kT)$.

It shows as follows:

$$\Delta u(kT) = K_p \times \Delta e(kT) + K_i \times e(kT) + K_d \times [\Delta e(kT) - \Delta e(kT - T)]$$

$\Delta e(kT) = e(kT) - e(kT - T)$ is the variation of error at present period.

$\Delta e(kT - T) = e(kT - T) - e(kT - 2T)$ is the error variation of the last period.

Though the incremental arithmetic only changed a little in the aspect of arithmetic, it brings up a few advantages:

1) The digital regulator only has the output of increment, the bad efforts which are caused by the misapplication of computer is not much worse.

2) The switch between manual and automatic has less impact, and it is easy to realize the switch without movement.

3) It is unnecessary to accumulate in the arithmetic. The increment is only related to the recent several samples, so the better control is much easy to realize.

3 The Option of Controller's Param eters

3.1 The selection of sample period[8]

Besides confirming K_p, T_i, T_d, the tuning of digital PID regulator's parameters is still needed to make sure the sample period of system T [9]. The sample period T is an important parameter in the digital control system, and the selection of sample period influences a lot on the performance of system. Standing on the point of the signal's fidelity, the sample period T should not be too long, that is to say: the frequency of sample's angle $\omega_s (\omega_s = 2\pi / T)$ should not be too low. Sample theory has already shown the minimum's frequency $\omega_s \geq \omega_m$, ω_m is the highest frequency of the old signal Considering the control performance, the sample period should be as short as possible, that is to say: the frequency of sample angle ω_s should be as high as possible, but the higher the sample frequency is, the faster the calculating speed of computer should be and the bigger the capacitance of memorizer should be. So the working time and workload increase. And when the frequency of the sample becomes high to some extent, the improvement of system capacity has not been so prominent.

The selection of sample period T is closely related to the control system's dynamic index, the controlled object's dynamic characteristics, the spectrum of disturbing signal and the capacity of computer, etc.. The control process of the PID regulating is finished in the state of intermitting at timing in the single chip microcomputer's control system of soccer robot. So the length of sample period T must guarantee that the programme which is in the intermittence is under the normal procedure. Without influencing the working procedure of the interrupted progarmme, the sample period $T = 0.1\tau$ (τ is the time that electromotor has been relatively delayed).When the working time of the interrupted programme T_Z exceeds 0.1τ, then $T = T_Z$.

So the sample period can be made certain as follows:

$$T = \begin{cases} 0.1\ \tau & T_Z \leq 0.1\tau \\ T_Z & T_Z > 0.1\tau \end{cases}$$

In the system of soccer robot, the time to finish the service programme of intermittence needs about 2~3ms, and the pure delayed time of electromotor is less than 4ms. Therefore, the sample period can be chosen as 4ms. down and the accuracy of control will be improved. When T_d is leaning to be bigger, solodovnikov σ_P is bigger and the adjusting time is longer. When T_d is leaning to be small, σ_P is bigger, the adjusting time is still longer. The perfect transiting producer can be made when T_d is appropriate.

3.2 The selection of pid parameters based on the expanded critical ratio method

The expanded critical ratio method is one kind of tuning methods that can determine the PID regulator's parameters.

3.3 The effects of pid regulator's parameters on the control performance

The simulated PID regulator's tuning is to determine the regulator's parameters *Kp,Ti,Td*, which is required for the control capacity by the technology and is used universal in the projects and known by the technicians.

3.3.1 The Efforts of the Proportional Control on the System Performance

In dynamics, to increase the proportion control will make the systemic movement more active, the speed faster, K_p bigger, the times of vibration more and the adjusting time longer. When K_p is too big, the system will tend to be unsteadily, but if the K_p is too small, it will make the systemic movement slow; In stable characteristic: when the system is steady, to increase the proportion control K_p will reduce the error of steady-state and improve accuracy of control. But to increase K_p can only reduce the error of steady-state and

can not totally eliminate it.

3.3.2 The Efforts of Integral Control on the Control Performance

The integral control always acts on with the proportion control and differential control, which consitutes the PI control or PID control. In dynamic characteristics: the integral control T_i usually descends the steadiness of the system. When T_i is leaning to be smaller, the times of vibration will become more; when T_i is too smaller, the system will be constable; when T_i is too bigger, its efforts on system will be reduced; and when T_i is suitable, the interim will be perfect. In stable characteristics: the integral control T_i can eliminate the error of steady-state and improve the control accuracy of control system. But if T_i is too big, the integral function will be too weak to reduce the error of steady-state.

3.3.3 The Efforts of Differential Control on the Control Performance

The differential control usually works with the proportion control and integral control, which make PD or PID control. The differential control can improve the dynamic characteristics, such as: solodovnikov reduces; the adjusting time becomes shorter; and it allows to increase the proportion control so that the error of steady-state will be cut The tuning steps are as follows:

1) The appropriate sample period T=4ms is selected, and the regulator is controlled by the pure proportion K_p.

2) The proportion K_p is gradually increased to make the expanded critical vibration in the control system. As it shows in Figure 1, during the vibrating procedure, the corresponding period of expanded critical vibration T_s and the increasing expanded critical vibration K_s is taken down. (The expanded proportion is $\delta_s = 1/K_s$)

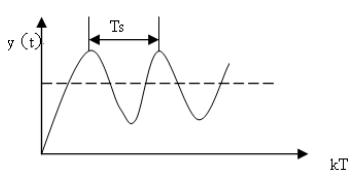


Figure 1 The Experimental Waveforms

3) The control magnanimity is selected. Its definition is the integral ratio of errors' square, which are all produced in the procedure of digital regulator and simulated regulator's transition. In fact, the errors' square integration is taken as assessing function of capacity

$$control\ magnanimity = \frac{\left[\min \int_0^\infty e^2 dt\right]_D}{\left[\min \int_0^\infty e^2 dt\right]_A}$$

4) After choosing the control magnanimity, the control parameters K_p, T_i, T_d are selected as the following rules.

$$K_p = 0.6\delta_s, \quad T_i = 0.5T_s, \quad T_d = 0.125T_s$$

5) According to the obtained tuning parameters, the system is working and the efforts of control are observed. And the parameters must be rectified till the better satisfactory effects of control is obtained.

4 The Improvement of PID Control Arithmetic

4.1 The integral discretion arithmetic

After the integral rectification in the system, the over-big solodovikov will be produced, which is disadvantageous. That is the reason integral discretion arithmetic is introduced, According to the size of deviation in practical production, Integral discretion PID control can be used either to add or to cancel integral effect so as to promote the stability and rapidity of the system, It can not only keep efforts of integration for the system itself, but also reduce the solodovikov that make great improvement of control perfromance.

The integral discretion arithmetic installs a separating improve the capacity of control system. In Figure 3, there is experimental curve, which shows the control of the left wheel of small car of soccer robot that is controlled by the improved PID arithmetic. The fixed

value of epigyny computer is 50. from these two Figs, there are still some errors between the emulated curve and experimental curve, for there are some influencing factors, such as: the gearclearance, filter, etc.integral valve E_0 .When $|e(KT)| \le E_0$, that refers to deviation $e(KT)$ is comparatively small，the application of PID control can guarantee the control precision of system.

When $|e(KT)| \ge E_0$, that refers to deviation $e(KT)$ is comparatively big, the application of PID control can reduce the solodovikiov in larger-scale.

It shows as follows:

$$\Delta u(kT) = K_p \times \Delta e(kT) + K_i \times K_I \times e(kT) + K_d \times [\Delta e(kT) - \Delta e(kT - T)]$$

$\Delta e(kT) = e(kT) - e(kT - T)$ is the error variation of present period.

$\Delta e(kT - T) = e(kT - T) - e(kT - 2T)$ is the error variation of last period.

when $|e(KT)| \le E_0, K_I = 1$;

when $|e(KT)| \ge E_0, K_I = 0$.

4.2 The PID control eliminating the changing impacts of fixed value

Epigyny computer will send an order of fixed speed 40ms/time, and the fixing of small car's speed is frequently under the changing state. For eliminating the frequent changing impact of fixed value, the PID control arithmetic which banishes the changing impact of fixed value is applied here[10]. And the output can only be differential, but the fixed value should not be differential. Such differential control of output can work in any place where the fixed values frequently go up and down, and it can avoid the over-vibration of solodovikov when the fixed value is frequently going up and down.

In a word, the integral separating arithmetic and the PID control which eliminates the changing impacts of fixed value are applied here.

$$\Delta u(kT) = K_p \times \Delta e(kT) + K_i \times K_I \times e(kT) + K_d \times [\Delta y(kT) - \Delta y(kT - T)]$$

$\Delta e(kT) = e(kT) - e(kT - T)$: the variation of present period.

$\Delta y(kT) = y(kT) - y(kT - T)$: the output variation of present period.

$\Delta y(kT - T) = y(kT - t) - y(kT - 2T)$: the output variation of the last period.

5 The Simulated Experimental Waveforms

In Figure 2, the small car's left wheel of soccer robot is taken as example and its simulated curve of the PID control is shown. According to Figure 2, the simulated curve of PID control does a great help to improve the PID control arithmetic and also

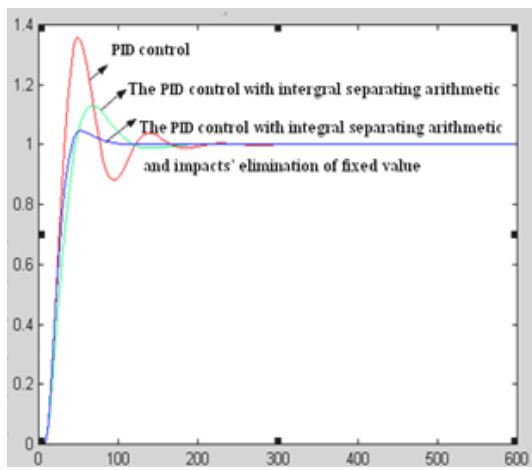


Figure 2 The Simulated Curve of PID Control of Left Wheel

6 Conclusion

In this paper, the model of small car of soccer robot is distinguished by the tested dynamic parameters of soccer robot. The emulated curve of distinguished model and experimental curve are almost the same. And this model is taken as controlled object

to do some simulated researches on PID control arithmetic. Anyhow, such kind of method is finally applied in the practical system and the control effects are much better because of it. The system of soccer robot based on the PID control arithmetic has been successfully applied in the 2006 FIFA Cup China, and brought up the third-place of MiroSot5vs5.

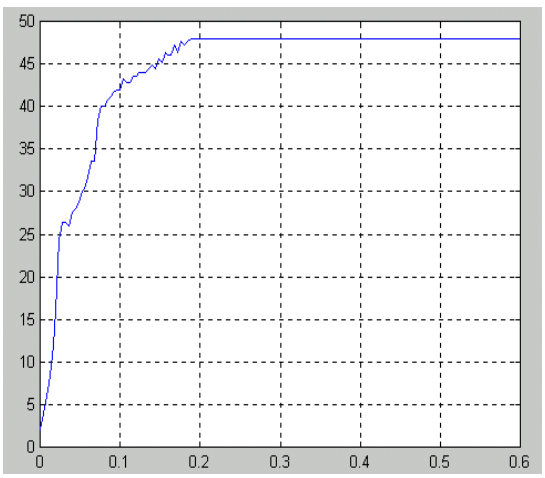


Figure 3 The Experimental Cure of PID Control of Left Wheel

References

[1] ZHANG Xue-yan, ZHANG Jian-xia, The Design of a Single Neuron Adaptive PID Controller and The Simulation of Matlab[J]. Techniques of Automation and Applications，2007,26(9):pp.52-53

[2] LI Jing ,YANG Zhou, A New Designing Method for the Optimal PID Controller[J]. Machinery & Electronics，2007(12):pp.57-59

[3] ZHANG Xing-hua,LI Wei,ZH0U Liu-xi, Particle swarm optimization algorithms for parameter tuning of PID controllers[J]. Computer Engineering and Applications, 2007,43(33):pp.227-229

[4] LIU Zaixin,Wang Jinge,Zhu Weibing. Soccer Robot Path Planning Based On Sinusoid[J]. Journal of Huaqiao University，2006,27(4)：pp.426-428

[5] Liu Zaixin, Wang Jinge, Zhu Weibing, Shooting Algorithm of Soccer Robot Based on Bi-Arc, Journal of Xi'an Jiaotong University, 41(11) , 2007,pp..12-14

[6] GAO Feiyan, TANG Yaogeng. Researching the Tuning Optimization of PID Parameters Based on Attenuated Frequency Characteristic[J] . Process Automation Instrumentation, 2007,28(12) pp.:26-28

[7] Xie Jianying, The Microcomputer Control Technology, Beijing: National Defence University Press, 1996

[8] He Kezhong , Li Wei, Computer Control System, Beijing: Tsinghua University Press, 1998

[9] Wang Furui, Single-microcomputer Testing Control System, Beijing: Beijing University of Aeronautics & Astronautics Press, 1999

[10] Zhang Guofan , Hu Liang, The Microcomputer Control Technology, Shenyang: Northeast University Press, 1995

Research of Human Body Deformable Model Based on Simple Spring-Mass System

Yongqiang Chen Lihua Pen

School of Computer Science, Wuhan University of Science and Engineering, Wuhan, Hubei 430073, China
Email:chenyqwh@hotmail.com

Abstract

Aimed to the stiff and static disadvantages of traditional geometric modeling, we adopted a kind of physical-based geometric modeling, simple spring-mass system, in the human body deformable model. Particle force and equation system's computing methods were researched. Based on the characteristics of human body, the characteristic parameters were managed and the skin surface was discretized into simple spring-mass system. Proved by our experiment, this system could stimulate the deformation change in time and have good display result.

Keywords: Geometric Modeling, Spring-Mass System, Characteristic Parameter

1 Introduction

The geometric modeling of human body is a high-tech technology that sets up an expression of body's data model in computer and supplies good methods to manipulate it by human-computer interaction[1]. The realistic body moves feel abundance and complex. In order to acquire the realistic effect, the body gesture and skin deformation are moveably simulated by computer.

In the common 3D modeling software, objects are modeled by the one method of Wire-frame, Surface and Solid[2]. These methods describe the object's external geometric characters and can not express the physical characters and outside environment factors. The still rigid object is successfully expressed by the traditional geometric modeling technology, on the contrary, the dynamic deformation of elastic object is unsuccessfully described. This is the reason that the human in cartoon and game feels stiff and static.

The physical-based geometric modeling combines the physical and activity features into the traditional modeling. It is a modeling technology that includes geometric information and physical performance. There are some modeling methods in the physical-based geometric modeling, such as Elastic Deformable Model, Particle System, Spring-Mass System, and so on. The Spring-Mass System is a simple linear elastic system and can approximately express the object deformation[3,4]. Many deformations can be simplified to simple linear systems in engineering practice, so the Spring-Mass System can be applied widely.

2 Simple Spring-Mass System

The Spring-Mass System simplifies a deforming object into a linear elastic mass system connected by spring and expresses the deforming process by motional rule of Spring-Mass system. In the Spring-Mass system the motion of a mass is limited by the springs and the spring force generated by mass's movement is calculated through the hooke's law. In the simple Spring-Mass System the elastic deformation fore generated by spring's tensile compression will be added into calculation except the bend and cut deformation forces.

* This paper is partially supported by Hubei Provincial Department of Education Grant # D20081707 and Hubei Digital Textile Equipment Key Laboratory Grant #DTL200702 to Chen Yongqiang.

2.1 Mass Lagrange's equation of motion

In the Spring-Mass System, the motion of a mass must be satisfied to Lagrange's Equation of Motion:

$$m\frac{\partial^2 X}{\partial t^2}+\gamma\frac{\partial X}{\partial t}+\delta_x\varepsilon=f \tag{1}$$

In the equation, X is the position vector of mass. m andγ are quality and sticky density. $\delta_x\varepsilon$ is the elastic internal force and expresses the variational form of elastic internal energy. f is the external force.

The Eq.(1) is a second order partial differential equation. In the left of this equation, the first item is inertia force F_g . The second is damping force F_r generated by obstructive action of medium and the third is deforming force F_d suffered in the mass.

2.2 Mass's forces calculation

Inertia force: $F_g=ma=m\frac{\partial^2 X}{\partial t^2}$.

Damping force: $F_r=C_r v=C_r\frac{\partial X}{\partial t}$, C_r is damping coefficient.

Deforming force: $F_d=\sum_{i=1}^{m}f_{t_i}$. $f_{t_i}=C_{E_i}\cdot\Delta l_{S_i}\cdot n_{S_i}$, f_{t_i} is the tensile compression force generated by the connected spring S_i, Δl_{S_i} is the deforming value of S_i; C_{E_i} is the equivalent elasticity modulus of S_i; n_{S_i} is the unit vector which direction being from P_0 to P_i .

Pull force: $f_l=c_l\cdot dis\cdot n$. The pull force is generated by pulling a point in the model, c_l is the pull index and depended on the deformation performance, *dis* is the movement distance and *n* is the unit vector of pull force.

Gravity: $f_g=mg$. Whether the gravity is added depends on the complex level and realistic effect of model. In the simple condition we can neglect the gravity.

2.3 Discrete data's solving

Suppose that the spring-mass system be made up of n nodes and every motional node must be satisfied by the Lagrange's Equation of Motion. So the system will meet the differential equation system:

$$\mathbf{M}\frac{\partial^2 X}{\partial t^2}+\mathbf{D}\frac{\partial X}{\partial t}+\mathbf{K_{(X)}}X=\mathbf{f(X)} \tag{2}$$

In the Eq.(2), M is a $n\times n$ system quality matrix which being a diagonal matrix, and D is a $n\times n$ system inertia matrix, a diagonal matrix too. $\mathbf{K_{(X)}}$ is a $n\times n$ system rigidity matrix which being a sparse band matrix and **f(X)** is a $n\times 1$ column matrix which shows composite external force of a mass point.

The Eq.(2) is a system of second order partial differential equation about time history. We need add some boundary conditions to solve this system. The boundary conditions are followed as:

$$X\big|_{t=0}=X_0\quad \frac{\partial X}{\partial t}\big|_{t=0}=V_0\quad \frac{\partial X^2}{\partial t^2}=a_0 \tag{3}$$

The usual method to solve this system is to adopt difference and increment method based time in order to change the nonlinear partial differential equation system to linear equation system. In the process, we transform the balance computation of the mass system to that of a series of independent masses.

FDM(finite difference method) is used in the every step. The two previous items can be simplified as follows:

$$m\frac{\partial^2 X}{\partial t^2}=m\frac{1}{\Delta t^2}(X_{t+\Delta t}-2X_t+X_{t-\Delta t}) \tag{4}$$

$$\gamma\frac{\partial X}{\partial t}=\gamma\frac{1}{2\Delta t}(X_{t+\Delta t}-X_{t-\Delta t}) \tag{5}$$

To a mass, the deforming force of spring may be put to the right of Eq.(1), and Eq.(4) and Eq. (5) would be input to the equation, so that the Eq. (1) is changed to as follow:

$$\begin{aligned}&m\frac{1}{\Delta t^2}(X_{t+\Delta t}-2X_t+X_{t-\Delta t})\\&+\gamma\frac{1}{2\Delta t}(X_{t+\Delta t}-X_{t-\Delta t})\\&=f_t(X_{t+\Delta t})+f_l(X_{t+\Delta t})\end{aligned} \tag{6}$$

For further simplification, we use the increment method. When the time step Δt is short enough, $f_t(X_{t+\Delta t})$ and $f_l(X_{t+\Delta t})$ would be replaced by $f_t(X_t)$ and $f_l(X_t)$ severally, and so the equation

(6) is changed to a simple linear equation:

$$\begin{aligned}&(m\frac{1}{\Delta t^2}+\gamma\frac{1}{2\Delta t})X_{t+\Delta t}\\&=f_t(X_t)+f_l(X_t)+m\frac{2}{\Delta t^2}X_t\\&-(m\frac{1}{\Delta t^2}-\gamma\frac{1}{2\Delta t})X_{t-\Delta t}\end{aligned}\qquad(7)$$

Suppose there is a static condition without internal force in the start, the system initial conditions are as follows:

$$\begin{aligned}&X_{t=0}=X_0\qquad\frac{\partial X}{\partial t}\Big|_{t=0}=0\qquad\frac{\partial X^2}{\partial t^2}=0\\&X_{-\Delta t}=X_{t=0}=X_0\end{aligned}\qquad(8)$$

In the Eq.(8), X_0 is the initial position vector.

According to Eq.(7) and system initial conditions in Eq.(8), $X_{t+\Delta t}$ can be calculated very easily. After computing position coordinates of a mass in one time, this mass' coordinates will be renewed and the position coordinates of all masses in the $t+\Delta t$ time can be gotten through iterative computations.

3 Human Body Deformable Model

The simple Spring-Mass System used to human body model needs some steps, such as acquiring model's points, dispersing surface to triangles, parameterizing characters, building spring-mass system's computation module, changing partial characteristic sizes, computing the deformable force after pulling characteristic points, and displaying deformable model.

3.1 Data acquirement and curved surface discretization

The model point's data can be acquired through CT, MR, ultrasonic, and so on. We used slice images gotten from a dead human to reconstruct 3D skin curved surface.

The human skin surface is simply expressed to a triangular grid through dispersing curved surface. In every triangular element, the vertexes form spring-mass system's masses and the edges form springs that every one is connected by two masses. So, every mass in the spring-mass system is connected by some adjacent masses through springs, showed as Figure 1.

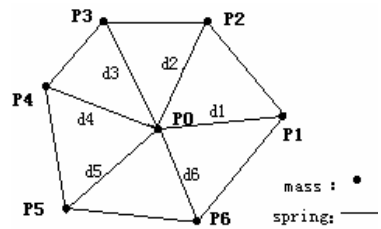


Figure 1 triangular element's spring-mass system

3.2 Characteristic parameterization

The key of characteristic parameterization is to do variable description of character and it's sizes. When changing the characteristic variables, the positions of other point in the architectural feature and partial geometric shape would be changed.

According to the body structure, shape traits and modeling conveniences, the human feature can be classified as architectural feature and modeling feature. The architectural features include head, trunk, left arm, left hand, left foot, right arm, right hand and right foot. The model features include height, shoulder breadth, circumference, and so on.

3.3 Computation and display

In the Spring-Mass System's computation module, suppose that the all of surface is homogeneous and is treated as uniform damping. To every triangular element, the mass's equivalent quality can be seen as $m_{ij}=\frac{1}{N_j}\iint_{\Omega_s^i}\rho ds$, m_{ij} is the value of quality that the quality in the *i* grid is distributed to the *j* mass in the *i* grid, N_j is the mass number of the i grid, and ρ is the surface density function.

Suppose that the damping do directly action to the vertexes, the equivalent damping in mass can be seen as $c_{ij}=\frac{1}{N_j}\iint_{\Omega_s^i}\gamma ds$, c_{ij} is the value of damping that the damping in the *i* grid is distributed to the *j* mass in the *i*

grid, N_j is the mass number of the *i* grid, and γ is the surface distribution function.

After determining the parameter ρ and γ as an unit value, computing all of triangular element in turn, accumulating quality and damping values of the same number mass, we can get the quality value m_i and damping value c_i .

After building the spring-mass system's differential equation system, we change some characteristic variables and compute iteratively according to the linear Eq.(7) and initial conditions in Eq.(8), in the last the final deformable model can be displayed, showed as Figure 2.

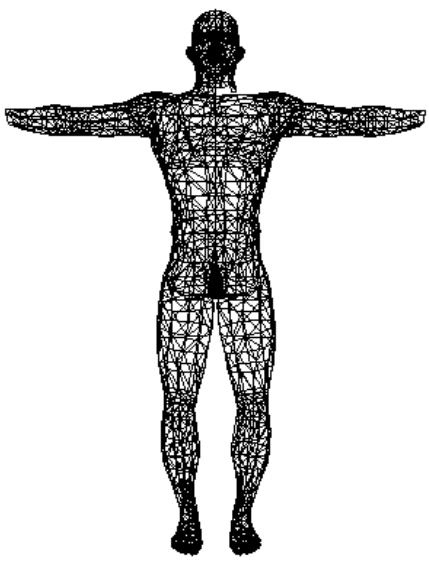


Figure 2 deformed model

4 Conclusion

Compared with the traditional geometric model, the built human body model has more realistic effect. After combined with time variable, this model can express the simulation.

The computational efficiency depends on the precision of curved surface discretization and parametric choice. The model results from the rational simple conditions.

References

[1] Alex Mohr, Michael Gleicher. Building efficient accurate character skins from examples. ACM Trans. Graphics, 2003, 21(3): 562～568

[2] B. Allen, B. Curless, Z. Popovic. Articulated body deformation from range scan data1 In : Proc. SIGGRAPH 2002 , Reading, MA: Addison-Wesley, 2002

[3] Provot X. Deformation constraints in a mass-spring model to describe rigid cloth behavior. Graphics Interface, 1995, 20(3):147~154

[4] Fan qin. Research of physical-based modeling and it's application in the GCAD: [Ph.D. Thesis]. Wuhan: CAD Center of HUST, 1998

Speech Application System Based on MS Agent

Yu Weihong

Transportation Management College of Dalian Maritime University, LiaoNing, Dalian, 116026,China
Email: yuwhlx@163.com

Abstract

Speech recognition and speech synthesis provide a new way of human-computer interaction, and they have shown some significant advantages in a lot of areas such as information processing, education, commerce and so on. The principle of MS Agent has been analyzed and the development steps of speech system based on MS Agent have been discussed in this article. According to the theories of speech recognition and speech synthesis, a speech application system based on MS Agent was developed. And with some illustrations, the running results of the system have been given. So as a conclusion, the advantages of MS Agent were confirmed by our system.

Keywords: MS Agent, Speech Recognition, Speech Synthesis

1 Introduction to Agent and MS Agent

In the field of computer and artificial intelligence, an agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors. In order to achieve some goals, the agent is able to initiate actions.

According to the professor Jennings and Dr.Wooldridge, an agent is an encapsulated computer system that is situated in some environment and that is capable of flexible, autonomous action in that environment in order to meet its design objectives.

MS Agent is a set of programmable software services that supports the presentation of interactive animated characters within the Microsoft Windows interface. It enables software developers and Web authors to incorporate a new form of user interaction, known as conversational interfaces, that leverages natural aspects of human social communication. In addition to mouse and keyboard input, MS Agent includes optional support for speech recognition so applications can respond to voice commands. Characters can respond using synthesized speech, recorded audio, or text in a cartoon word balloon.

The conversational interface approach facilitated by the MS Agent services does not replace conventional graphical user interface (GUI) design. Instead, character interaction can be easily blended with the conventional interface components such as windows, menus, and controls to extend and enhance your application's interface.

2 The Mechanism of MS Agent

MS Agent belongs to user interface agent; that is to say, it is the interface between the computer and its user. As shown in Figure 1, the purpose of the user interface agent is to assist and cooperate with a human user in the performance of some computer-based tasks. This implies, among other things, that the agent must be able to communicate with and observe the actions of the human user and must be able to interact with whatever application programs are used to perform the task.

The agent interacts with shared application programs through the same interface used by a human user in a way that can be observed by a human user. This approach facilitates the reuse of existing applications and supports collaboration by making it easy for the user to know what the agent is doing.

The components of MS Agent for developing a speech application system mainly include:

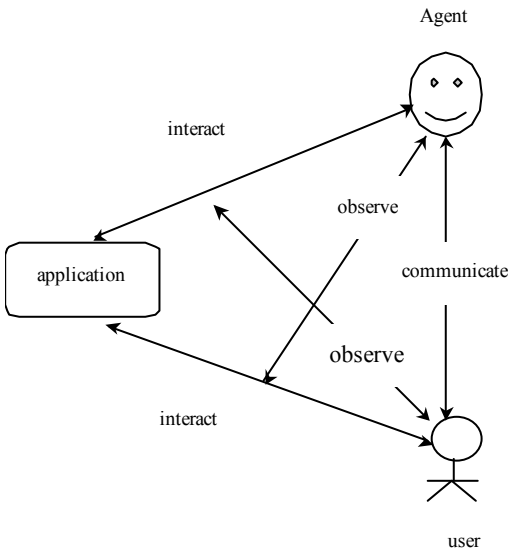


Figure 1 the principle of the user interface Agent

(1) The core components of MS Agent and localization support: msagent.exe.

(2) MS Agent character files. There are mainly four characters for our using: Genie, Merlin, Peedy and Robby. The developers can also use other characters that have been created by third parties with the Microsoft Agent Character Editor.

(3) Text-to-speech engines: cgram.exe.

(4) SAPI 4.0 runtime support, that is to say, speech application programming interface.

(5) Speech recognition engines: Actcnc.exe.

The above components can be downloaded from Microsoft web site. The user must install them before developing the speech application system based on MS Agent.

3 The Development of Speech Application System Based on MS Agent

3.1 Defining agent and character

The easiest way to load the agent control is to select it from the controls menu and just drop it on your form.

Alternatively you can add a reference to the Microsoft Agent Control 2.0 from the references menu item and create an object for the control at runtime, which includes creating an instance of agent object and initialize it, defining the character and the character's request.

3.2 Loading a character and showing or hiding it

Before loading a character, the Agent object must open a connection to the agent sever, that is to say, myagent.connected=True. Otherwise MS Agent will return this error information: Server access failure. *. The attempt to connect with the server failed. Please verify that the server is running and available. And then we can use Load method to load the character, associate the character with the variable referencing the IagentCtlCharacter interface. Load statement must specify Character ID and what character file to open (*.acs file).

To enable text-to-speech output and speech synthesis, a language type must be declared, for example,

myagentchar.LanguageID=&H409

&H409 indicates American English. Agent will automatically attempt to load a TTS engine that matches the character's LanguageID.

3.3 Speech synthesis

Speech synthesis means text to speech,which can convert normal language text into speech for output. Generally, there may be three modules in a speech synthesis system,text analysis module,prosody analysis module and acoustic module.The principle of speech synthesis is shown as Figure 2.

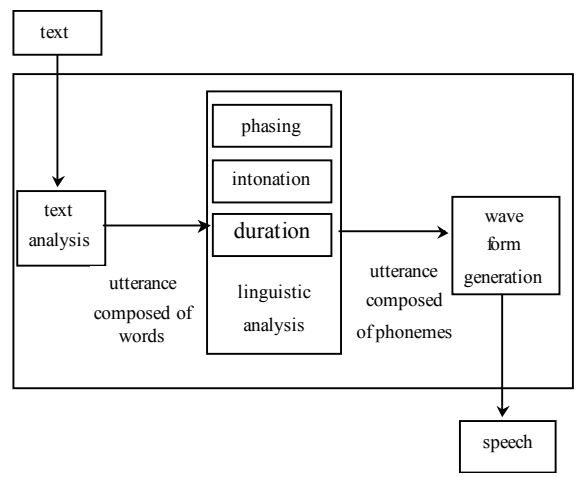


Figure 2 the principle of speech synthesis

MS Agent supports English text to speech. Give an English article, the corresponding speech signal will be generated and be played via sound card and sound box. The function of speech synthesis can be implemented by calling the method *IAgentCtlCharacter::Speak( )*. If you ask the character to read a Chinese text, a .wav file is needed. In speak( ) method you should provide the TEXT parameter with the Chinese text what the character says and specify the location of an audio file (.WAV or .LWV format) in the Url parameter.

The running result of the speech synthesis program is shown as Figure 3.

Figure 3 the running result of the speech synthesis program

First we should choose a character to read for us. If we select a text file, the character can read it for us and the contents what the character reads can be shown in the word balloon.

3.4 Speech recognition

Speech or voice recognition is the ability of a machine or program to recognize and carry out voice commands or take dictation. In general, speech recognition involves the ability to match a voice pattern against a provided or acquired vocabulary.

A complete speech recognition system can be divided into three modules: (1) Voice feature extraction: its purpose is to generate a sequence of acoustic feature vectors that represent temporal and spectral behavior of the speech input. In theory, it should be possible to recognize speech directly from the digitized waveform. However, because of the large variability of the speech signal, it is a good idea to perform some forms of feature extraction that would reduce that variability. (2) Acoustic model and pattern matching (recognition algorithm): To get the best recognition result, The pattern matching algorithm matches a series of acoustic feature vectors with the patterns contained in a acoustic model. (3) Voice and language processing model: This module can deal with language syntax and semantic analysis. The processing flow of speech recognition is shown as Figure 4.

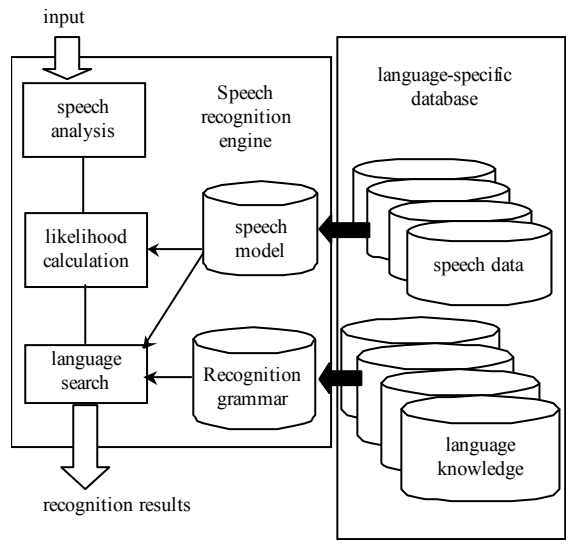


Figure 4 the processing flow of speech recognition

If the speech recognition engine was installed in your computer, you have to press the Scroll Lock key in order to activate the speech recognition or listening capability of the agent control. When the user presses Scroll Lock a status box will display under the character that represents that the character is listening for commands. The agent command event will be activated if the user inputs some commands via the microphone.

Commands are the words or sentences that the user speaks through the microphone, but the character will not recognize any command given by the user unless you program a command into it. To add a command to your character, you should use IAgentComands::Add () method, for example, add this to your code:

mychar.Commands.Add "Time", "What time is it?", "What time is it?", True, True

This adds the command "Time" to your character. To make a character respond to the "Time" command, you should program in command event.

The running result of the speech recognition

program is shown as Figure 5.Figure 5 illustrates that the user presses the Scroll Lock key and the agent character is listening for commands. When the user inputs the command “one” via the microphone, the character responds to this command, plays the action “wave” and speaks “Yes, sir!”.

Figure 5 the running result of the speech recognition program

Speech recognition and speech synthesis jointly constitute a complete human-machine conversation system.

4 Conclusion

This article develops a speech application system based on MS Agent and implements speech recognition and speech synthesis. The system will be applied into the development of educational software. There are some advantages of MS Agent:

(1) MS Agent is simple to program and fun to use. MS Agent's programming interfaces make it easy to animate a character to respond to user input. Animated characters appear in their own window, providing maximum flexibility for where they can be displayed on the screen.

(2) MS Agent should be supportable from any language that supports the ActiveX interface. It includes code samples for Visual Basic, VBScript, JScript, C/C++, and Java.

(3) Each agent character can have a lot of vivid actions and some intelligent features.

(4) Developers can use characters as interactive assistants to introduce, guide, entertain, or otherwise enhance their Web pages or applications in addition to the conventional use of windows, menus, and controls.

References

[1] M. Wooldridge and N. R. Jennings, “Intelligent agents: theory and practice”, Knowledge Eng. Rev., vol. 10(2), 1995,pp. 115-152

[2] Li Tingjun, “Developing an English Number Listening-practicing with MS Agent in VB6.0 Environment”, Computer Engineering and Applications, Vol.41, No.11, 2005, pp.97-99

[3] Dai Chunyan, “ Programming technology of .net ActiveX based on Microsoft Agent”, Journal of Chongqing Technology and Business (Natural Sciences Edition), Vol.21, No.1, 2004, pp.54-56

[4] Yi Ding. “Model and Application of Agent-based Voice Interactive Interface”, Microcomputer Applications. Vol.22, No.3, 2006, pp.39-42

[5] Gu Xuejing, Shi Zhiguo, Wang Zhiliang. “ Research of Affective Robot Based on BDI Agent”. Application Research of Computers. Vol.20, No.4, 2003, pp.24-46

Identifying the Mesophilic And Thermophilic Proteins From Their Amino Acid Composition With V-Support Vector Machines

Yanrui Ding[1,2] Yujie Cai[2,3] Jun Sun[1] Wenbo Xu[1]

1 School of Information Technology, Jiang Nan University, Wuxi, 214122, China

2 Key Laboratory of Industrial Biotechnology, Jiang Nan University, Wuxi, 214122, China

3 School of Biotechnology, Jiang Nan University, Wuxi, 214122, China
Email: yr_ding@yahoo.com.cn; yujie_cai@126.com; sunjun_wx@hotmail.com; xwb_sytu@hotmail.com

Abstract

Many researchers had proved that both single amino acid composition and dipeptide composition can influence protein thermostability. We use ν-support vector machines approach to predict hyperthermophilic protein, thermophilic protein and mesophilic protein from single amino acid composition, dipeptide composition and the combination of the two factors. For the prediction accuracies, we conclude that, single amino acid composition is suitable for prediction of mesophilic protein; dipeptide composition is suitable for prediction of hyperthermophilic protein and thermophilic protein; when considering the combination of the two factors, the prediction accuracy of hyperthermophilic protein is 84.1%, thermophilic protein is 83.4%, mesophilic protein is 84.4%, average accuracy is 84.0%. It shows that the protein thermostability can be predicted properly based on the combination of single amino acid composition and dipeptide composition. Obviously, dipeptide composition is correlative significantly to protein thermostability based on the prediction accuracies.

Keywords: protein thermostability, amino acid composition, ν-support vector machines

1 Introduction

In 2006, Japanese researchers found a protein called “CutA1”, which can act in 148.5℃. As we know, both mesophilic proteins and thermophilic proteins are composed of the same kinds of amino acids. Why thermophilic proteins can maintain their activities at high temperatures? There are many factors that influence the thermostability of proteins[1-13], Such as single amino acid composition[1], disulfide bond[2,3], hydrophobic interactions[4~6], aromatic interactions[7], hydrogen bond[4,5,8,9], ion pairs[4,5,8,10~12], prolines and decreasing the entropy of unfolding[14,15], intersubunit interactions and oligomerization[16], packing and reduction in solvent-accessible hydrophobic surface[5,17,18].

Among these factors, single amino acid composition has long been thought to be correlated significantly to its thermostability [19, 20]. Several investigations [19~24] have been carried out to illustrate the influence of amino acid composition on protein thermostability. These studies showed that thermophilic protein prefers to contain charged, aromatic, and hydrophobic residues comparing to mesophilic protein.

From the facts that mutation of the residues in the mesophilic enzyme to those observed in the thermophilic enzyme (i.e. Ser->Ala and Thr->Ala) produces a mutant enzyme which is 20℃ more stable than the wild type [23], and the tertiary structures of pig heart (37℃) and Thermoplasma acidophilum (55℃) citrate synthases have a high degree of structural homology but only 20% sequence identity[18], we can also know that single amino acid composition play an dominant role on protein thermostability.

In our previous work [13], we studied the influence of dipeptide composition on protein thermostability. At the same time, the influence of single amino acid composition also was studied for comparison. We found the influence of single amino acid composition could be deduced from the influence of dipeptide composition. The characteristic dipeptides not only describe the dipeptide that influence protein thermostability significantly but also show the relationship among significant single amino acids that influence protein thermostability.

Up to now, there is no method to identify the mesophilic and thermophilic proteins based on the primary structure. The first use of the Support Vector Machines (SVMs) approach to predict protein thermostability from single amino acid composition, dipeptide composition, and the combination of the two factors is described here. From the prediction accuracy, we not only know if the SVMs can predict protein thermostability from these factors, but also can deduce which factor that examined is correlative significantly to protein thermostability.

2 Material and Method

2.1 Dataset

At present, there are 10 hyperthermophilic organisms, 3 thermophilic organisms and 52 mesophilic organisms in NCBI COG database[25]. We selected the prokaryotic organisms from them and retrieved useful protein sequences of each organism from NCBI database (http://www.ncbi.nlm.nih.gov/COG). Then, the final dataset was composed of 15187 hyperthermophilic protein sequences, 3974 thermophilic protein sequences and 101868 mesophilic protein sequences.

2.2 N-Support Vector Machines

In this paper, there are three kinds of vectors.

The first kind of vector is defined as: $X_1 = (x_1, x_2, x_3, \ldots, x_{20})^T$ where xi (i=1,2,3,…,20) is the composition of each amino acid in the protein.

The second kind of vector is defined as $X_2 = (x_1, x_2, x_3, \ldots, x_{400})^T$ where xi (i=1,2,3,…,400) is the composition of dipeptide (AA, AC, AD, …, AY, CA, CC, CD, …,CY, …, YA, YC, YD, ... , YY) in the protein.

The third kind of vector is defined as $X_3 = (x_1, x_2, x_3, \ldots, x_{400}, x_{401}, \ldots x_{420})^T$ where xi (i=1,2,3,…,420) is the composition of dipeptide and single amino acid (AA, AC, AD, …, AY, CA, CC, CD,…, CY,…YA, YC, YD, …, YY, A, C, D, …, Y) in the protein.

In ν-SVMs, three labels (1, 2, 3) were used to represent hyperthermophilic proteins, thermophilic proteins and mesophilic proteins separately.

2.3 The training and predicting process

The regularization parameter controls the complexity of the learning machine to a certain extent and influences the training speed. To solve the classification problem properly, it's important to select optimal regularization parameters. In addition, if the data came from different class for training is unbalanced, the prediction system would not good. The 10 fold cross-validation procedure is employed to estimate the classification accuracy for selecting suitable regularization parameter and examining the influence of unbalance data on prediction accuracy. A grid search on regularization parameter using 10 fold cross-validation was carried out on training data. The training data were selected from dataset randomly. The proportion of hyperthermophilic proteins, thermophilic proteins, and mesophilic proteins of training data was examined here. The training data was divided into 10 subsets of (approximately) equal size. Sequentially one subset is tested using the classifier trained on the remaining 10-1 subset. Thus, each instance of the whole training set is predicted once so the cross validation accuracy is the percentage of data which are

correctly classified. Basically pairs of (υ, γ) are tried and the one with the best accuracy of 10 fold cross-validation is picked. Figure 1 describes the process of ν-SVMs training and predicting protein thermostability.

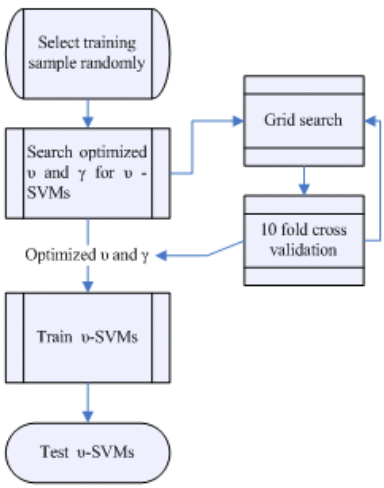


Figure 1 Training and Predicting Process of ν-SVMs

Here, the average accuracy (AA) and the prediction accuracy of each class are used to assess the prediction system.

$$accuracy_i = \frac{p_i}{n_i} \tag{5}$$

$$AA = \frac{\sum_{i=1}^{k} \frac{p_i}{n_i}}{k} \tag{6}$$

Where, p_i is the number of correctly predicted proteins in class i, n_i is the number of proteins in class i, k is the class number.

3 Results and Discussion

3.1 Regularization parameter selection

As we mentioned, 10 fold cross-validation is used to select the optimal parameters, for a certain training sample size, the optimal parameters is selected through "grid search" method, the highest prediction accuracy of 10 fold cross-validation, the most optimal parameters. The result is listed in table 1.

From table 1, we know all the prediction accuracies of 10 fold cross-validation are larger than 79%. This indicates that when the regularization parameter were selected properly, the hyperthermophilic proteins, thermophilic proteins, and mesophilic proteins can be well separated based on single amino acid composition, dipeptide composition, or the combination of the dipeptide composition and single amino acid composition. All prediction accuracies based on single amino acid composition are smaller than the others, this shows that the influence of single amino acid composition on protein thermostability is smaller than dipeptide composition. From No. 1 to No. 3 the prediction accuracies of 10 fold cross-validation ascend, while from No. 4 to No. 9, the prediction accuracies of 10 fold cross-validation decrease, when the training sample size is 3000:3000:3000, most of the prediction accuracies are highest. It is a good proportion for training sample to get a good prediction system.

Table 1 The prediction accuracies of 10 fold cross-validation and optimal parameters of different training samples

sample Size	Amino acid composition (%)	Dipeptide composition (%)	amino acid composition +dipeptide composition (%)
1000:1000:1000	79.8	79.9	80.7
2000:2000:2000	80.6	82.4	82.8
3000:3000:3000	81.2	83.1	83.6
4000:3000:4000	81.9	82.9	83.4
5000:3000:5000	81.7	82.1	82.8
6000:3000:6000	80.7	81.3	82.0
7000:3000:7000	80.1	80.5	81.3
8000:3000:8000	79.7	80.1	80.8
9000:3000:9000	79.7	79.9	80.1

3.2 Prediction result based on single amino acid composition

From table 2, we can easily find the prediction accuracies for mesophilic protein are higher than the other proteins, this shows the single amino acid composition of mesophilic protein is very different from hyperthermophilic proteins and thermophilic proteins.

Table 2 Prediction result based on single amino acid composition

sample Size	H_A (%)	T_A (%)	M_A (%)	AA (%)	Optimized (ν, γ)
2000:2000:2000	78.2	73.8	85.7	79.2	(0.5,155)
3000:3000:3000	80.0	75.7	85.8	80.5	(0.5,100)
4000:3000:4000	80.4	63.3	88.6	77.4	(0.5,135)
5000:3000:5000	81.7	57.4	89.4	76.2	(0.5,155)
6000:3000:6000	85.4	59.2	89.1	77.9	(0.5,130)
7000:3000:7000	86.7	54.2	89.4	76.8	(0.5,130)
8000:3000:8000	85.7	49.9	90.1	75.2	(0.5,140)
9000:3000:9000	87.5	46.9	89.9	74.8	(0.5,140)

Train sample size=hyperthermophilic:thermophilic:mesophilic

Because the number of thermophilic protein sequences is very small comparing with hyperthermphilic protein sequences and thermophilic protein sequences, we have to increase the amounts of hyperthermophilic and mesophilic protein sequences to consider the influence of training sample size on prediction accuracy. Although, average accuracy has only a little change, the prediction accuracies for thermophilic protein decreased dramatically. The sequence amount is more unbalanced, the accuracies for the thermophilic protein are lower. From No. 3 to No. 9, the average accuracy decreased only 5.7%, but the accuracies for thermophilic protein decreased 28.8%. As we know, microorganisms can be classified according to their optimal growth temperature [26], T_{opt}, roughly into four groups: psychrophilic ($0< T_{opt} <20$℃), mesophilic ($20< T_{opt} <50$℃), thermophilic ($50< T_{opt} <80$℃) and hyperthermophilic ($80< T_{opt} <120$ ℃). Obviously, thermophilic protein is a transition protein between mesophilic protein and hyperthermophilic protein, and if the training data is unbalanced, ν-SVMs will receive more information from hyperthermophilic and mesophilic protein and less 'noise' from thermophilic, then ν-SVMs can predict hyperthermophilic and thermophilic protein with the accuracy around 90%, but the average accuracy is relative lower. Also, we had checked the selection process of ν-SVMs parameters carefully. The prediction results under unbalanced dataset were not suffered from overtraining. Considering better prediction accuracy of each class and average accuracy, when the training size is 3000:3000:3000, the prediction result is best.

3.3 Prediction result based on dipeptide composition

In our previous work [13], we had found that the dipeptide composition could provide more information of protein thermostability than single amino acid composition. In order to compare with results in table 2, we use the same protein sequences as in table 2 to train the ν-SVMs.

Table 3 Prediction result based on dipeptide composition

sample Size	H_A (%)	T_A (%)	M_A (%)	AA (%)	Optimized (ν, γ)
1000:1000:1000	78.8	83.9	77.5	80.1	(0.5,300)
2000:2000:2000	81.7	84.6	79.1	81.8	(0.5,340)
3000:3000:3000	83.0	85.1	83.0	83.7	(0.5,280)
4000:3000:4000	86.9	77.1	86.4	83.5	(0.5,260)
5000:3000:5000	89.3	71.6	87.9	82.9	(0.5,280)
6000:3000:6000	90.7	68.9	88.1	82.6	(0.5,280)
7000:3000:7000	91.7	64.4	88.7	81.6	(0.5,280)
8000:3000:8000	91.9	58.6	89.0	79.8	(0.5,320)
9000:3000:9000	92.3	57.7	89.1	79.7	(0.5,320)

Train sample size=hyperthermophilic: thermophilic:mesophilic

Table 3 shows there are higher prediction accuracies for hyperthermophilic protein and thermophilic protein than mesophilic protein. Comparing with table 2, the unbalanced data have the same influence on prediction accuracies as that in table 2. When the sample size is balance, the predict accuracies for mesophilic protein in table 3 is lower than that in table 2, but when the sample size is unbalance, the predict accuracies for mesophilic protein in table 3 is similar as that in table 2. Because the dipeptide composition is 400 dimensions, it includes more information than single amino acid composition. Then, average accuracy based on dipeptide composition have an improvement than those based on single amino acid composition. Obviously, the training sample size, 3000:3000:3000 is the best.

3.4 Prediction result based on the combination of dipeptide composition and single amino acid composition

From the above results, we can find single amino acid composition is better to predict mesophilic proteins and dipeptide composition is better to predict hyperthermophilic and thermophilic proteins. Here, we combined these two factors to predict protein thermostability. The protein sequences for training and predicting in table 4 are same as that in table 2 and in table 3. The results were list in table 4.

In table 4 H_A: hyperthermophilic protein accuracy; T_A: thermophilic protein accuracy; M_A: mesophilic protein accuracy.

Table 4 Prediction result based on the combination of dipeptide composition and single amino acid composition

sample Size	H_A (%)	T_A (%)	M_A (%)	AA (%)	Optimized (ν, γ)
1000:1000:1000	80.4	81.2	79.8	80.5	(0.5,500)
2000:2000:2000	82.9	81.7	83.3	82.6	(0.5,560)
3000:3000:3000	84.1	83.4	84.4	84.0	(0.5,560)
4000:3000:4000	87.2	79.1	87.2	84.5	(0.5,740)
5000:3000:5000	88.9	71.0	88.8	82.9	(0.5,660)
6000:3000:6000	90.1	70.2	89.0	83.1	(0.5,580)
7000:3000:7000	91.3	66.6	89.4	82.4	(0.5,620)
8000:3000:8000	90.9	62.6	90.3	81.3	(0.5,680)
9000:3000:9000	92.1	62.6	90.3	81.7	(0.5,680)

We find the predict accuracies for three kinds of protein is balance and the average accuracy is higher than that in table 2 and table 3 with balanceable data. For the unbalanced data, although the prediction for hyperthermophilic and mesophilic proteins in table 4 is similar to table 3, the prediction for thermophilic proteins in table 4 is higher than those in table 2 and table 3. Obviously, the overall prediction result based on the combination of dipeptide composition and single amino acid composition is highest. For the training sample size, 3000:3000:3000, the prediction accuracy of hyperthermophilic protein is 84.1%, thermophilic protein is 83.4%, mesophilic protein is 84.4%, and average accuracy is 84.0%. It's a better result for predicting protein thermostability using ν-SVMs. After all, there are many factors influence protein thermostability.

We consider the training sample size is 3000 is enough for predicting each kind of proteins; larger sample size will improve the prediction CPU time significantly.

4 Conclusion

In this article, we predicted the thermostability of protein by collected together all the sequence. The high prediction accuracy proved that there was the overall trend in mesophilic and (hyper)thermophilic proteins which is implicated in the protein primary structure.

Thermophilic microorganisms are the source of novel thermostability enzymes. Some thermophilic enzymes such as DNA polymerases, amylases from thermophilic microorganisms had been used successfully. For enzymes which can't be found in thermophilic microorganisms, modern techniques like mutation genesis and gene shuffling will lead to convert mesophilic enzyme to thermophilic enzyme. Here, we provide a powerful method (ν-SVMs) which is easily to predict thermostability of protein from primary structure.

Acknowledgements

This work is supported by the innovation team-building project of jiangnan university, JNIRT0702.

References

[1] K Suhre, JM Claverie, "Genomic correlates of hyperthermostability, an update," *J Biol Chem*, 278(19), 2003, pp. 17198-17202

[2] G. Cacciapuoti, M. Porcelli, C. Bertoldo, M.D. Rosa, V. Zappia, "Purification and characterization of extremely thermophilic and thermostable 5'-methylthioadenosine phosphorylase from the archaeon Sulfolobus solfataricus. Purine nucleoside phosphorylase activity and evidence for intersubunit disulfide bonds," *J. Biol. Chem*, 269, 1994, pp. 24762-24769

[3] M. Matsumura, G. Signor, B.W, "Matthews, Substantial increase of protein stability by multiple disulphide bonds," *Nature*, 342, 1989, pp. 291-293

[4] M Robinson-Rechavi, A Alibes, A Godzik, "Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of Thermotoga maritime," *J Mol Biol*, 356(2), 2006, pp. 547-557

[5] M Sadeghi, H Naderi-Manesh, M Zarrabi, B Ranjbar, "Effective factors in thermostability of thermophilic proteins," *Biophys Chem*, 119(3), 2006, pp. 256-270

[6] K Saraboji, MM Gromiha, MN Ponnuswamy, "Importance of main-chain hydrophobic free energy to the stability of thermophilic proteins," *Int J Biol Macromol*, 35(3-4), 2005, pp. 211-220

[7] L. Serrano, A.R. Fersht, "Aromatic-aromatic interactions and protein stability. Investigation by double-mutant cycles", *J. Mol. Biol*, 218, 1991, pp.465-475

[8] G Vogt, S Woell, P Argos, "Protein thermal stability, hydrogen bonds, and ion pairs", *J Mol Biol,* 269(4), 1997, pp.:631-643

[9] E Querol, JA Perez-Pons, A Mozo-Villarias, "Analysis of protein conformational characteristics related to thermostability", *Protein Eng*, 9(3), 1996, pp. 265-271

[10] E Bae, GN Jr Phillips, "Structures and analysis of highly homologous psychrophilic, mesophilic, and thermophilic adenylate kinases", *J Biol Chem*, 279(27), 2004, 28202-28208

[11] A Fish, T Danieli, I Ohad, R Nechushtai, O Livnah, "Structural basis for the thermostability of ferredoxin from the cyanobacterium Mastigocladus laminosus", *J Mol Biol*, 350(3), 2005, pp. 599-608

[12] GI Makhatadze, VV Loladze, DN Ermolenko, X Chen, ST Thomas, "Contribution of surface salt bridges to protein stability: guidelines for protein engineering", *J Mol Biol,* 327(5), 2003, pp.1135-1148

[13] Y.R. Ding, Y.J. Cai, G.X. Zhang, W.B. Xu, "The influence of dipeptide composition on protein thermostability", *FEBS lett,* 569, 2004, pp. 284-288

[14] C. Li, J. Heatwole, S. Soelaiman, M. Shoham, "Crystal structure of a thermophilic alcohol dehydrogenase substrate complex suggests determinants of substrate specificity and thermostability", *Proteins Struct. Funct. Genet*, 37, 1999, pp. 619-627

[15] B.W. Matthews, H. Nicholson, W.J. Becktel, "Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding", *Proc. Natl. Acad. Sci.* 84, 1987, pp. 6663-6667

[16] H. Moriyama, "The crystal structures of mutated 3-isopropylmalate dehydrogenase from Thermus thermophilus HB8 and their relationship to the thermostability of the enzyme", *J. Biochem*, 117, 1995, pp. 408-413

[17] J Chen, WE Stites, "Replacement of staphylococcal nuclease hydrophobic core residues with those from thermophilic homologues indicates packing is improved in some thermostable proteins", *J Mol Biol*, 344(1), 2004, pp.271-280

[18] IN Berezovsky, EI Shakhnovich, "Physics and evolution of thermophilic adaptation", *Proc Natl Acad Sci U S A*. 102(36), 2005, pp. 12742-12747

[19] S. Kumar, C.J. Tsai, R. Nussinov, "Factors enhancing protein thermostability", *Protein Eng,* 13, 2000, pp. 179-191

[20] C. Vieilie, G.J. Zeikus, "Hyperthermophilic enzymes: Sources, uses, and molecular mechanisms for thermostability", *Microbiol. Mol. Biol. R*, 65, 2001, pp. 1-43

[21] P.J. Haney, J.H. Badger, G.L. Buldak, C.I. Reich, C.R. Woese, "Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic Methanococcus species", *Proc. Natl. Acad. Sci*, 96, 1996, pp. 3578-3583

[22] A.Szilagyi, P. Zabodszky, "Structural differences between mesophilic, moderately thermophilic and extremely thremophilic protein subunits: results of a comprehensive survey", *Structure*, 8, 2000, pp. 493-504

[23] M.M. Gromiha, M. Oobatake, A. Sarai, "Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins", *Biopys. Chem*, 82, 1999, pp. 51-67

[24] R.J. Russell, G.L. Taylor, "Engineering thermostability: lessons from thremophilic proteins", *Curr. Opin. Biotech*, 6, 1995, pp. 370-374

[25] R.L. Tatusov, E.V. Koonin, D.J. Lipman, "A genomic perspective on protein families", *Science*, 278, 1997, pp. 631-637

[26] K. Andrey, L. Rudolf, "Ion pairs and the thermotolerance of proteins from hyperthermophiles: a 'traffic rule' for hot roads", *Trends in biochemical sciences*, 26, 2001, pp. 550-556

The Research and Application of Freeport Communication of SIEMENS PLC

Jie Chen Xuejun Hu Lixin Xu

School of Electronic & Information Engineering, Wuhan Institute of Technology, Wuhan, Hubei 430073, China
Email: ch58j@163.com

Abstract

The Communication between various PLC or PLC and special intelligent devices can be solved by programmed protocol under freeport mode of SIEMENS S7-200 PLC. Based on freeport communication mode function,PC/PLC long-distance control the smoothly startup and stop of asynchronous motor.

Keywords: S7-200PLC, Freeport, Protocol, PC/PLC Communication, asynchronous motor

1 Introduction

With the improvement of PLC technology, PLC will instead of the traditional relay gradually as the greater function, faster speed, smaller cubage, lower cost and higher reliability and has to be the standard equipment of industrial control. PLC can constitute distributed control system and be medium or lower equipment of layer control process system as it is easy to realize collecting, disposing and controlling data. Such as product line control system, equipment running control system, flexibility machining and produce system and so on. The application area of PLC not only include the area of electric, oil, car, chemical industry, iron, mining, building materials, engine machining, transportation, light spin industry, environment protection, but also will turn to using PLC gradually in some situation where using computer.

The requirement for interconnection between various PLC and PLC and the other control equipment also should be improve for the improvement of factory's automatic. At present PLC's relevant communicate interface(such as RS-232,RS-422,RS-485 and so on) can solve ASCLL code's communication between the same machine and host computer.

2 The Communication Function of Siemens S7-200 Series PLC

The communicate ability of S7-200 series' PLC can realize from both software and hardware. In hardware, except for CPU 210, all types of CPU integrate one or two communication port interiorly. The communication port is the standard port of RS485. The definition of chip pin can refer to from the reference. At first, in physically, we should make use of relevant cable to constitute the complicated communication network with many kinds of communication function to ensure that do not need the outer communication module. S7-200 can connect to spot' equipment to compose different communication network by various method with the communication port, realize the control of factory's spot. Besides, by PROFIBUS's spot bus, it can exceed the control of spot to transfer the control information quickly, and connect to workshop or factory's manage system, can constitute modern control and manage network. It can't be compared with a normal minitype PLC.

There are 3 work modes for S7-200 series' PLC communication port: PPI mode, Freeport mode and PROFIBUS-DP mode[1].

2.1 PPI mode

PPI communication protocol is a special one for

s7-200 series; It adopts RS485 signal level in physical. PLC is default in this mode. PPI is one principal and subordinate protocol. Primary station equipment send requests to subordinate station equipment, then subordinate station equipment answers. And subordinate station device never send message on its own, just waiting for the requests from primary station equipment and answer them. It doesn't need extra hardware module and software by PPI mode to communicate, so the convenience is out of question. But as PPI protocol is not open, peripheral equipments are required for supporting PPI protocol. Therefore it has some restriction for PLC's applicability which use for current controller. To solve the problem, The S7-200 PLC series' RS485 communication port also provides another work mode: Freeport.

2.2 Freeport mode

Freeport mode is one communication protocol with completely opened work mode. If PPI mode is that peripheral equipment adapts PLC, then Freeport mode is that PLC adapts peripheral equipment. In the Freeport mode, peripheral equipment is not restricted by PPI protocol, and the equipment which doesn't support PPI protocol can also communicate with S7-200 series. Freeport mode makes S7-200 PLC can communicate with any other equipment and controller which has open communication protocol. That's communication port depends on peripheral equipment in S7-200 PLC's Freeport mode, PLC adapts peripheral equipment by program, user themselves can also define communication protocol[2]. So the area of corresponding increasing amazingly, makes control system's configure more flexible, convenience and with high cost performance.

2.3 PROFIBUS-DP mode

PROFIBUS was constituted in German in 1986 and completed in 1990. At present, It's the most extensive bus which applies to face to open industrial spot. To constitute based on spot bus's centralized and discrete control network, connect to PROFIBUS spot bus network by PROFIBUS-DP communication port[3]. The character is to enlarge PLC's control ability and scope by bus's communication ability.

3 Freeport Communication Mode's Development and Application

As there are many PLC's producers, and every company has their own specific network communication method, every type of PLC doesn't compatible with each other, and there doesn't have an explicit and same standard. Enterprises often buy different producers' different type's PLC as the whole set's equipments are required or considering some specific requirements. It causes the communication is very difficult between the different company's PLC. Siments Company's S7-200 PLC's Freeport's Freeport communication mode can communicate with any equipment which has open communication protocol and controller by it's custom protocol, solving the communication problem between every company's PLC and PLC with other intelligent equipment.

Through the standard RS485's series communication network and custom or independent choice's communication protocol, constitute PLC and industrial control computer's distributed network. In the practical application, combine system's development cost and function's practicability, this method has better flexibility and cost performance, therefore it can be used widely in the systems' design interiorly and overseas. The applying is very successful in some area which requires complex communication requirement, such as city illumination's centralized control, train door's air-conditioning's centralized control.

When S7-200 CPU's communication port work in the Freeport mode, the host computer Step7-Micro/WIN32 software can't set up the normal programming communication and software monitoring with it. Then it can by program and special function register SM0.7 to control communication port's communication mode, that's

define communication port as ordinary PPI communication to realize programming communication. The following is Freeport work mode which using communication port, custom protocol realize PC/PLC control's scheme design of the winding asynchronous motor start and stop smoothly.

3.1 The relay control scheme of winding asynchronous motor's start smoothly

The winding asynchronous motor rotor series resistance starting, it can not only restrict starting current, but also increase starting torque, it's widely use in manufacture engine which need heavy haul and starting frequently. The work procedure is as following. Press the starting button SB1, contactor gets electric then pull in and self-locking. Motor stator get through power, rotor series get through all resistance starting, at meantime timer KT1 loop starts to timing, KT1's normal open contacts close when the delay time is up. Then KM5 gets electric and close, short-circuit the first level's resistance, KM5's auxiliary contact gets through KT2, when the giving time is up, close KM4, short-circuit the second level's resistance, starting timer KT3, it goes on until four level's resistance short-circuit all. Motor starts when has rated voltage.

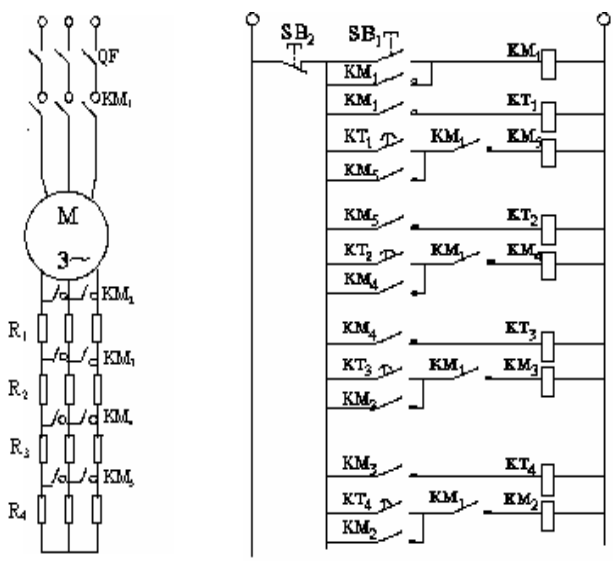


Figure 1 Relay control scheme

3.2 Freeport mode to realize the control of winding asynchronous motor's start smoothly PC/PLC control scheme

3.2.1 In Freeport mode, self-define protocol to realize PC/PLC control controlled device's programming design

Freeport mode makes S7-200 PLC can communicate with any other equipment and controller which has open communication protocol. That's communication port depends on peripheral equipment in S7-200 PLC's Freeport mode, PLC adapts peripheral equipment by program, user themselves can also define communication protocol. So the area of corresponding increasing amazingly, makes control system's configure more flexible, convenience and with high cost performance.

The Self-define communication protocol can be finished by the most two instructions XMT and RCV. The others are initial instructions, interrupt instructions, mistake identify instructions and so on. Figure 2 is the main program.

3.2.2 Test and Operate

The scheme is Freeport communication realize host computer's software supervising. When testing S7-200CPU's Freeport's communication, with PC/PPI cable to connect CPU and PC, running serial debugging software, such as the HyperTermianl application which intergrated by Windows operation system, send data to CPU, or receive data from CPU, get through the communication between host computer and slave computer.

Firstly, in PLC's stop mode, PC/PPI cable gets PC connect to PLC by PPI communication protocol, after downloading self-define protocol and motor control program, cut the connection between Step7-Micro/ WIN32 and CPU. Open Hyper Terminal of Windows starting menu, select icon and specify one connecting name, choose connect to PC and connect PC/PPI cable's series communication port(This passage chooses COM2); Secondly, set up communication port's

parameters, communication speed is 9600 bits/s, data bit is 8, no parity check, stop bit is 1, no flow control.

Then make S7-200CPU's mode optional switch on Run, do the following steps. Send ON command on Hyper Terminal, program begins. When Q0.0, the message send out, contactor KM1 gets through, rotor series all resistance starting, delay 5s, Q0.1 gets through, contactor KM5 gets through, short-circuit rotor resistance R4 makes the third resistance starting; Then delay 5s again, Q0.2 gets through, contractor KM4 gets through, short-circuit rotor resistance R5 makes the second resistance starting. Then delay 5s again, Q0.3 gets through, contractor KM3 gets through, short-circuit rotor resistance R2 makes the first resistance starting; Then delay 5s again, Q0.4 gets through, contractor KM2 gets through, makes direct-on-line start. Send OFF command on Hyper Terminal causes all contractor cut and motor stop.

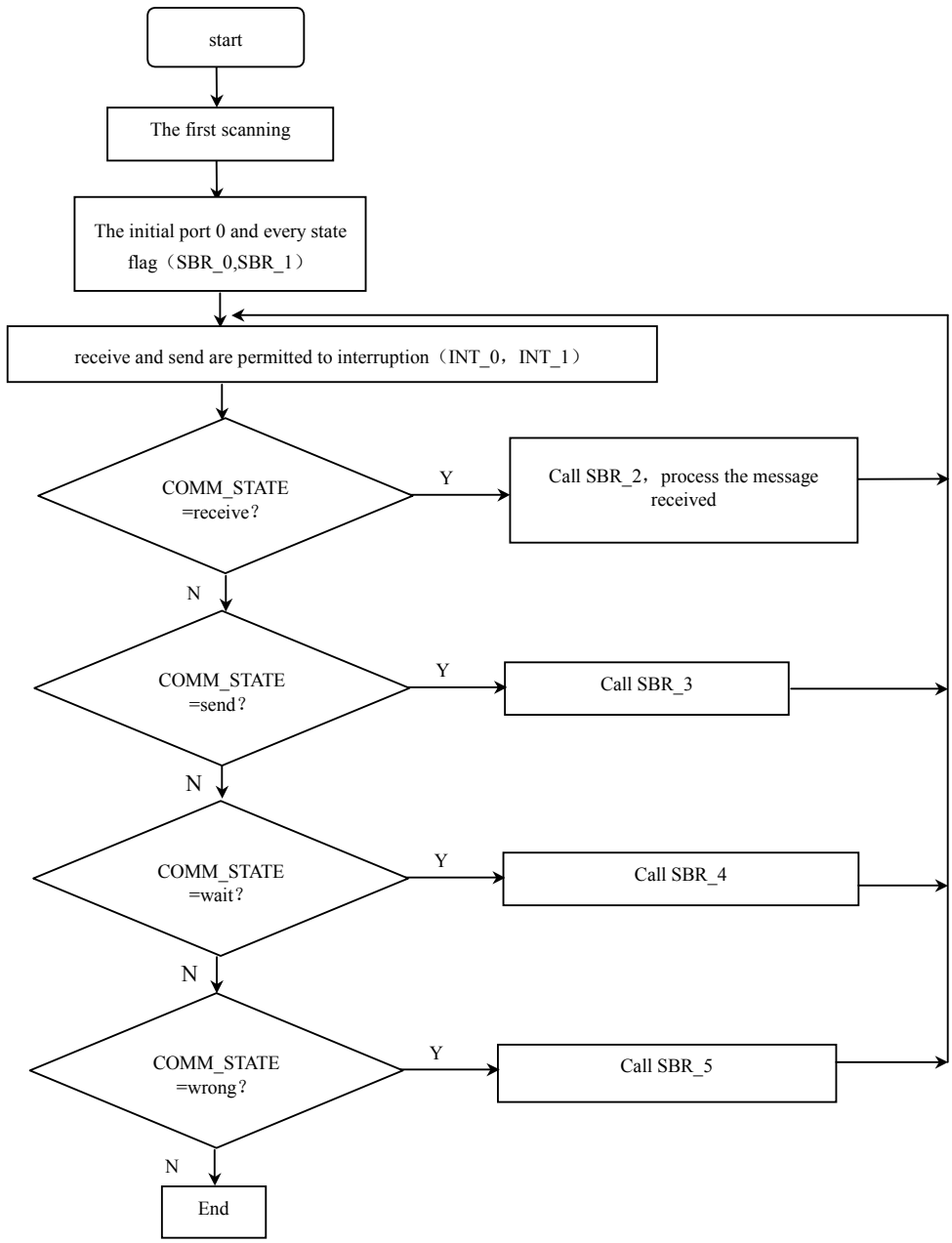


Figure 2 PC/PLCThe motor start-stop control program diagram by freeport mode communication

So with self-define protocol, in Freeport communication mode, PC/PLC master and slave computers control winding asynchronous motor's start and stop smoothly comes true. In practice, it proves that control is more flexible and simple by S7-200CPU's Freeport communication port.

4 Conclusion

With the development of factory's automatic, The requirement for interconnection between various PLC and PLC and the other control equipment also should be improve. The Communication between various PLC or PLC and special intelligent devices can be solved by programmed protocol under freeport mode of SIEMENS S7-200 PLC. This method has better flexibility and cost performance. The applying is very successful in some area which requires complex communication requirement, such as city illumination's centralized control, train door's air-conditioning's centralized control.

Reference

[1] SIMATIC S7-200 PLC System Manual, 2002.3

[2] Xingjian Cai, SIMATIC S7-200 PLC, Press of Beihang University,Beijing,2000

[3] Hongfang Tian, Yinhong Li, "Serial communication between PLC and computer", Microcomputer Information ,2001,(3)

[4] Renguang Yuan, Application technology and example of PLC, Press of South China University of Technology, 2001

[5] Guang liu, Ping Wang & Jianchun Xing, "Newest development of PLC", PLC&FA, 2002,3(10)

[6] Jinkun Liu, Intelligent Control, Beijing: Electronic Industrial Public, 2005

[7] Meixian Wu, Xueliang Zhang, Summarization of BP Neural Networks Improvement, Taiyuan University of Science & Technology Transaction, Vol.26,No.2,2005

[8] Guoyong Li, Intelligent Control Beijing: electronic industrial public, 2005

[9] Xiancai Gui, realization of BP Networks And Their Application on MATLAB, Zhanjiang Normal College Transaction Vol.25,No.3,2004

[10] W.J. Palm, Modeling Analysis and Control, 2nd ed.,John Wiley & Sons, New York, 2000

[11] R.C. Dorf and R.H. Bishop, Modern Control Systems, 10th ed., Science Press, Beijing, 2005

A Load Balancing Algorithm Based on The Initiative Feedback and Nearby Service

Fan Yang Qingping Guo

School of Computer Science and Technology, Wuhan University of Technology, Wuhan, Hubei 430063, China
Email: jiessie_yf@126.com

Abstract

Load balancing is the key of distributed cluster system research, load balancing algorithm is the most important factor of the symmetrical distribution of tasks .Introducing some common scheduling algorithm and analyzing them. Considering the characteristic of Video Grid Cluster, and researching on the model of user behavior, we create the initiative feedback and nearby service load balancing algorithm, which uses the method of dynamic feedback in a certain cycle, and adjusts the allocation of the server nodes based on the spare capability of them, to provide customer a good video service. In the end, we prove that this algorithm adapts to the Video Grid System.

Keywords: Initiative feedback, Load balancing, The nearby Service, Video Grid, User behavior

1 Introduction

Multimedia Network Server have to face rapid growth in number of visit because of the rapid growth of Internet, the server should be of concurrency accessing capability, so the processing and I/O capability of server became the bottleneck of serving. To solve this problem, we can use the high-powered or SMP computer, or connect many servers to the whole, through the parallel processing and the fast information intercourse between each other to enlarge the capability, the later become the main method of constructing high-performance server. The background this paper research is the Video Grid System, which take distributed cluster as framework, to provide the high quality video service by the algorithm what we research.

Load balancing algorithm of cluster server has two main aspects, which are static and dynamic algorithm, the static one doesn't consider the real load condition of the server node, but dynamic algorithm concentrate on it. Research has mainly focused on dynamic load balancing study in recent years, DNS-based load balancing algorithm is one of the hot research, in reference [1], DNS collect the load information of the server node for load balancing, but DNS need to exchange information with server nodes constantly, which increase network load. This paper shows us a new strategy that the server nodes reflect the load information status positively, which simplifies load balancing algorithm for the allocating request, to avoid grate network load caused by exchanging information constantly. In order to enable services to marginalized, and reduce network bandwidth load in a greater extent, we propose the nearby service load balancing algorithm.

2 The Model Based on the Initiative Feedback of Load Balancing of Video Grid

2.1 Structure of video grid system

In the existing network, there are large number of internet user, but the frequencies of the use of computer resources is quite low, based on the target of resource sharing, the concept of grid had been imposed for many years, which hope that users can make use of idle computer resource on local area network or internet, providing service to users like power grid [2]. The target of Video Grid System is to set up an infrastructure of video service,

which takes dynamic data information of streaming media as managed objects, and sees all the streaming servers which are in various parts as a whole, to achieve the sharing of program source and server capability. The structure of Video Grid system is shown as Figure 1.

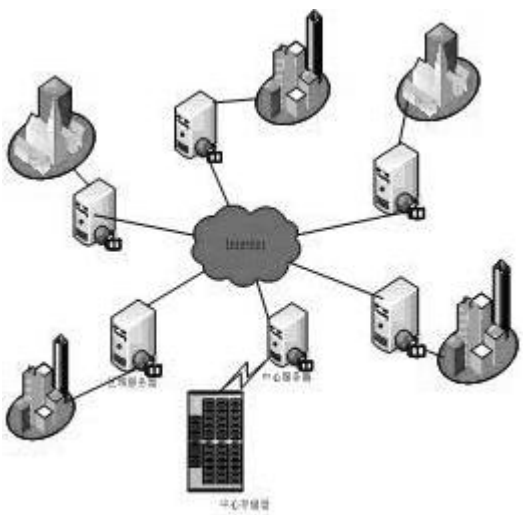


Figure 1 structure of Video Grid system

Video Grid System is made up of one central server node and a number of edge server nodes, the former keeps the load information, the server capability and the network load information of the edge server nodes. When a client request reaches, according to the load balancing algorithm, the central server node inquiry the load information of each edge server nodes, choose the most suitable node to provide service. So the suitable load balancing model and algorithm is the key of the success of this system. Besides, considering the limit of the network bandwidth, the system provide client the service by the nearby sever as far as possible, in other words, not only considering the capability and load balancing information of the server nodes but also the distance between the client and the edge server node, for this perspective, we propose the nearby service load balancing algorithm.

2.2 The model based on the initiative feedback load balancing

The common model of load balancing takes the load balancer as the carrier, which sends the inquiry request to each server node positively in a certain period, to acquire the load information of the server nodes. The dynamic algorithm based on this model such as Weighted Least Connection method (it's the default scheduling way of LVS system). Define N servers as S1,S2, ……Sn, of which the processing capacity are C1, C2,……Cn, the number of requests of each server are R1, R2,……Rn at present, the ratio between the server capability and the number of requests are Wm=Cm/Rm, then allocate the current task to the server, whose weights Wm is the largest. This algorithm allocates the task according to the current requests and server capability and the condition of the load balancing of the server. But this kind of algorithm has some weak point as following:

(1) The number of requests can't reflect the load condition of server node, because every request has different requirement for the utilization of the CPU or memory, so just the number of the requests can't reflect the real load of the server node.

(2) There is some inaccuracy about the number of requests between load balancer's record and the real number, and the inaccuracy will increase with the accumulation of time. If the difference of each request is great, such differences will be more obvious, lead to the apparent uneven distribution between the server nodes.

(3) Because the system is controlled centralized by the load balancer fully, the load of the load balancer also becomes hard with the number of request increasing, the load balancer will be of the bottleneck of the service finally.

Through the above analysis, the load balancing model and scheduling algorithm should achieve the following objectives: ① consider all server nodes and the handling capacity of the current load fully. ② transfer the work of information collection and calculation of the server nodes to themselves to avoid the load balancer itself becoming the bottleneck. ③ try to provide the service nearby, which avoid network bandwidth bottleneck.

The core of initiative feedback model is to change the state that the load information are collected by the central node, which replace it with collected by the server nodes themselves and report their load

information according to the status of themselves. So the central server node dispatch the task only according to the load information that has been sent by the server nodes, no need to collect load information by itself, which reduce the additional communication costs due to the load that caused by gathering load information, and so reduce the load of the central server node. Each server node shouldn't inquiry the information and calculate only when the requests reach, they should calculate according to their capability and load status, and report the calculated value to load balancer in a certain period. Load balancer gives the availability weights W of each node according to the certain algorithm, and dispatch the task to the node whose weights is the largest. Since each load status information feedback will have a time-T, if assign the request that reach in a time-T to a node simply, which will lead to overloading of this node. So the load balancer should evaluate the real load of the server nodes according to a certain algorithm, it should consider the performance parameters of each nodes, including static parameters and dynamic parameters, and the forecast of dynamic load increment of each server nodes. Based on these parameters the comprehensive analysis and calculation, we give a reasonable, simple algorithm, which determine the weights of each server node and achieve the purpose of load balancing.

3 The Model of User Behavior

Video scheduling strategy is actually a services strategy of user behavior, so the key of the research about the mathematic model of the scheduling algorithm is to create the model of user behavior, and study the rule of the reaching of user's request, the way users choose the program, and the way users wait for the program[3][4].

Under the objective conditions, the accesses of users are not average, but adhere to a certain probability of distribution, it following the Zipf distribution

$$p_k = \frac{k^{-\theta}}{c}, c = \sum_{k=1}^{n} i^{-\theta}, \quad 0.271 \le \theta \le 1 \qquad (1)$$

The above formula express the probability of access to the ith movie in n movies, the θ is a constant, call it depth factor, if θ is great, the inclination of the popular program of the users is high. To the different user group and program set, the value of θ is different, in order to enable universal certification, we will set the range of Zipf depth factor as [0.273, 1].

Under different circumstances, the choice of parameter is different when it follows the Zipf distribution, and we can derive a mathematical expression of parameter θ through the way of parameter estimation, the parameters of the system are of point estimation, gather the information in a certain period, and update as time flies. Define the existing data streaming media as x1, x2, …, xn from the most popular to the least , so the probability of visit is {p1, p2, …, pn}, secondly , we get a certain θ of the model of data distribution through the calculation of parameter moments estimates. Derivation process as following:

First, get the Mathematical expectations of real access probability of n movies $E_1 = \sum_{k=1}^{n} k \times p_k$, by the formula (1) can be drawn $k = (p_k \times c)^{-1/\theta}$,

Define Mathematical expectations E_2, which is estimated by the real P_k and parameter θ :

$$E_2 = \sum_{k=1}^{n} p_k \times (p_k \times c)^{-1/\theta}$$

The objective now is the appropriate choice of parameters θ , which make E_2 and E_1 close to full, it can be achieved by the way of binary search, which make the deviation of the parameter estimates in the controllable range, through the contrast between estimate of the probability of mathematical expectation and the original of the probability of mathematical expectation, then we can get the value of parameter θ . We found that the web site is entirely consistent with the distribution of the Zipf of θ . See as Figure 2-1, so Zipf distribution has a high credibility when it used to describe the model of visit.

The Zipf distribution has amazing properties that its hot spots are much focused, most of the movies

have little visitors, and the videos that most people visit are hot. The probability of visit of former m of the hot films is:

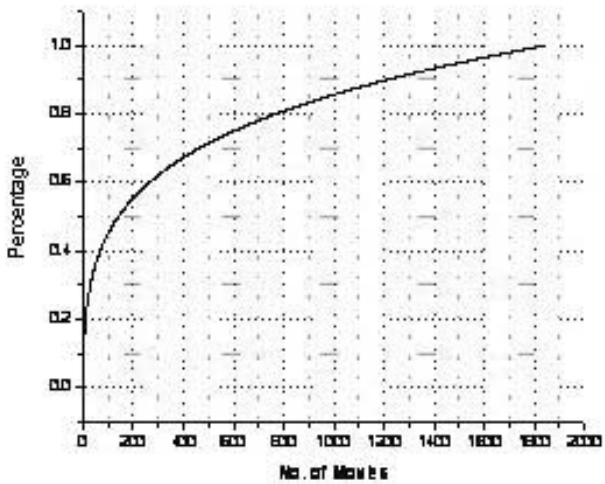


Figure 2 the probability of visit of the former N movies

$$P(M)=\sum_{i=1}^{m}p(i)=\sum_{i=1}^{m}\frac{i^{-\theta}}{\sum_{j=1}^{n}j^{-\theta}}=\frac{\sum_{i=1}^{m}i^{-\theta}}{\sum_{j=1}^{n}j^{-\theta}}$$

Its effect shown in Figure 2-1, when θ=0.8, in 1854 of the films, the number of visit of one of the most popular films is 5.33% of the total amount, the second is 3.18%, the former 20 is 26.06%, and the former 100 is 45%, the former 200 is 55.31, the former 500 is 71.34%. In other words, the 27% of the all movies has 71.34% of the total visit, and 73% of the left only has 28.66%. This characteristic means that the visit of local film is very good; it is suitable to use the model of nearby service to achieve higher efficiency in Video Grid system.

4 The Load Balancing Algorithm Based on The Initiative Feedback and Nearby Service

Through the analysis of the model of initiative feedback load balancing and the model of the film selection of users, we propose the load balancing algorithm that based on the initiative feedback and nearby service [5].

4.1 The calculation of the capability of the server node

Define the process capability of node S_i as C (S_i), consider it mainly in these indicators, the number of CPU ni, the type of CPU, memory capacity C (Mi), the I/O rate of disk is C (Di), the network throughput is C (Pi). The process probability of node can be expressed in the function follow as:

C′(Si)=k1×niC(Ci)+k2×C(Di)+k3×C(Mi)+k4×C(Ni)+k5×C(Pi)，I=0,1,…，n-1，∑k=1 (3-1)

Because the various indicators of nodes are different, and the process capabilities of the various services are also different, so we introduce the parameter k to impress the dependence of the various indicators of a certain kind of service [6].

4.2 The load of server nodes by initiative feedback

Define the occupancy rate of CPU as L (Ci), the occupancy rate of disk I/O as L (Mi), the occupancy rate of network bandwidth as L (Ni), the occupancy rate of the number of process as L (Pi). So the load of node L’ (Si) can be expressed in following function:

L′(Si)=r1 × L(Ci)+r2 × L(Di)+r3 × L(Mi)+r4 × L(Ni)+r5×L(Pi), i=0,1,…,n-1, ∑k=1 (2)

Because of the difference of the type of service, there is some difference between the loads of each part of nodes. So we introduce parameter r, to impress the different degree of impact of each part of this service.

4.3 The incremental load of server nodes

The incremental load refers to task Ri will increase INC (Ri) load on the node which accept the work. the calculation of the incremental load of node can be expressed in following function: INC(Ri)=L(Si)/n,

L (Si) is the load feedbacked by the node positivly, n as the number of connections of the node.

Then we can get the correction method of the load of node: define that there is a load feedback of node in t1 moment, the server node Si accepts a request at t2 moment (t1<t2<t1+T, T is the initiative feedback cycle

of node), L (Si) is the load of Si before t2 moment. For heterogeneous nodes, because of the difference of process capability, the incremental load also has some difference, so we should adjust in accordance with the handling capacity of a node C (Si). C is the process capability of node of the calculation of incremental load, the current load L″(Si) following as:

$$L''(Si) = L(Si) + C/C(Si) * INC(R) \quad (3)$$

In the same, if Si finishes a request of task in t3 moment (t1<t3<t1+T), the current load L″ (Si) following as:

$$L''(Si) = L(Si) - C/C(Si) * INC(R) \quad (4)$$

L″ (Si) is the load after node Si accepts the task, in the next time of calculation of load balancing; the load of this node should use L″ (Si).

According to the above analysis, we can get the way of calculation of load of the server node L (Si): define that do the initiative feedback of load of nodes in t1 moment, so calculate the load L′(Si) to correct the current load L (Si) of Si by the information that node feedback:

$$L(Si) = L'(Si) \quad (5)$$

If the server node Si accepts a request in t2 moment (t1<t2<t1+T), use the load L″ (Si) after accepting request to correct the current load of Si L (Si):

$$L(Si) = L''(Si) \quad (6)$$

4.4 Load balancing algorithm

(1) Check that whether there is a need for operations copies of movies in the area of task of server node, if not go to (3), and otherwise continue.

(2) Make sure that the load of the server node is not overload in the area of the task; if yes return this node, and otherwise continue.

(3) Find the other server node.

(4) Check that whether there is a sever node for need, if not go to (8), and otherwise continue.

(5) Check that whether the network bandwidth of this server node is overload; if yes go to (8), and otherwise continue.

(6) Check that whether the load of this server node is overload; if yes go to (8), and otherwise continue.

(7) If the load of this server node LOAD is lower than the least load LOWESTLOAD defined, then LOWESTLOAD=LOAD, clear the linklist LOWESTLOADQ, and add this node to the LOWESTLOADQ. If LOWESTLOAD equal LOAD, add this node to the LOWESTLOADQ.

(8) If there is another available server node, return to (3), and otherwise continue.

(9) If the least linklist of load LOWESTACQ is not empty, choose a node from LOWESTACQ in randomly to return.

(10) Check that whether the network bandwidth from user to the center is overload, if not go to (12), and otherwise continue.

(11) If the loads of central sever node is not overload, return the central node, and otherwise continue.

(12) Return null.

5 Conclusions

We discuss a load balancing algorithm that based on the initiative feedback and nearby service, take the Video Grid System as the background, through the analysis of the mathematical model of user behavior; we verify that the nearby service load balancing algorithm can be more efficient for the user to provide video services. The strategy of initiative feedback is to change the state that the load information are collected by the central node, which replace it with collected by the server nodes themselves and report their load information according to the status of themselves, it also reduces the additional communication costs that due to the load that caused by gathering load information and the load of the central server node. This algorithm is suitable to solve the load balancing problem of Video Grid System.

References

[1] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal

Processing, 2000, 10(1): 19-41

[2] Liu Peng, Yu Zhihui Chen, Yu .Grid computing. Available online at: http://www.chinagrid.net, 2003

[3] Dong, Yingfei. Efficient resource management in multimedia streaming networks [J]. Dissertation Abstracts International, 2003, Volume: 64-06

[4] Zamora, Javier. Video-on-demand systems and broadband networks: Quality of service issues [J]. Dissertation Abstracts International, 1998, Volume: 59-11

[5] Liu Jian Tian Shaoliang. Based on the dynamic feedback load balancing algorithm. Computer Engineering and Design. Vol.28. No.3. 2007

[6] n. A load balancing algorithm based on dynamic feedback. Computer Engineering & Science. Vol.25. No.5 .2003

[7] Varadarajan, Srivatsan. End-to-end Quality of Service (QoS) management for continuous media delivery over distributed network systems [J]. Dissertation Abstracts International, 2004, Volume: 65-10

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

[9] Sonesh Surnana, Brighten Godfrey. Loading balancing in dynamic structured peer-to-pear system. Available online at www.sicencedirect.com. Performance Evaluation 63(2006) 217-240

[10] Zhao, Yinqing. Design and analysis of server scheduling for video-on-demand systems [J]. Dissertation Abstracts International, Volume: 66-11, Section: B, page: 6190

[11] Rumade, Shraddha. Distributed streaming for video on demand [J]. Masters Abstracts International, 2005, Volume: 44-01, page: 0406

[12] Rotithor H G. Taxonomy of dynamic task scheduling schemes in distributed computing system[J].IEEE Proceedings-Computers and Digital Techniques,1994,141(1): 55~59

Yang Fan, male, born in 1983. He is a master degree candidate of School of Computer Science and Technology, Wuhan University of Technology. His research interests are Parallel distributed and high-performance computing.

Guo Qingping is a Full Professor and a head of Parallel Processing Lab, dean of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology. He graduated from Wuhan University in 1968; from Huazhong University of Science and Technology in 1981 with specialty of wireless technology. He is a holder of K. C. Wong Award of UK Royal Society (1994); was a visiting scholar of City University and University of West Minster (1986~1988), Visiting Professor of the UK Royal Society (1994), Visiting Professor of Queen Mary and Westfield College, London University (1997~2000), Visiting Professor of National University of Singapore (2000), Visiting Professor of University Greenwich (2003). He is one of the DCABES international conference founder, was the chairman of DCABES 2001, co-chair of DCABES 2002, and will be the chairman of DCABES 2004. He has published two books, over 80 Journal papers, edited two DCABES Proceedings. His research interests are in distributed parallel processing, grid computing, network security and e-commence.

Research on the Disambiguation with Ontology in MT

Wei Tang Qingping Guo

School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China
Email: wildducktoby@163.com

Abstract

Disambiguation is one of the hardest questions in MT [2,3].Research indicates that the difficulty comes from the machine's lack of capability for comprehending [6,7] semantics. Ontology is a new technique for expressing knowledge which comes from AI. Ontology has great ability for expressing semantics, and also can infer semantically with itself. Thus it has a great sense in disambiguation. Naturally we want to apply ontology to MT, make full use of its ability to express and process semantics. This passage discusses what effect ontology can make on disambiguation. It also tries to find out the reasons why there wound be ambiguity in MT, and classify different kinds of ambiguity, then correspondingly discuss the methods for disambiguation by using ontology to get the right translation. After that ,we discussed the performance of the MT system based on ontology, and how the completeness and the structure of ontology affect it. This paper tries to figure out the prospect of ontology's use in MT for disambiguation.

Keywords: Ontology, MT, Disambiguation

1 Introductions

1.1 Introduction to MT

MT(machine translation)is a subject of translating a kind of language to another language using computer [3]. This new subject is also a new kind of technique. It is very typical subject which refers to many fields such as glossology, science of computer, mathematics. To let computers understand human's language is a very difficult but charming subject for research. Because of the great complexity of natural language, MT is very complex and hard, it is one of the ten most difficult problems.

The core question of translation is comprehending [4,6]. Although the conventional NLP based on formal grammar claimed that it can "understand" natural language. But because it doesn't refer to the semantics, it can not reach true comprehending. Comprehending must be built upon certain knowledge.1966,Language Automatic Processing Consultation Committee of US National Academy of Sciences publicized a report named "Language and Machine", which pointed out that MT has a "semantic barrier" which is hard to overcome [2]. The result of this is that research on MT fell to the bottom all over the world. In the 1980s the rise of processing based on Statistical method brings new way for the development of MT. But the problem of expression of semantics is still not well solved yet. In recent years, the technique of ontology arising from the field of AI provides a new prospect for solving this problem.

1.2 Introduction to ontology

Ontology is defined as "a formal, explicit specification of a shared conceptualization". In 1990s it appears with the development of knowledge base technique. It is the entity of knowledge that is composed of relations between concepts, is the conceptual description of world knowledge or domain knowledge[3].

1.3 Introduction to MT based on ontology

The main task of MT based on ontology is using the knowledge owned by ontology to help parse and

build the language, extract the meaning of the text, translate it into the inner form of the ontology, which could be translated into forms of different natural languages. Simply speaking, this process resembles manual translation very much, which has a process of comprehending and translation.

Ontology not only has the ability to conceptually describe domain or world knowledge [3], but also includes the knowledge for inference. Thus they could be used for disambiguation and inference of natural language processing [6].

2 Analyses for Ambiguity in MT

Processing for ambiguity is the difficult part of MT. The essence of it is the contradiction of language's form and meaning, in other words, one form associated with different meanings. Massive existences of ambiguity is a important feature of natural language which differentiated it from formal language. Research on ambiguity is usually the point from which theory of MT grows.

The disambiguation we talked here, means clearing up the ambiguity which may appear in MT, not the ambiguity that is hard to understand in man's reading. That's because if some ambiguity can't be understood in man's reading, that means even with a completely understanding of current semantic, we still can't clear up it. Obviously, machines should not try to clear up this kind of irresolvable ambiguity.

Here we will discuss the source of several kinds of main ambiguity. Then we try to discuss how them could be cleared up with the help of ontology's capability to express semantics.

2.1 Lexical ambiguity

Lexical ambiguity is "same word, different meanings", that means a word could attach to more than one kind of meaning. e.g. "seal" can mean "sea dog" or "print".

2.2 Structural ambiguity

Structural ambiguity takes effect on the structure formed by words, belongs to the syntax layer. A classical example in NLP is "I saw a girl with a telescope". "With a telescope "can be judged to be the attribute of "a girl" or the Adverb of "saw". This could not be decided at the syntax layer, thus a structural ambiguity.

There are three main kinds of structural ambiguity in English:

1. In the structure "VP+NP1+Prep+NP2".the phrase "Prep+NP2"could be both the attribute of NP1 and the adverb of VP, which makes the ambiguity, such as the example "I saw a girl with a telescope".

2. When there is an "and" in the phase, for "and" could has different control areas, the structural ambiguity appears. e.g. the structure of "young boys and girls" can be "(young boys) and girls" or "young (boys and girls)".

3. The meaning of a phrase formed by several nouns could be interpreted differently, thus the ambiguity forms. e.g. "stone hammer" could be interpret to "a hammer made of stone" or "a hammer for processing stone".

2.3 Pronoun ambiguity

When pronouns appears in the phases, machine often can't decide which thing in the context this pronoun want to point to. This phenomenon is just very common.

3 Disambiguation Based on Ontology

With the powerful capability to express semantics and infer through semantics, we could make use of it to clear up ambiguity in MT at several levels.

3.1 Disambiguation by knowledge of ontology

Here we will discuss how to clear up different kinds of ambiguity with the knowledge of ontology.

3.1.1 Clearing up Lexical ambiguity by knowledge of ontology

When we hear a word, we could obtain a series of information of the concept this word expresses. Such information could be called “background information”. e.g. When we hear the word “bike”, much information about “bike” appears in our brain: it is a traffic tool ,has two wheels, mainly made of steel, the height and the width about… This back ground information is usually the key to judge whether some expression is reasonable. But when a MT system which lack such background knowledge hears the word ,it sees it as only a meaningless symbol. So we could first build such inherent information into ontology, then using it to help comprehending and inference, to clear up some kind of lexical ambiguity.

The process of disambiguation is shown as follow. First we associate the ontology with the vocabulary. In this process we need to consider the relationships between concepts and vocabulary. A word that has several meanings would be mapped semantically to different concepts. So every concept has a list of words formed by thesauruses. e.g. The ontology of WordNet takes such mechanism. When analyzing the text, the system could infer and clear up the ambiguity.

A famous example given by American Symbolic logician Bar-Hillel in 1959:” John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.” He said, to discover the right translation of the polysemous word “pen” is a very hard thing. Here ”pen” must be translated into “playpen”, can’t be translated into “the writing tool fountain pen” .To get such right translation depends on the common knowledge of the world, but we can’t add such knowledge into computers.

While now we have the ontology that could express such common knowledge of the world, so we get the chance to clear up such ambiguity. First we find multi-concepts of the word “pen” through the semantic mapping of vocabulary and ontology, two of which are “fountain pen” and “playpen”. Then we add the two concept into the concept network of the sentence where it comes from, and infer to check the semantic consistency of it. The Second-order concept “be in" should has this semantic restraint :”if the two variables of this concept are material object, then volume of the first variable should not be bigger than the cubage of the second variable.” By this semantic restraint we find in the ontology the “volume” property of “toy box” and the “cubage” property of “fountain pen”, then we find that the two concepts don’t obey this restraint. Then we find the concept “playpen” could pass. So we can determine in this example the word “pen” should be translated into the concept “playpen”.

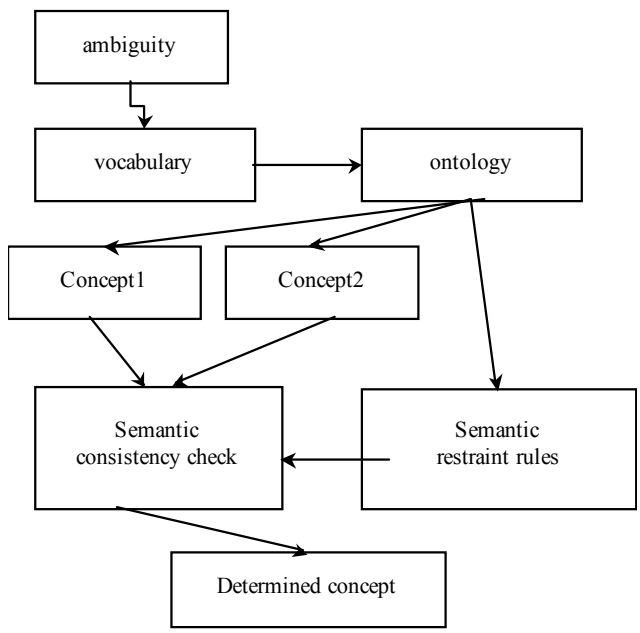


Figure 1 process of Clearing up Lexical ambiguity by knowledge of ontology

3.1.2 Clearing up some structural ambiguity by knowledge of ontology

Because the semantics could help to decide the structure of sentences, some certain structural ambiguity could be cleared up with the help of ontology’s knowledge. This process takes place on the phase of syntax parsing. Here we have two sentences: “ I saw a boy with a telescope” and “I saw a boy with a hat”. The two sentences have a structural ambiguity in MT. “I saw a boy with a telescope” can be translated into “I saw a boy using a telescope.”, or “I saw a boy who was taking a

telescope." In this sentence the ambiguity is impossible to be cleared up even by manual translation, thus we should let computer try to do it. But the structural ambiguity in the one "I saw a boy with a hat" can be cleared up by the knowledge of the ontology. Its structure can be decided, for "hat" can't be used to "see the boy". While meeting this kind of structural ambiguity, we can separately build the different forms of the concept attached to different syntax structures. Then in each case we check the semantic consistency, thus the right structure can be determined. The process of disambiguation for some other kinds of structural ambiguity is similar.

3.2 Disambiguation by context

A great problem in MT is that computers are hard to process the relations in the context [7]. The context is the environment where the language structural element such as word, phrase, sentence lies in.

In articles especially narration such as novels, things develops in order of time, states of objects described varies continuously. If we can grasp the information about the development of the objects in context, we could make a judgment for the state of the objects, thus to clear up some ambiguity. To achieve the goal, a feasible method is adding some additional flags of state for each individual that appears in the context. If some event changes some individual's state, the flag of that state would be altered correspondingly. By this way we can record the information of conditions of the individuals at that time. Then we can use this information to infer semantically, to achieve the goal of disambiguation. E.g. If a man "dies" in some place of the context, we set the flag of living condition to "dead". In the following part of the context that individual should not take any action which must be done by alive ones.

Disambiguation by context can clear up many kinds of ambiguity, including some lexical ambiguity that could not be cleared up only using the knowledge of the ontology, pronoun ambiguity ,some kind of structural ambiguity etc.

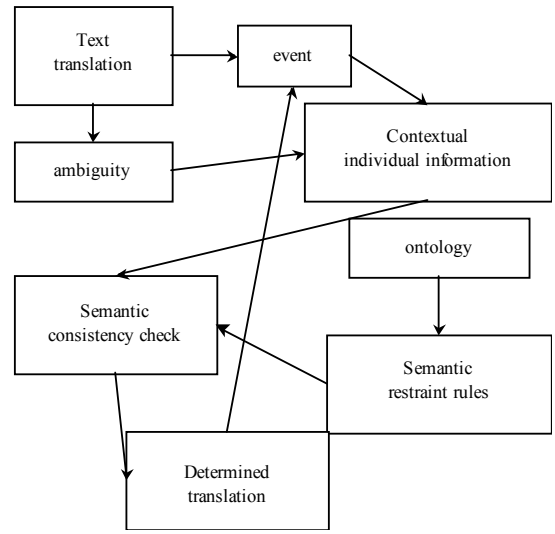


Figure 2 the mechanism of Disambiguation by context

4 The Influence of Design of Ontology to Performance of MT System

The natural language is very rich and ambiguous. Although ontology has a powerful ability to express semantics, it can only express semantics that is specific and formal, which is also the feature of computer. So it is very hard for machine to fully comprehend natural language. We could see that, a reasonable design of the ontology is very important for the machine's ability to understand natural language.

4.1 The completeness of ontology

Because the concepts and relations between concepts are very complex, the designed ontology should has a complete coverage, with subtle enough concepts and complete semantic restraints. So the MT system based on this ontology can has a more complete semantics-understanding ability to process the rich semantics of natural language.

4.2 The structure of ontology

Concepts in natural language have rich inherent structure, with plenty of similar concepts and relations of inheritance. We should take use of these characteristics to organize the concepts in ontology, let

the structure of the ontology as compact as possible, with less redundancy, thus accord with human's thinking custom. A system with a ontology of a fine structure would work more efficiently.

5 Conclusions

The technique of ontology has been used and paid great attention in many fields, and has got many fruits. Yet using it in MT is a great challenge, for the covered area of MT is too wide, including the knowledge of common sense. Now the main difficulties are in two aspects. First, to construct a ontology for common use upon which the MT system works is a very giant task, including countless concepts which should be defined efficiently, countless semantic restraints for deduction which should well express common sense and other knowledge. Second, MT system should try its best to dig as much as possible the semantics from the text. By dong this the ontology's knowledge can be fully used.

The using of ontology in MT has a very promising prospect. If the two problems are well solved, MT would exhibit a great intelligence. The ability of translation can be expected to approach the level of manual translation.

References

[1] Shanping Li. Overview of researches on Ontology[J].journal of computer research and development. 2004. 7

[2] Chongde Shi. A Research on the Ontology-based Chinese-English Machine Translation[J]. Library and Information Service. Vol.50. No 9. 2006. 9

[3] Xiaojie Wang. Using Ontology in a English-Chinese Machine Translation System. Journal of Chinese Information Processing Vol.14. No.5

[4] Zhiwei Feng. On Humanity Spirit of Natural Language Processing from the Viewpoint of Ontology[J]. Applied Linguistics. 2005, 4

[5] Yanfeng Sun. The application of the Ontology technology on the understanding of natural language[J]. Journal of Qinghai Normal University(Natural Science). 2003, 3

[6] Yubing Pan. Ontology-Based Natural Language Understand. Computing Technology and Automation. 2003,12

[7] Jun Ye. On Multi-meaning in Machine Translation. Journal of Yangtze University(Social Sciences). Vol. 28 No. 13

[8] Uschold M, Gruninger M. Ontology: Principles, methods and applications[J].The Knowledge Engineering Review, 1996,11(2):932136

Wei Tang was born in 1983, He works in Parallel Processing Lab of School of Computer Science and Technology, Wuhan University of Technology. His main research interests are in the area of distributed parallel processing, grid computing, network security and e-commence.

Qingping Guo is a Full Professor and a head of Parallel Processing Lab, dean of Computer Technology Institute in School of Computer Science and Technology, Wuhan University of Technology. He graduated from Wuhan University in 1968; from Huazhong University of Science and Technology in 1981 with specialty of wireless technology. He is a holder of K. C. Wong Award of UK Royal Society (1994); was a visiting scholar of City University and University of West Minster (1986~1988), Visiting Professor of the UK Royal Society (1994), Visiting Professor of Queen Mary and Westfield College, London University (1997~2000), Visiting Professor of National University of Singapore (2000), Visiting Professor of University Greenwich (2003). He is one of the DCABES international conference founder, was the chairman of DCABES 2001, co-chair of DCABES 2002, and will be the chairman of DCABES 2004. He has published two books, over 80 Journal papers, edited two DCABES Proceedings. His research interests are in distributed parallel processing, grid computing, network security and e-commence.

Remote Control Simulation for the Loitering Attack Missile based on Data Link

Shengzhi Yuan Xiaofang Xie Jian Cao Xiaoming Bai

Department of Science and Technology of Weapons, Naval Aeronautical Engineering University
Yantai, Shandong, China
Email: yuanshengzhi_hy@sina.com

Abstract

In order to master the remote control technology of the Loitering Attack Missile (LAM), and the application of the data link into the cruise missile, it is very important to develop a remote control simulation for the LAM. The structure of the simulation system was designed on the platform of the High Level Architecture (HLA).The key questions such as the basic control models, the simulation of Data Link , the simulation of track following and scene matching were strictly introduced. In practice, the feasible method -the Creator, Vega and VC++, was applied to the development of the remote control simulation system for the LAM. The method is very useful, not only in practice but in the national defense, which can be used in other applications.

Keywords: Loitering Attack Missile (LAM), Remote Control, High Level Architecture (HLA), and Data Link

1 Introduction

Along with the development of the technology of Data Link and its applications to the Weapon Systems[1], the remote control technology for Cruise Missile has great improvement. The Loitering Attack Missile (LAM) is coming forth. The LAM is a missile, which can fly to the target area according to the control software applied into the missile-borne computer before launching, exchange the information with the launch platform, the remote command and control center, the GPS Satellite through the bidirectional data link, also accept the instructions from the remote command and control center, change to attack more valuable targets. There are many types of LAM which have been used: BLOCK IV, AGM-136 of USA, ARW-10 of S. Africa, and Star-1 of Israel and so on.

In order to master the remote control technology of the Loitering Attack Missile (LAM), and the application of the data link to the cruise missile, a remote control simulation system for the LAM was proposed. In order to build the architecture of the system, the High Level Architecture (HLA) would be used. The paper was organized in the following way: the structure of the remote control simulation system for the LAM based on HLA was presented in Section 2. In Section 3, the key points of system designing such as basic control models, the simulation of the data link and controlling was strictly discussed. In Section 4 the method- "Creator, Vega and VC++" was applied to the realization of the remote simulation system. The simulation scenes were also shown. In Section 5, the issues mentioned in this paper plans for further work were summarized and described.

2 Structure of the Remote Control Simulation System(RCSS) for LAM

2.1 Analysis of combat process of lam

After Launching, The LAM can pass through all navigation pots and directly attack the scheduled target or preparative target based on the flight and control software. It can also patrol above the area of scheduled target based on the patrolling pattern in order to find a time-urgent target which is more valuable. If a time-urgent target is found in the patrol course, the LAM will

send the radar image of battlefield to the remote command and control center through Data Link. The more valuable target will be identified in the remote command and control center from the battlefield information accepted. Then a series of control instructions, which will control the LAM to strike the time-urgent target, will be formed and sent by the center. The control process of LAM was shown as Figure 1.

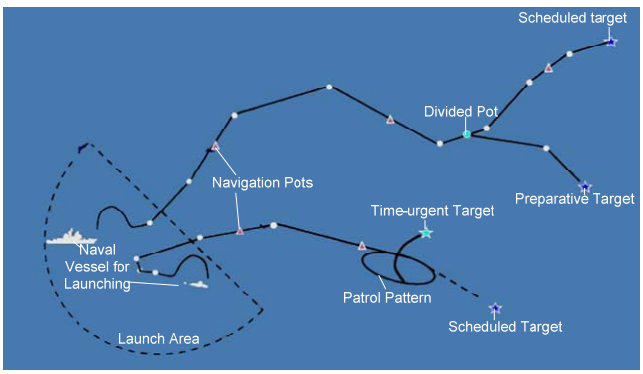


Figure 1 Combat process of LAM

During the process, there are four steps:

1)The step of launching: in the Launch Area, the information of the scheduled target, preparative target and the flight control software is input from the Naval Vessels before Launching;

2)The step of navigation: the LAM exchanges the information of navigation with the GPS through Data Link; with the navigation method of GPS, the LAM can flight according to Navigation Pots;

3)The step of patrolling: in order to find more valuable targets, the remote command and control center sends instructions through Data Link, makes the LAM patrol and find the Time-urgent Target according to the Patrol Pattern;

4)The step of attacking: After accepting the image which the LAM has gotten, the remote command and control center can identify the more valuable target, forms a series of control instructions and command the LAM to attack the time-urgent target.

2.2 Structure of the RCSS based on HLA

Based on the analysis of control process of LAM, it is found that the Remote Control of LAM depends on Data Link. With the equipments such as wireless broadcasting stations, the wireless control net is formed. Because of the long range, about thousands of miles, it is impossible in practice to form a remote control simulation for LAM by using the true equipments. Thus, it's necessary to found the simulation environment based on LAN. This method can not only save money and resources, but also realized easily.

According to the particularity of the Remote Control Simulation for LAM, which belongs to the category of the data link simulation, especially the diversity of targets, environment, the HLA was proposed to build its structure. Based on the Federation Development and Execute Process Model (FEDEP) of the HLA[4][5], we defined the remote simulation system as a Federation, which was composed of six federation members: the federation object of LAM, the naval vessels for launching, the battlefield environment, the GPS satellite, the simulation controller and the remote command and control center. The federation objects and run-time infrastructure form a distributed interactive simulation system-the Remote Control Simulation System (RCSS). The structure of RCSS was shown as Figure 2.

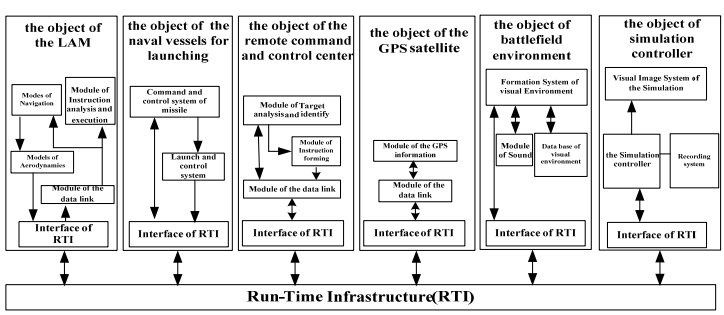


Figure 2 Structure chart of RCSS

In Figure 2, the federation object of the simulation controller is composed of four parts: Visual Image System of the Simulation, the Simulation controller, Recording system, and interface of RTI. The simulation controller takes charge of the simulation process, the initialization of the simulation, which is about the targets and the virtual image environment of the simulation. The Visual Image System of the simulation takes charges of the presentation of the virtual operation environment. It offers the scene of the simulation immediately through the projectors and the spar screen. Recording system can record all the scenes during the course of the remote control simulation, which can give us the chance to evaluate the effect of the simulation.

The federation object of operation environment can offer a real virtual environment. It is composed of four parts: Formation system of visual environment, Database of visual environment, Sound system and Interface of RTI. According to the targets and the process of the simulation which the simulation controller initialized, the different models of operation environment are chosen from Database of visual environment. It also provides us with the real sound environment through sound system.

The federation object of the LAM is the simulation object of RCSS. Based on the models of aerodynamics, it can offer a dynamic process. With the module of Data Link, it gets the navigation information from the federation object of the GPS satellite, also the instructions from the object of the remote command and control center. With the mode and information of Navigation, it can offer the track of the LAM during the step of navigation. During the steps of patrolling and striking, it can choose and strike the target according to the instructions from the remote center.

The federation object of the remote command and control center takes charge of target choosing and attacking for LAM. It is composed of four parts: Module of the data link, Module of Target analysis and identify, Module of Instruction forming and Interface of RTI.

The federation object of the GPS satellite takes charge of the Navigation information transferring. It is composed of three parts: Module of the GPS information, module of the data link, and interface of RTI. The federation object of the naval vessels for launching takes charge of launching and control of LMA in the step of Launching. It is composed of three parts: Command and control system, launch and control system, and interface of RTI.

2.3 Hardware platform of the RCSS

In the design process, the hardware as below was chosen: 1) 2 sets Sun Graph 6000 series VR workstation--Sun Graph 6000T (CPU: 4500MHz,.Hard Disk: 400G, 1000M Ethernet supported); 2) 4 sets slap-up computer--Dell Optiplex GX520 (CPU: 3000MHz, Main Memory: 1GB, System Bus: 533MHz, Hard Disk: 120G); 3) The solid projector--Sun Graph Dual 5500(4 sets), the 120° spar screen;4) The 100M Ethernet. The whole structure of hardware platform was shown as Figure 3.

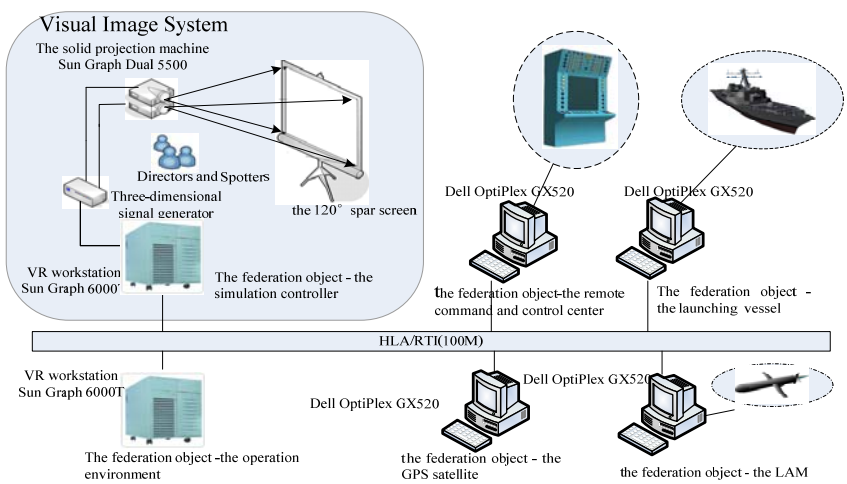


Figure 3 Structure chart of the hardware platform

3 Key Points and Answers About the RCSS for LAM

The RCSS is a typical man-in-loop control system. All parts: the LAM, the GPS, the remote command and control center, exchange information through the bidirectional data link. Thus, it is belong to the catalog of the Simulation of Data Link. Obviously, it is very significant to make research on the control method of LAM based on Data Link.

3.1 Basic control models of LAM

The research on the simulation of the aviation and control for LAM was developed on the models of aerodynamics and navigation. Comparing the LAM with the common cruise missile, the models of aerodynamics has not changed because the LAM comes up from the cruise missile. In order to research on the flight control for LAM, the models of aerodynamics were developed. In practice, 12 equations were chosen for engineering signification.

1) Particle motion kinematics equations of missile were shown as Eq.(3-1).

$$\begin{cases} \dot{x} = V\cos\theta\cos\psi_V \\ \dot{y} = V\sin\theta \\ \dot{z} = -V\cos\theta\sin\psi_V \end{cases} \tag{3-1}$$

2) Particle turn dynamics equations of missile were shown as Eq.(3-2).

$$\begin{cases} \dot{\omega}_x = \dfrac{M_x}{J_x} - \dfrac{J_z - J_y}{J_x}\omega_y\omega_z \\ \dot{\omega}_y = \dfrac{M_y}{J_y} - \dfrac{J_x - J_z}{J_y}\omega_x\omega_z \\ \dot{\omega}_z = \dfrac{M_z}{J_z} - \dfrac{J_y - J_z}{J_z}\omega_x\omega_y \end{cases} \tag{3-2}$$

3) Particle motion dynamics equations of missile were shown as Eq.(3-3).

$$\begin{cases} \dot{V} = \dfrac{XX}{m}\cos\alpha\cos\beta - \dfrac{YY}{m}\sin\alpha\cos\beta \\ \quad + \dfrac{ZZ}{m}\sin\beta - g\sin\theta \\ \dot{\theta} = \dfrac{XX}{mV}\left(\cos\alpha\sin\beta\sin\gamma_V + \sin\alpha\cos\gamma_V\right) \\ \quad - \dfrac{YY}{mV}\left(\sin\alpha\sin\beta\sin\gamma_V - \cos\alpha\cos\gamma_V\right) \\ \quad \dfrac{ZZ}{mV}\cos\beta\sin\gamma_V - \dfrac{g}{V}\cos\theta \\ \dot{\psi}_V = \dfrac{XX}{mV}\left(\cos\alpha\sin\beta\cos\gamma_V - \sin\alpha\sin\gamma_V\right) \\ \quad - \dfrac{YY}{mV}\left(\sin\alpha\sin\beta\cos\gamma_V + \cos\alpha\sin\gamma_V\right) \\ \quad - \dfrac{ZZ}{mV\cos\theta}\cos\beta\cos\gamma_V \end{cases} \tag{3-3}$$

4) Attitude corner dynamics equations of missile were shown as Eq.(3-4).

$$\begin{cases} \dot{\alpha} = -\dfrac{XX}{mV\cos\beta}\sin\alpha - \dfrac{YY}{mV\cos\beta}\cos\alpha \\ \quad + \dfrac{g}{V\cos\beta}\cos\theta\cos\gamma_V + \omega_z - \omega_x\tan\beta\cos\alpha + \\ \quad \omega_y\tan\beta\sin\alpha \\ \dot{\beta} = -\dfrac{XX}{mV}\cos\alpha\sin\beta + \dfrac{YY}{mV}\sin\beta\sin\alpha \\ \quad + \dfrac{ZZ}{mV}\cos\beta + \dfrac{g}{V}\cos\theta\sin\gamma_V \\ \quad + \omega_x\sin\alpha + \omega_y\cos\alpha \\ \dot{\gamma}_V = \dfrac{XX}{mV}(\sin\alpha\tan\beta - \cos\alpha\sin\beta\tan\theta\cos\gamma_V \\ \quad + \sin\alpha\tan\theta\sin\gamma_V) + \dfrac{YY}{mV}(\tan\beta\cos\alpha \\ \quad + \sin\alpha\sin\beta\tan\theta\cos\gamma_V + \cos\alpha\tan\theta\sin\gamma_V) \\ \quad + \dfrac{ZZ}{mV}\cos\beta\tan\theta\cos\gamma_V - \dfrac{g}{V}\cos\theta\cos\beta\cos\gamma_V \\ \quad + \omega_x\dfrac{\cos\alpha}{\cos\beta} - \omega_y\dfrac{\sin\alpha}{\cos\beta} \end{cases} \tag{3-4}$$

In the Eq.3-1, 3-2, 3-3, 3-4, x, y, z are the co-ordinates of the missile location; V is the speed of the missile. $\omega_x, \omega_y, \omega_z$ are the angular velocity; M_x, M_y, M_z are the torque; J_x, J_y, J_z are the moment of inertia; P is the thrust of the engine; X, Y, Z is aerodynamic drag, aerodynamic lift, pneumatic side force; Q is the kinetic pressure($\rho V^2/2$); ρ is the pressure. In addition, the rest required formulae were

shown as Eq.(3-5).

$$\begin{cases} XX = P - X\cos\alpha\cos\beta + Y\sin\alpha - Z\cos\alpha\sin\beta \\ YY = X\sin\alpha\cos\beta + Y\cos\alpha + Z\sin\alpha\sin\beta \\ ZZ = -X\sin\beta + Z\cos\beta \\ X = QSC_X, Y = QSC_Y, Z = QSC_Z; \\ M_x = QSbC_x, M_y = QSbC_y, M_z = QSLC_x; \end{cases} \quad (3\text{-}5)$$

Additional remarks: the impact of earth's rotation and curvature as well as the change of acceleration of gravity is not considered in the basic models of the LAM. The influence of wind and atmosphere is ignored.

3.2 Simulation of data link

In the operation process of LAM, Data Link forms a wireless control network. The control center for LAM is not changeless. In the step of launching, the control center is the naval vessels for launching; in the step of navigation, the control center is the LAM; in the step of patrolling and striking, the control center is the remote command and control center. In this net, how to distribute the channel is a very important problem which should be solved. The layer of MAC (Medium Access Control) is the key to solve the above problem. As you know, in the communication network which the channel is shared, the capability of the net such as the throughput, efficiency and delay, mostly lies on the MAC protocol. The FDMA (Frequency Division Multiple Access) protocol, the TDMA (Time Division Multiple Access) protocol and the CDMA (Code Division Multiple Access) protocol are primary protocols of the MAC sub layer. Because of the favorable capability in the real-time communication, the TDMA protocol is usually chosen [6]. The Link 16 of USA, which is realized by the Joint Tactical Information Distribute System (JTIDS/TADIL J/Link 16)[1][7], is a typical example to which the TDMA protocol is applied. Because of different tasks, different environment, different command and control members, the RCCS for LAM should solve the problem of the time slice synchronization of the TDMA protocol, the cute node of reference time, the mechanism of control center changing. In the mode of time slices distributing, in order to fulfill the operation environment and reality, dynamic distribution method was applied. In practice, the Link 16 of USA was chosen for the simulation of the data link in the battle simulation of the LAM.

In the project of the simulation software, a separate module of Data Link was designed for the application as function transferring. Based on the HLA/RTI, the function of Data Link module was realized by RTI interface function. The naval vessel for launching, the GPS satellite, the LAM and the remote command and control center are the federation members of the RCSS for LAM. On the platform of Windows, multi-thread structure was selected. The main thread, the thread of control instruction forming, the transfer thread were designed by VC++. The class of the program has three kinds: 1) the class of time synchronization which is used to solve the problem of federation synchronization; 2) the class of instruction coding which is used to solve the problem of R_S coding; 3) the class of the network commutation which is used to solve the problem of data stream sending and receiving after coding.

3.3 Simulation of track following and scene matching

Obviously, the track of cruise missiles is mostly at the same height and same speed. With the technology of terrain following, the cruise missile can fly above low altitude or super low altitude. In different terrain, the fight height is also different. In the offing or plain, the fight height is 5.20m; in the foothill, the fight height 50m; in the hilly country, the fight height 100m.During the process of TF/TA2, the missile can flight through the valley based on the terrain gurgitation. With the technology of terrain following, the cruise missile can fly above low altitude or super low altitude. In different terrain, the fight height is also different. In the offing or plain, the fight height is 5.20m; in the foothill, the fight height 50m; in the hilly country, the fight height 100m.During the process of TF/TA2, the missile can flight through the valley based on the terrain gurgitation.

Because of the long range, the viability of cruise missiles will face the intimidation of antiaircraft forces more and more. Thus it is necessary to do research on the technology of track programming. In the simulation of the trace programming, the problems about intimidation modeling and programming algorithm choosing should be solved necessarily. Commonly, there are three kind of models for the intimidation: the mountain model, the electrical potential model, the equivalent column model [6].There are many programming algorithms such as A* algorithm, dynamic programming algorithm, genetic algorithm, Dijkstra algorithm. For incertitude of the battlefield situation, flexibility and concealment of the intimidation, it is necessary to get the real time dynamic programming ability. Therefore, in practice, the species of intimidation were set by the simulation controller; the location of intimidation was produced stochastically; the dynamic programming algorithm was chosen.

In the steps of patrolling and striking, the image information of the battlefield, which the LAM gets through its camera equipment, will be sent to the remote command and control center. How to identify more valuable targets from the image information should be discussed strictly. The technology of scene matching, which lies on the comparison between the real time image and the reference image from the little different environment, is selected as the terminal guidance law [8]. With the compare, the location or property difference was found. As a rule, the image information memorized in the remote command and control center is called as the fiducially image. The image information got by LAM is the real time image information. In order to identify the target quickly and exactly, the searching strategy and the guide line choosing should be discussed carefully. A simulation test conclusion about the matching time and veracity was introduced in the reference paper eight. In the environment of the reference image 279×252, the real time image 50×50, the FFT algorithm, Hansdroff algorithm, gene algorithm about binary coding, gene algorithm about decimal coding, the layered algorithm about wave transform, and the layered ameliorated algorithm about dimensional domain was analyzed and compared. The layered ameliorated algorithm about dimensional domain was most effective than others. In fact there is a vertical angle between the real time image and the fiducially image. The algorithm above is not adapted. The matching algorithm about orbicular projection was provided in the reference paper eight. In practice the matching algorithm about orbicular projection was chosen and got prefect effect.

4 Realization of the Remote Control Simulation System for LAM

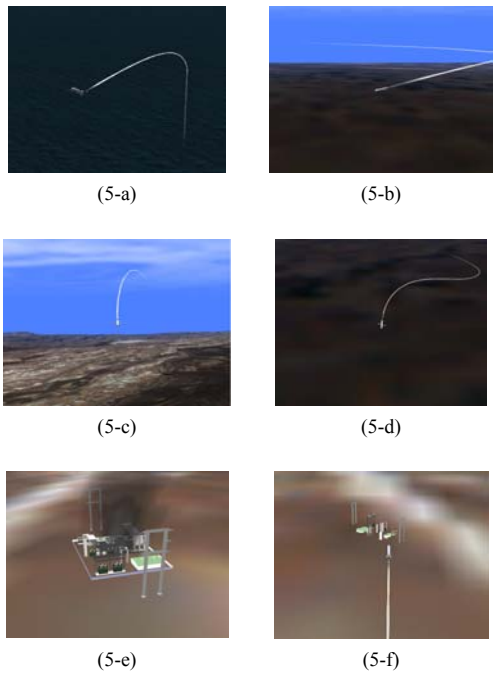
After the key questions solved, the RCSS for LAM was realized by the feasible method-the Creator, Vega and VC++. BLOCK IV and Link 16 of USA were chosen as typical simulation objects in the remote control simulation system of LAM. The software Creator was used to build simulation models [9]. The simulation model of LAM was shown as Figure 4-a, the simulation model of the transformer substation was shown as Figure 4-b.

(4-a) (4-b)

Figure 4 Simulation models developed by Creator

With the software Vega [10], the simulation models were import for further development. In order to fulfill the needs of the RCSS of LAM, different simulation scenes are designed. In order to reuse the simulation scenes developed by Vega, the support of Vega environment should be founded. Based on the software platform VC++, for the further application without Vega environment, we should make the Vega library actively. The typical scenes were shown as below. In Figure 5, the launching scene was shown as Figure 5-a; the middle guidance flight scene was shown as Figure 5-b; the track following scene was shown as

Figure 5-c; the dive-attack scene was shown as Figure 5-d; the scene of approaching the transformer substation was shown as Figure 5-e; the scene of hitting the transformer substation was shown as Figure 5-f.



(5-a) (5-b) (5-c) (5-d) (5-e) (5-f)

Figure 5 Typical Simulation scenes developed by Vega

5 Summary and Future Work

With the improvement of remote control technology, the LAM has been equipped in some country. The LAM will be improved quickly because of the control ability and use easily. In order to mater the remote control technology of LAM, and the application of the data link to the cruise missile, it is very important to make research on the remote control simulation for LAM. It has great meanings, not only in practice but also in the national defense.

In the paper, the structure of the battle Simulation System was designed on the platform of the HLA. In the process of developing the RCSS, the Multi-Gen Creator was used to build the 3D simulation models, Vega used to develop the Visual simulation scenes, and VC++ to integrate the whole system. In the aspect of theory researching, how to improve the effect of the remote command and control center, how to reduce the delay of the simulation system, how to apply HLA to the largest scale simulation faultlessly is further work for us.

References

[1] Chen Ying and Cheng Xiao-lane, “Application and Key Technology of Data Link in Armament of Foreign Army”, Tontics Technology, 2004.11, pp.43-50

[2] DMSO ,IEEE P1516.1 Standard for M&S High Level Architecture, Federate Simulation Interoperability Workshop, April 20, 1998

[3] Kevin Cox, A Framework-based Approach to HLA Federate Development, Simulation Interoperability Workshop, 98F-SIW-036, March 1998

[4] Zhou Yan and Dai Jianwei, Design of Simulation Program based on HLA, Publishing House of Electronics Industry. Beijing, china, 2002

[5] Huang Liuxin, “Distributed Interactive Scene Simulation of Cruise Missile based on HLA”, The master paper of Huazhong University of Science and Technology, Wuhan, china, 2004.5, pp .12-40

[6] Zhang hui, “Simulation on Data Link for Experiment in UAV Mission Control”, The master paper National University of Defense Technology, Changsha, Hunan, china. 2005.11, pp.16-44

[7] JTIDS (Joint Tactical Information Distribution System)-Link 16[J/OL], 2003.11

[8] Zou lin, “Research on Cruise Missile's Precise Air-to-Terra Attacking”, The master paper of Nanjing University of Aeronautics and Astronautics, Beijing, china, 2006.2, pp.20-47

[9] MultiGen Paradigm Inc, MultiGen Paradigm Inc.Creator3.0 tutors Guide, 2001

[10] MultiGen Paradigm Inc, MultiGen Paradigm Inc.Vega3.70 programmer’s Guide and Lynx help, 2001

Model Parameter Identification of a Coupled Industrial Tank System Based on A Wavelet Neural Network

Allam Maalla[1] Chen Wei[1] Mohammed H. Hafiz[2]

1 School of Information Engineering, Wuhan University of Technology, Wuhan, Hubei, 430070, China
Email: Allam.ahmed76@yahoo.com,

2 Department of Production and Metallurgy Engineering, University of Technology Baghdad, Iraq
Email:drmhh1962@gmail.com

Abstract

Liquid tank systems play important role in industrial application such as in food processing, beverage, dairy, filtration, effluent treatment, pharmaceutical industry, water purification system, industrial chemical processing and spray coating. In this paper describes the modeling process and parameters identification of the coupled industrial tank system based on WNN algorithm. Second transfer function for the LTS is suggested at first. In the investigated coupled industrial tank system, measurement was performed to obtain step response used for the identification of the parameters. The simulation results show good accuracy of proposed method. The results show that the obtained model and its parameters are good enough to represent the coupled industrial tank system

Keywords: WNN, coupled industrial, parameter, identification , tank system

1 Introduction

Liquid tank systems play important role in industrial application such as in food processing, beverage, dairy, filtration, effluent treatment, pharmaceutical industry, water purification system, industrial chemical processing and spray coating. A typical situation is one that requires fluid to be supplied to a chemical reactor at a constant rate. An upper tank can be used for filtering the variations in the upstream supply flow.

In order to achieve high performance, feedback control system is adopted. Classical PID controller based is widely used in controlling industrial liquid level control application [1,2]. Advanced control methods also have been proposed by several researchers such as sliding mode control [3] and nonlinear back stepping control [4]. Accurate model and its parameters which capture the characteristic of the coupled tank system is required for designing its controller for the achieving a good performance.

This paper presents WNN method to identify coupled industrial tank system based on the field test results. Wavelet transform is used as a preprocessor to optimizing the input features of the neural network [5]. Wavelet functions are embedded in the neural network to be stimulating functions of neurons or weight functions between two layers of neurons [6,7]. The validation result shows that the obtained model is good enough to represent the behaviour real industrial tank system.

2 System Description and the Test Procedure

This system consists of two tanks with orifices and level sensors, a DC motor driven pump and a liquid basin. The two tanks have same diameters and can be fitted with different diameter outflow orifices. The input is supplied by a DC motor driven variable speed pump which supplies fluid to Tank 1. The orifice, which is connected between Tank 1 and Tank 2, allows the fluid

to flow into Tank 2. The other orifice at Tank 2 flows back the fluid to the liquid basin. Two potentiometers are used to measure level of the fluid in the both Tank 1 and Tank 2.

System description

The schematic diagram of the lab-scale coupled tank system is shown in Figure 1. Each tank has and outlet port with q_{12} and q_{out} flow rates. The first tank is fed with q_{in} flow rate supplied by DC motor driven pump. The system is configured as a SISO system in which the input and outputs are q1 and h_2 respectively.

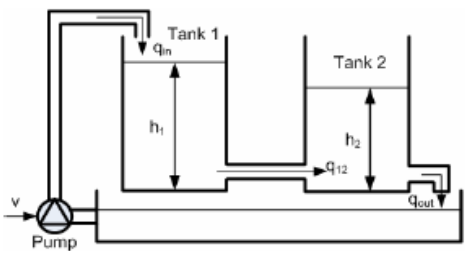


Figure 1 Schematic diagram of lab-scale couple industrial tanks

The tank model of Figure 2 can be represented as the following:

$$\frac{\hat{H}_2(s)}{Q_m(s)} = \frac{R_2}{R_1R_2A_1A_2s^2 + (R_1A_1 + R_1A_1 + R_2A_2)s + 1} \quad (1)$$

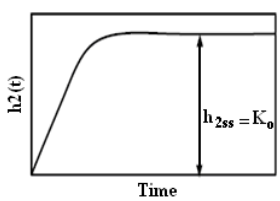


Figure 2 Step response of unknown system G(s)

Where R_1 and R_2 is the resistance of the tank 1 and tank 2 orifices respectively. In addition, A_1 and A_2 are the base surface area of each tank. Finally, by assuming the input flow rate is proportional to the input voltage of the pump, the transfer function of the industrial coupled tank system is expressed as follows:

$$\frac{H_2(s)}{V(s)} = \frac{CR_2}{R_1R_2A_1A_2s^2 + (R_1A_1 + R_1A_1 + R_2A_2)s + 1} \quad (2)$$

Where V(s) and C is input voltage to the pump and pump constant respectively.

In this paper, the parameters are identified based on the experimental step response introduced by Dorsey [5]. For the purpose of parameters identification, transfer function of Eq. (2) is expressed as follows:

$$G(s) = \frac{H_2(s)}{V(s)} = \frac{K_0}{a_2s^2 + a_1s + 1} \quad (3)$$

and is shown in Figure 1. The parameters K_0, a_2, and a_1 are

$$\begin{aligned} K_0 &= CR_2 \\ a_1 &= R_2A_1 + R_1A_1 + R_2A_2 \\ a_2 &= R_2A_1R_2A_2 \end{aligned} \quad (4)$$

Test procedure

In the defined test procedure, $H_2(kT)$ and $V(kT)$ are the samples of the system input and output with constant sampling period T. For parameters identification, a 5 volt step input is applied to drive the industrial tank system. The measured step response of tank system is shown in the Figure.2.

Another Tests were performed by another two different step signals (4 and 6 volt step input), which are used to verify the effectiveness of the system parameter identification.

3 Identification Method

Wavelet neural network (WNN)

In this paper the Wavelet neural network (WNN) is studied from the point of view of explicit parameter estimation in Coupled Industrial Tank system.

Wavelet transform can provide more precise decomposition in whole frequency band of an original signal. It is applied generally to analyze the complicated signals in many fields, such as pattern recognition, image processing and fault detection because of good capability of feature extraction and localization in time and frequency domain. The main advantage of wavelets is that they have a varying window size, being wide for low frequencies and narrow for the high ones, thus leading to an optimal time-frequency resolution in all the frequency ranges.

A family of wavelet is derived from the translations and dilations of a single function. If $\Psi(t)$ is the original function, referred to as the mother wavelet, the members

of the family are given by

$$\psi_{a,b}(t) = |a|^{\frac{1}{2}} \psi(\frac{t-b}{a}) \qquad a,b \in R, \quad a \neq 0 \tag{4}$$

They are indexed by two labels a and *b*, where a indicates the dilation and *b* the translation of the mother wavelet $\Psi(t)$. Additionally, *a* scaling function $\varphi(t)$, is used that can be translated and dilated in the same way. The wavelets are derived from a so-called mother wavelet by dilation and translation factors. The mother wavelet is normalized with zero average and meets the following admissible condition

$$C_w = \int_0^\infty \frac{|\psi(w)|^2}{w} dw < \infty \tag{6}$$

The continuous wavelet transform of a given signal $f(t) \in L^2(R)$ is defined as the inner product of the wavelet function and the signal, i.e.

$$W(a,b) = \langle f(t), \psi_{a,b}(t) \rangle = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \psi_{a,b}\left(\frac{t-b}{a}\right) dt \tag{7}$$

where *f(t)* is the signal to be analyzed and $\psi_{a,b}(t)$ is the wavelet function.

If we select $a=2^j$ and $b=k2^j$, $(j, k) \in Z_2$ in equation (3), W (a,b) translates into a binary wavelet transform. MRA (multi-resolution analysis) is used to carry out the transform of *f(t)*, which is divided into two parts at different scales: the low frequency part and the high frequency part. The low frequency part is called an approximation of the original signal, while the high frequency part called details respectively. The decomposition equation is

$$\begin{cases} c_{j-1} = \sum_{k \in z} h^*(k-2n) c_j(k) \\ d_{j-1} = \sum_{k \in z} g^*(k-2n) c_j(k) \end{cases} \tag{8}$$

Where *h(k)* and *g(k)* are two-scale series; *h*(k)* and *g*(k)* are complex conjugate functions of *h(k)* and *g(k)*.

A wavelet base is employed in the hidden layer of WNN rather than a sigmoid function, which discriminates it from general back propagation neural networks [7].

The training algorithm of WNN is elucidated as following steps:

Step1. Data acquisition: The output response voltage signals of test point are sampled in terms of the SY running under different input step voltage. Then optimal features for training neural networks are obtained by wavelet decomposing coefficients and normalization [8].

Step2. Select the parameters of wavelet neural network: The dimension of fault feature vectors and the circuit fault pattern determines the number of input node and output node. The output number is equal to the number of fault classes, while input number is equal to the dimension of feature vectors. The number of neurons on hidden layer is set greater or equal to $\sqrt{M+N}+a$, where M and N is the node number of input layer and output layer and a is set 1~10.

Step3. Training of wavelet neural network: The different input step voltage vectors, as input vectors of training pattern, are used to train the wavelet neural network. The output vectors reflect the states of SY. The gradient descent algorithm [9] is employed to minimize the error function so as to adjust the weights of network. The error function is

$$E = \frac{1}{2} \sum_{p=1}^{p} \sum_{j=1}^{j} \left(y_j^p(t) - \overset{*\,p}{y}_j(t) \right) \tag{9}$$

Where P is the total number of training patterns $\overset{*\,p}{y}_j(t)$ and $\overset{p}{y}_j(t)$ are the desired and real output associated with the j*th* feature for the p*th* neuron.

Parameter identification method based on WNN

The parameter identification method based on Wavelet neural network (WNN) is a method of explicit parameter identification [10]. The rule of the method is the mean square error method commonly used in system identification. The block diagram of the identification principle based on Wavelet neural network is given in Figure 3.

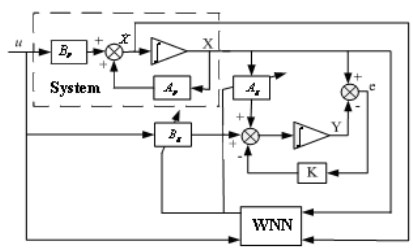


Figure 3 The block diagram of the identification principle based on WNN

The object function used to identify the transfer function of Coupled Industrial Tank system can be calculated eq.(9):

The transfer function of Coupled Industrial Tank

system can be described as shown in Figure 3. The optimization process is to get the optimal parameters A_g, B_g which can make Q minimum.

4 Results and Discussion

Parameters Identification

For parameters identification, a 5 volt step input is applied to drive the industrial tank system. The measured step response of tank system is shown in the Figure 5. The system parameter K0 is obtained from the steady state response of the system which is 17.8 and the tank parameters a_1 and a_2 are 196 and 5169 respectively.

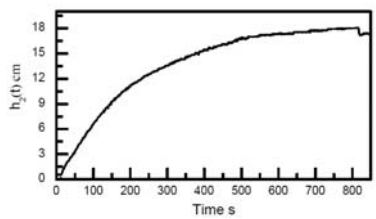


Figure 4 Response to 5V step input

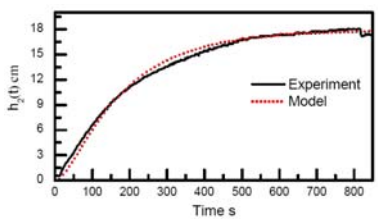


Figure 5 Measured and simulated responses

Model Validation

Validation process is done to examine the effectiveness of the obtained model to represent the industrial tank system. The validation process is done by comparing between real industrial tank system response and simulated response using the obtained model for the same input voltage.

5 Conclusion

Parameter estimation in time domain based on Wavelet neural network (WNN) is a simple and straight approach of dynamic systems identification. The characteristic of this approach is that the outputs of the neural cells are corresponding to the parameters to be identified, according to the mechanism of dynamic system. That is to say the process of Wavelet neural network to convergence is just the process of parameters to be identified. The other characteristic of WNNs is that it will be implemented using electronic circuit possibly, so it will have distinctly applications in online identification.

References

[1] Visioli, "A new design for a PID plus feedforward controller", Journal of Process Control, Vol. 14 pp. 457–463, 2004

[2] K.K. Tan, S. Huang and R. Ferdous, "Robust self tuning PID controller for nonlinear systems", Journal of Process Control, Vol. 12, pp. 753–761, 2002

[3] N.B. Almutairi and M. Zribi, "Sliding mode control of coupled tanks", Mechatronics, Vol. 16, Issue 7, pp. 427-441, 2006

[4] H. Pan, H. Wong, V. Kapila and M.S. de Queiroz, "Experimental validation of a nonlinear back stepping liquid level controller for a state coupled two tank system", Control Engineering Practice, Vol. 13, pp. 27–40, 2005

[5] B.Q. Lv, T.Z. Li, " A Fault Detection Approach Based on Wave-net", Journal Control Theory and Application, 1998, 15(5):802-805

[6] M. Aminian, F. Aminian, "Neural-network based analog-circuit fault diagnosis using wavelet transform as preprocessor", IEEE Transaction Circuits and Systems II: Analog Digital Signal Process, 2000, 47(2):151-156

[7] I. Statish, "Short-time Fourier and Wavelet transforms for fault detection in power transformers during impulse tests", Proceeding IEEE Sci. Meas. Technol, 1998, 145(2):77-84

[8] Dahai Zhang; Yanqiu Bi; Yanbing Bi; Yantao Sun; "Design and initialization algorithm based on modulus maxima of wavelet transform for wavelet neural network", Power System Technology, 2004. PowerCon 2004. 2004 International Conference on. Volume 1, 21-24 Nov. 2004 ,1 (s):897 – 901

[9] Faa-Jeng Lin; Rong-Jong Wai; Mu-Ping Chen; "Wavelet neural network control for linear ultrasonic motor drive via adaptive sliding-mode technique". Ultrasonics, Ferroelectrics and Frequency Control, IEEE Transactions on Volume 50, Issue 6, June 2003 pp. 686 - 698

[10] Dazhi Wang; Jie Yang; Xiaoqin Liu; Qing Yang; Kenan Wang; "Wavelet Neural Network Approach for Fault Diagnosis to a Chemical Reactor". Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on Volume 2, 21-23 June 2006 pp.5764 – 5768

Analysis and Improvement of Linux File-system

Ping Xiong

Information School, Zhongnan University of Economics and Law, Wuhan 430074, China
Email: pingxiong01@126.com

Abstract

Linux operating system is widely used in all kinds of servers and supports manifold file-systems. But the reliability of the file-systems has not been validated completely. Furthermore, the stability of file-systems is extremely important to the security of servers. After using test cases in syscall-level to verify the I/O reliability, several bugs in the kernel of Linux operating system are discovered. By analyzing these critical bugs, an amelioration proposal is brought out in detail which would improve the I/O reliability of file-systems distinctly. Finally, the method of testing the reliability of file-system is discussed to meet the requirement of automatic testing in bug analysis.

Keywords: Linux, file-system, reliability

1 Introduction

File-system is an indispensable component in all kinds of operating systems. It plays a key role in high reliability systems, such as backup systems and databases[1]. The file- system is an important part of any operating system. After all, it's where users keep their stuff. As the primary repository for permanent data on a computer system, a file-system must shoulder the heavy burden of 100% reliability. Testing of a file system must be thorough and extremely strenuous.

For a database system, in order to ensure the safety of the important information, some renewing points should be set in regular intervals. Once an accident occurs, the status of the system at the last renewing point could be resumed. At the renewing point, the system must use a specific interface provided by file-systems to ensure that data could be written to the disk dependably. However, if some I/O errors occur and the file-system can not return exact results at the moment, the renewing process could be interrupted unpredictably. Therefore, the reliability of file-systems should be ensured for a stable operating system.

Generally, the operating systems supporting manifold file-systems, such as UNIX, Linux, and Solar, are adopted in all kinds of high reliability systems[2]. Among them, Linux operating system is widely used in various servers. It is supported by a large group of engineers contributing back into the open source (as they do into the FSF's GNU tools). This makes Linux a very dynamic and fast moving operating system. The 2.6.11 kernel of Linux supports the mainstream file-systems including EXT2, EXT3, JFS, ReiserFS, and XFS etc[3].

However, it is not certain that data could be written correctly in any case with the file-systems and proper error type could be returned while any error occurs. For example, EXT2 performs almost all operations in memory until it needs to flush the buffer cache to disk and makes no guarantees about consistency of the file system or whether an operation is permanently on the disk when a file system call completes[4]. Thus, the reliability of the file-systems should be validated completely.

The methods of code audit and software testing could be applied in reliability analysis of file-system generally[5]. code audit is a comprehensive analysis of source code in a programming project with the intent of discovering bugs, security breaches or violations of programming conventions. Software Testing aims at executing a program or system with the intent of finding errors and evaluating an attribute or capability of a program or system and determining that it meets its required results. But for Linux operating system, code

audit could hardly find problems because of the kernel's complexity. In addition, it is difficult to make unit test and system test for the coupling of functions in Linux kernel[6].

In this paper, we use test cases to validate the I/O reliability of file system in syscall-level. Several bugs in the file system are discovered and analyzed, and an amelioration proposal is brought out in detail which would improve the I/O reliability of file-systems distinctly.

The subsequent sections are organized as follows. Section 2 describes the test project including designing test cases, test method and test objects; the test results are analyzed in section 3 and an amelioration proposal is brought out in section 4. Conclusions are given in the last section.

2 System Testing

There are three methods of tests we can run against a file system: synthetic tests, real-world tests, and end user testing. Synthetic tests are written to expose defects in a particular area (file creation, deletion, etc.) or to test the limits of the system (filling the disk, creating many files in a single directory, etc.)[7]. Real-world tests stress the system in different ways than synthetic tests do and offer the closest approximation of real-world use. Finally, end user testing is a matter of using the system in all the unusual ways that a real user might in an attempt to confuse the file system. In this paper, we apply the method of user testing.

2.1 Test cases

Test cases are the specific inputs that will be tried and the procedures that we'll follow when we test the software. And designing good test cases is a complex art[8]. The design of tests is subject to the same basic engineering principles as the design of software. Good design consists of a number of stages which progressively elaborate the design. Good test design consists of a number of stages which progressively elaborate the design of tests: test strategy, test planning, test specification, test procedure. In the system testing, the test cases should be concise to avoid additional errors. At the same time, the cases must include every syscalls which correlated to I/O[9]. Therefore, our cases consisted of following sequences:

i. read(2)/readv(2)

ii. write(2)/writev(2) asynchronism mode

iii. write(2)/writev(2) synchronization mode — open(2) using O_SYNC

iv. write(2)/writev(2) synchronization mode — open(2) using O_DSYNC

v. write(2)/writev(2) + fsync(2)

vi. write(2)/writev(2) + fdatasync(2)

vii. mmap(2) + read(2) equivalence opration

viii. mmap(2) + write(2) equivalence opration + msync(2) using MS_SYNC

ix. mmap(2) + write(2) equivalence opration + msync(2) using MS_ASYNC

2.2 Test method

We simulated I/O errors by unplug USB storage device in the testing process. The process includes several steps as follows:

i. Compile the test programs on local disk.

ii. Set up the file-systems which would be tested on an USB device(EXT2,EXT3,JFS).

iii. Create test files on each file-systems to be tested.

iv. Restart the host and switch to the single-user access mode to avoid the process such as Haldeamon which may disturb the tests.

v. Load the file-system on the USB device and run the test program. Unplug the USB device in each interval (the default time is 3 seconds) of syscalls and record the results returned by the test program at the same time.

vi. Unloaded file-system. Repeat the step vi to test the other file-systems until all file-systems and test cases are completely tested.

2.3 Test objects

The main file-systems in Linux 2.6.11 kernel including EXT2 / EXT3 / JFS / XFS / ReiserFS were

tested in the project. Moreover, in EXT3, we tested three different mode that set data=journal/ordered/ writeback, also tested three kinds of data and log modes when set data=ordered (Data to disk, log to USB. Data to USB, log to disk. Data to USB, log to USB). For Resier4 was not included in Linux official kernel, we didn't test it.

3 Test Results

The test results is shown in Table 1.

Table 1 Test results

Test	File-systems				
	EXT2	EXT3	JFS	XFS	ReiserFS
Test 1	E1 E2 E3	E1 E2 E3 E6	E2 E3	√	E2 E3 E8
Test 2	E2 E4	E2 E4 E5	E2 E4	√	E2 E5
Test 3	E2	E2 E5	E2	√	E5 E10
Test 4	E2	E2 E5	E2	√	E5 E11
Test 5	E2	E2 E5	E2 E7	√	√
Test 6	E2	E2 E5	E2	√	√
Test 7	E1 E2 E3	E1 E2 E3	E2 E3	E3 E9	E2 E3 E9
Test 8	E2	E2 E5	E2	E9	E9 E5
Test 9	E2 E4	E2 E4 E5	E2 E4	E4 E9	E4 E9 E5

The symbols in table 1 represent all kinds of errors in the test. The meaning of these symbols are listed as follows:

E1: Open operate returned –ENOENT instead of –EIO;

E2: For those repetitive open-read/write-close operations, if I/O error occurred between read/write and close, then the next open operation would work normally. If the error happed before the next open operation, then the open operation would return –EIO;

E3: OS didn't detect any error because of page buffer;

E4: Errors occurred when the asynchronous I/O operation couldn't be detected;

E5: Remount in read-only mode automatically;

E6: When data is on disk and the log is on the USB storage device, the I/O errors couldn't be detected;

E7: Fsync can not return error type;

E8: Read returns the page filled with zero;

E9: Mmap returns the page filled with zero;

E10: O_SYNC write error;

E11: D_SYNC write error;

√ : No error.

Generally, the file systems should report –EIO in the next I/O syscall if any I/O error has taken place[10]. However, the test results denote that the file systems reported a wrong error type, or report nothing, which certainly are the bugs of the file systems.

According to the analysis of the errors, the eleven kinds of bugs could be classified into two groups:

i. E1-E6. These bugs might have slight influence to system and some people claims that they should not be thought as errors, such as E5, which caused intense debate on Linux Kernel Mailing List. But in user's point of view, they expect the result should have no relativity to the file-systems. Whatever the file-system was, it must return –EIO instead of –NOENT in EXT2 and –EIO in XFS.

ii. E7-E11. The obvious errors which would cause unpredictable results.

The test results validated that bugs exist in all the file-systems supported by Linux 2.6.11. Each file-system can not achieve user's operation faultlessly. Comparatively, the capability of XFS is somewhat better and ReiserFS is the worst.

4 Improvement on File-System

In our research project, all the errors are analyzed in detail and some patches are designed to get rid of the bugs. In this paper, we only discussed E7 as example.

E7 represents that Fsync can not return the error type. Fsync(2) is a forced synchronizing operation. When an I/O error occurs and the file-system can not return –EIO, the application process which calls fsync(2) will consider the data has been written into the disk correctly, but in fact, the write operation has been interrupted by the I/O error already.

According to the analysis of the function sys_fsync, the synchronization of file consists of three steps as follows[11]:

i. Submit all the pages marked with

PAGECACHE_TAG_DIRTY to I/O system of the disk.

ii. Special synchronization of file system (call the function file->f_op->fsync(file, file->f_dentry, 0)).

iii. the pages marked with PAGECACHE_TAG_DIRTY write back(set the "writeback" field to blank).

In the first step, when the operation is complete, mpage_end_io_write would cut the page from PAGECACHE_TAG_WRITEBACK tree which is mapped from the process because all the I/O operations are asynchronous. When an error occurs in I/O operation, the PG_error flag of pages will be set on by mpage_end_io_write, while the AS_EIO flag of mapping->flags are not set. Thereby, in the third step, sys_fsync will call the function wait_on_page_writeback_range as follows(we have inserted some comments in the codes):

```
static int wait_on_page_writeback_range(struct address_space *mapping, pgoff_t start, pgoff_t end)
    ...
    int ret = 0;
    while ((index <= end) && (nr_pages = pagevec_lookup_tag(&pvec, mapping, &index, PAGECACHE_TAG_WRITEBACK,min(end - index, (pgoff_t) PAGEVEC_SIZE-1) + 1)) != 0) { ← this loop wasn't executed when an error occurs on I/O
    ...
    }
    /* Check for outstanding write errors */
    if (test_and_clear_bit(AS_ENOSPC, &mapping->flags))
    ret = -ENOSPC;
    if (test_and_clear_bit(AS_EIO, &mapping->flags)) ← this flag wasn't set up
        ret = -EIO;
    return ret; ← function returned zero instead of –EIO
```

The patch we designed in following can set the AS_EIO flag on.

```
diff -uNp linux-2.6.11.11-orig/fs/mpage.c linux-2.6.11.11-new/fs/mpage.c
--- linux-2.6.11.11-orig/fs/mpage.c
+++ linux-2.6.11.11-new/fs/mpage.c
@@ -79,8 +79,11 @@ static int mpage_end_io_write(struct bio
        if (--bvec >= bio->bi_io_vec)
                prefetchw(&bvec->bv_page->flags);
-       if (!uptodate)
+       if (!uptodate){
                SetPageError(page);
+               if(page->mapping)
+                       set_bit(AS_EIO, &page->mapping->flags);
+       }
        end_page_writeback(page);
    } while (bvec >= bio->bi_io_vec);
    bio_put(bio);
```

5 Conclusion

Through the test, several bugs in Linux file-systems are discovered. On the basis of error analysis, several patches are designed to get rid of the bugs.

In the test, when I/O errors occurred in system, they could be detected in a time segment of [0,N](N>0), in ideality, N=+∞, all the errors E1 to E11 can be detected at any time when USB equipment is unplugged. But in fact, N was a limited number, maybe nothing could be checked out if N is too small for people to have available reactivity in time.

On the other hand, the test isn't so convenient for it need artificial intervention to pull out USB equipment on time. Developer of Samba gave another advice that building a ram disk and simulating I/O errors via amending ramdisk_aops function in kernel which could test automatically.

References

[1] MK McKusick, WN Joy, SJ Leffler, RS Fabry, "A Fast File System for UNIX", ACM Transactions on Computer Systems, 1984, pp.181－197

[2] GA Bigley, KH Roberts, "The incident command system: High-reliability organizing for complex and volatile task environments", Academy of Management Journal, 2001, pp.1281-1300

[3] Marco Cesati, Daniel P.Bovet, Understanding the Linux

Kernel, 2nd Edition, Published by O'Reilly & Associates. 2005

[4] Dominic Giampaolo, Practical File System Design with the Be File System, 1st edition, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1998

[5] JR Douceur, WJ Bolosky, "A large-scale study of file-system contents", Proceedings of the 1999 ACM SIGMETRICS international conference on measurement and modeling of computer systems, 1999, pp.59-70

[6] Hong Zhu, Patrick A.V.hall, John H.R.May, "Software Unit Test Coverage and Adequacy", ACM Computing Surveys, 1997, pp. 366-427

[7] S.R. Dalal, A. Jain., N. Karunanithi, J.M. Leaton, C.M.Lott, G.C. Patton, B.M. Horowitz, "Model Based Testing in Practice", Proc. 21st International Conference on Software Engineering, pp. 285-294, 1999

[8] Menzies T., et all. "Prediction & verification: Validation methods for calibrating software effort models", In Proceedings of the 27th international conference on Software engineering (ICSE'05), ACM Press, pages 587-595. 2005

[9] Mohagheghi P., Anda B., Conradi R. "Effort estimation of use cases for incremental large-scale software development", In Proceedings of the 27th international conference on Software engineering (ICSE'05), ACM Press, pp. 303–311, 2005

[10] Philippe Kruchten, "Tutorial: introduction to the rational unified process", In ICSE'02: Proceedings of the 24th International Conference on Software Engineering, New York, ACM Press, pp.703–703, 2002

[11] M.G. Baker, J.H. Hartman, M.D. Kupfer, K.W. Shirriff, and J.K. Ousterhout, "Measurements of a Distributed File System", Proceedings of the ACM Symposium on Operating System Principles, 1991, pp. 198-212

Authenticated Multiparty Quantum Secure Direct Communication with Dense Coding

Wenjie Liu[1,2] Hanwu Chen[2] Tinghuai Ma[1]

1 School of Computer & software, Nanjing University of Information Science & Technology Nanjing, Jiangsu 210044, China

2 School of Computer Science and Engineering, Southeast University Nanjing, Jiangsu 210096, China
Email: wenjieliu@163.com; hw_chen@seu.edu.cn; thma@nuist.edu.cn

Abstract

In this paper, three efficient authenticated multiparty quantum secure direct communication protocols are presented by using multi-particle GHZ states and dense coding, each of which is suitable to a special scenario. These protocols show more efficiency on message transmission, and the anticipators' identity key can be repeatedly used for a long term. Security analysis shows they are secure against outer Eve's attacks and inner anticipators' attacks.

Keywords: Quantum Cryptography, Quantum Secure Direct Communication, Quantum Authentication, Dense Coding

1 Introduction

After Wiesner[1] and Bennett et al.[2] found that quantum effects can be used to transmit secrete information in an open quantum channel, quantum cryptography communi- cation has became a hot topic. Since the pioneering work of Bennett and Brassard was published in 1984 [3], a variety of Quantum key distribution (QKD) protocols have been proposed [4-6]. QKD provides a novel way for two legitimate parties to establish a common secret key over a long distance with unconditional security. And the security of some QKD protocols was theoretically proven [7-8].

Recently, a new concept of quantum cryptography communication, quantum secure direct communication (QSDC), was proposed. Different from QKD, the QSDC protocol is to transmit directly the secret message without generating a secret key to encrypt them, so it is more demanding on security: no message is leaked before Eve was detected. In 2002, Beige et al. presented the first QSDC scheme with single photons [9], and soon Bostrom and Felbingeer put forward a "Ping-pong" protocol with EPR state[10]. Since then, many other QSDC protocols have been proposed and actively pursued [11-20]. These QSDC protocols can protect against Eve's eavesdropping attacks, but they cannot prevent other kind of attack, active attack, such as the impersonation attack, the man-in-the-middle attack etc. Recently Lee, Lim and Yang proposed two QSDC protocols with authentication [21], which can resist the anticipators' active attack. But these protocols were proven to be insecure against the authenticator Trent's attacks [22]. We presented two revised effective authenticated QSDC protocols [23], which are secure against Trent's attacks and show more efficient on message transmission.

In this paper, we present three efficient authenticated multiparty quantum secure direct communication (AMQSDC) protocols based on Ref [23]. These protocols show the same efficiency as Ref.[23] on message transmission, and the anticipators' identity key can be repeatedly used. Security analysis shows they are not only secure against outer Eve's attacks, but also against inner anticipators' attacks (including Trent's attacks).

We will depict the three efficient AMQSDC protocols in Sec. 2, analyze the security of them in Sec. 3, and make conclusions in Sec. 4.

2 The Efficient Amqsdc Protocols

For simplicity, assume there are four parties: Alice, Bob and Charlie are three legitimate users of the communication; Trent is the third-party trusted center who will be used to supplies the GHZ states and authenticate the users.

The three proposed AMQSD Protocols all include two phases: the authentication process and the direct communication process. At first, we describe some notations that will be used in this paper.

2.1 Some notations

(1) $|0\rangle, |1\rangle$: The rectilinear basis states.

(2) $|+\rangle, |-\rangle$: The diagonal basis states. They are defined as $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$, $|-\rangle = \frac{1}{\sqrt{2}}(|0\rangle - |1\rangle)$.

(3) $|\phi^{\pm}\rangle, |\psi^{\pm}\rangle$:The four Bell states. They are defined as $|\phi^{\pm}\rangle = \frac{1}{\sqrt{2}}(|00\rangle \pm |11\rangle)$, $|\psi^{\pm}\rangle = \frac{1}{\sqrt{2}}(|01\rangle \pm |10\rangle)$.

(4) $|P^{\pm}\rangle$, $|Q^{\pm}\rangle$, $|R^{\pm}\rangle$, $|S^{\pm}\rangle$: The eight GHZ states. They are defined as $|P^{\pm}\rangle = \frac{1}{\sqrt{2}}(|000\rangle \pm |111\rangle)$, $|Q^{\pm}\rangle = \frac{1}{\sqrt{2}}(|001\rangle \pm |110\rangle)$, $|R^{\pm}\rangle = \frac{1}{\sqrt{2}}(|010\rangle \pm |101\rangle)$, $|S^{\pm}\rangle = \frac{1}{\sqrt{2}}(|011\rangle \pm |100\rangle)$.

(5) $h(.)$: The one-way hash function. The mapping of it is $\{0,1\}^* \to \{0,1\}^N$.

(6) K_{TU}: The n-bit secret key shared between Trent and a user U, such that K_{TA} is the secret key shared between Trent and Alice. It will be used repeatedly as a long-term secret key.

(7) r_{TU}: An l-bit random string chosen by the authenticate center Trent.

(8) $H, I, \sigma_x, i\sigma_y, \sigma_z$: The unitary operations. They are defined as $H = \frac{1}{\sqrt{2}}(|0\rangle\langle 0| - |1\rangle\langle 1| + |0\rangle\langle 1| + |1\rangle\langle 0|)$, $I = |0\rangle\langle 0| + |1\rangle\langle 1|$, $\sigma_x = |0\rangle\langle 1| + |1\rangle\langle 0|$, $i\sigma_y = |0\rangle\langle 1| - |1\rangle\langle 0|$.

Suppose Alice, Bob, Charlie and Trent share the four-particle GHZ state of $|\varphi\rangle_{ABTC} = 1/\sqrt{2}(|0000\rangle_{ABTC} + |1111\rangle_{ABTC})$, where the subscripts A, B, C and T correspond to Alice, Bob, Charlie and Trent, respectively. And $|\varphi\rangle_{ABTC}$ can be rewritten as follows:

$$\begin{aligned} |\varphi\rangle_{ABTC} &= \frac{1}{2}\{(|P^+\rangle_{ABC} + |P^-\rangle_{ABC})|0\rangle_T + \\ &\quad (|P^+\rangle_{ABC} - |P^-\rangle_{ABC})|1\rangle_T\} \\ &= \frac{1}{\sqrt{2}}(|P^+\rangle_{ABC}|+\rangle_T + |P^-\rangle_{ABC}|-\rangle_T) \end{aligned} \quad (1)$$

2.2 The authentication process

The purpose of the authentication process is to let the three users (Alice, Bob and Charlie) safely share the four-particle GHZ states. The authentication process is similar to Ref. [21, 23], and the details of the process can be depicted as follows (see Figure 1):

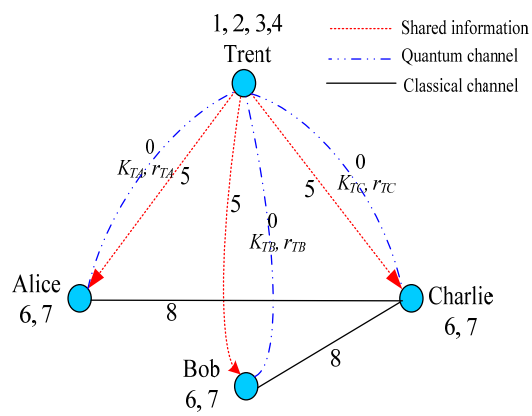


Figure 1 The authentication process of the MAQSDC protocols

(1) Prerequisite. (K_{TA}, r_{TA}) is shared between Trent and Alice in advance, the same with (K_{TB}, r_{TB}) and (K_{TC}, r_{TC}).

(2) Trent prepares $N(n + m_1 + m_2)$ four-particle GHZ states, and each of which is $|\varphi_i\rangle = 1/\sqrt{2}(|0000\rangle_{ABTC} + |1111\rangle_{ABTC})$ $(i = 1, 2, \cdots, N)$. We denoted the N ordered quadruplets as $[(P_1(A), P_1(B), P_1(T), P_1(C)), (P_2(A), P_2(B), P_2(T), P_2(C)), \cdots, (P_N(A), P_N(B), P_N(T), P_N(C))]$. Here the subscript indicates the ordering number of quadruplets, T, A, B and C represent the four qubits of each four-particle GHZ state.

(3) Trent takes particle A from each quadruplet to form an ordered particle sequence $[P_1(A), P_2(A),$

$\cdots, P_N(A)]$, called A-sequence. Similarly, the remaining partner particles compose B-sequence, T-sequence and C-sequence.

(4) With Alice's secret key K_{TA} and the random string r_{TA}, Trent computes $R_{TA} = h(K_{TA}, r_{TA})$ for A-sequence. Similarly, $R_{TB} = h(K_{TB}, r_{TB})$ for B-sequence and $R_{TC} = h(K_{TC}, r_{TC})$ for C-sequence can be computed.

(5) Trent encodes A-sequence according to R_{TA}:

1) If $(R_{TA})_i = 1$, then Trent makes a Hadamard operation H to the i-th qubit of A-sequence.

2) If $(R_{TA})_i = 0$, then identity operation I is applied to the i-th qubit of A-sequence.

Here, $(R_{TA})_i$ .denotes the i-th bit of the R_{TA} string. So the result of the operation on $P_i(A)$ is $P_i'(A) = \{(1-(R_{TA})_i)I + (R_{TA})_i H\}P_i(A)$. Trent encodes B-sequence and C-sequence in the same way.

(6) Trent distributes the encoded A-sequence, B-sequence and C-sequence to Alice, Bob and Charlie through the quantum channel, respectively, and keeps the remaining T-sequence.

(7) On receiving A-sequence, Alice computes $R_{TA} = h(K_{TA}, r_{TA})$ with K_{TA} and r_{TA}, and decodes the qubits by making the reverse unitary operations (I or H) with $(R_{TA})_i$. The result of $P_i'(A)$ becomes $P_i''(A) = \{(1-(R_{TA})_i)I + (R_{TA})_i H\}P_i'(A) = P_i(A)$. Bob and Charlie make the same operations as Alice does, respectively.

(8) Alice selects randomly a sufficiently large subset (m_1 particles) from the decoded qubits, and makes von Neumann measurement on them, so do Bob and Charlie, respectively.

(9) Alice, Bob and Charlie compare their measurement outcomes through the public channel. If the error rate is higher than expected threshold, then Alice, Bob and Charlie abort the communication. Otherwise, they confirm that their counter parts are legitimate and the channel is secure.

If the channel is authenticated to be safe and the users are legitimate, then Alice, Bob and Charlie execute the following message transmission procedures with the remaining $n + m_2$ four-particle states.

2.3 the efficient MAQSDC protocol 1

We assume the scenario is as follows: two or more senders, such as Alice and Bob, want to send secret messages to the distant receiver, Charlie, simultaneously, and there is a quantum channel between each sender and Charlie. To achieve the task, Alice and Bob encode their messages on their particles and transmit these particles directly to Charlie, respectively.

Our first MAQSDC Protocol is composed of two processes: the authentication process and the direct communication process. The authentication is depicted in Sec. 2.2, and the direct communication process is shown as follows (see Figure 2):

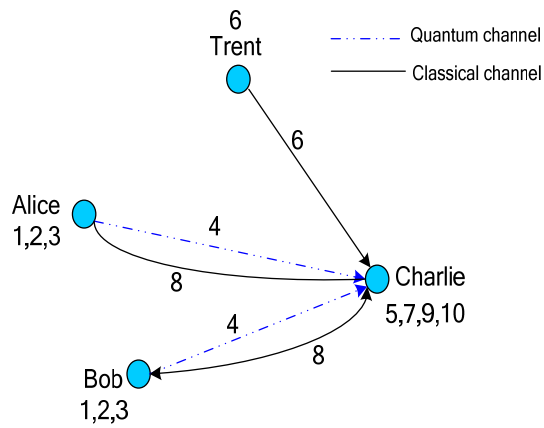


Figure 2 The direct communication process of the MAQSDC protocol 1

Senders:

(1) Alice selects randomly m_2 (sufficiently large) particles from the remaining $(n + m_2)$ A-sequence after the authentication process, and performs randomly one of the four operations $(I, \sigma_x, i\sigma_y, \sigma_z)$ on them. Alice records the positions and the operations of these sample particles. Bob makes random operations on the B-sequence particles in the same positions, and remembers these operations.

(2) Alice encodes the secret message with an error correction code (ECC), such as Hamming code, Reed-Solomon code, CSS code, or BCH code.

(3) Alice encodes the ECC-encoded message on the remanding n-length A-sequence by performing the

four unitary operations $(I, \sigma_x, i\sigma_y, \sigma_z)$. According to our dense coding strategy, two bits message corresponds to a unitary operation: 00 (I), 01 (σ_x), 10 ($i\sigma_y$), and 11 (σ_z). At the same time, Bob encodes his message on the remanding B-sequence particles, and one bit message corresponds to a unitary operation: 0 (I), 1 (σ_x).

1) If Alice's message is "00" and Bob's is "0", then

$$\begin{aligned} I_A I_B |\varphi\rangle_4 &= \frac{1}{\sqrt{2}}(|0000\rangle_{ABTC} + |1111\rangle_{ABTC}) \\ &= \frac{1}{\sqrt{2}}(|P^+\rangle_{ABC}|+\rangle_T + |P^-\rangle_{ABC}|-\rangle_T) \end{aligned} \quad (2)$$

2) If Alice's message is "01" and Bob's is "0", then

$$\begin{aligned} \sigma_{xA} I_B |\varphi\rangle_4 &= \frac{1}{\sqrt{2}}(|1000\rangle_{ABTC} + |0111\rangle_{ABTC}) \\ &= \frac{1}{\sqrt{2}}(|S^+\rangle_{ABC}|+\rangle_T + |S^-\rangle_{ABC}|-\rangle_T) \end{aligned} \quad (3)$$

3) If Alice's message is "10" and Bob's is "0", then

$$\begin{aligned} i\sigma_{yA} I_B |\varphi\rangle_4 &= \frac{1}{\sqrt{2}}(|0111\rangle_{ABTC} - |1000\rangle_{ABTC}) \\ &= \frac{1}{\sqrt{2}}(|S^-\rangle_{ABC}|+\rangle_T - |S^+\rangle_{ABC}|-\rangle_T) \end{aligned} \quad (4)$$

4) If Alice's message is "11" and Bob's is "0", then

$$\begin{aligned} \sigma_{ZA} I_B |\varphi\rangle_4 &= \frac{1}{\sqrt{2}}(|0000\rangle_{ABTC} - |1111\rangle_{ABTC}) \\ &= \frac{1}{\sqrt{2}}(|P^-\rangle_{ABC}|+\rangle_T + |P^+\rangle_{ABC}|-\rangle_T) \end{aligned} \quad (5)$$

5) If Alice's message is "00" and Bob's is "1", then

$$\begin{aligned} I_A \sigma_{xB} |\varphi\rangle_4 &= \frac{1}{\sqrt{2}}(|0100\rangle_{ABTC} + |1011\rangle_{ABTC}) \\ &= \frac{1}{\sqrt{2}}(|R^+\rangle_{ABC}|+\rangle_T + |R^-\rangle_{ABC}|-\rangle_T) \end{aligned} \quad (6)$$

6) If Alice's message is "01" and Bob's is "1", then

$$\begin{aligned} \sigma_{xA}\sigma_{xB} |\varphi\rangle_4 &= \frac{1}{\sqrt{2}}(|1100\rangle_{ABTC} + |0011\rangle_{ABTC}) \\ &= \frac{1}{\sqrt{2}}(|Q^+\rangle_{ABC}|+\rangle_T - |Q^-\rangle_{ABC}|-\rangle_T) \end{aligned} \quad (7)$$

8) If Alice's message is "10" and Bob's is "1", then

$$\begin{aligned} i\sigma_{yA}\sigma_{xB} |\varphi\rangle_4 &= \frac{1}{\sqrt{2}}(|0011\rangle_{ABTC} - |1100\rangle_{ABTC}) \\ &= \frac{1}{\sqrt{2}}(|Q^-\rangle_{ABC}|+\rangle_T - |Q^+\rangle_{ABC}|-\rangle_T) \end{aligned} \quad (8)$$

9) If Alice's message is "11" and Bob's is "1", then

$$\begin{aligned} \sigma_{ZA}\sigma_{xB} |\varphi\rangle_4 &= \frac{1}{\sqrt{2}}(|0100\rangle_{ABTC} - |1011\rangle_{ABTC}) \\ &= \frac{1}{\sqrt{2}}(|R^-\rangle_{ABC}|+\rangle_T + |R^+\rangle_{ABC}|-\rangle_T) \end{aligned} \quad (9)$$

(4) After finishing all operations, Alice and Bob send their encoded qubits to Charlie, respectively.

Receiver:

(5) After Charlie obtains Alice's and Bob's particles, he makes GHZ measurement on pairs of particles consisting of Alice's, Bob's and his own qubit.

(6) Trent measures his qubits in the X basis $\{|+\rangle, |-\rangle\}$ and publishes the measurement outcomes to Charlie.

(7) Charlie can infer both Alice's and Bob's operations from Trent's publication and his outcomes (shown in Table 1). For example, when Charlie's measurement outcome is $|Q^-\rangle_{ABC}$ and Trent's publication is $|+\rangle_T$, Charlie can deduce that Alice had performed a $i\sigma_{yA}$ operation, Bob had performed a σ_{xB} operation. So Charlie knows Alice's bit string is "10" and Bob's is "1".

(8) Alice tells Charlie the sample particles' positions and the relevant operations, and Bob informs his corresponding operations too. Then Charlie can get an estimate of error rate of the direct communication process.

(9) If the error rate is reasonably low, the users affirm the process is secure, and Charlie gets rid of the sample particles and gets Alice's and Bob's bit strings with their respective dense encoding strategy. Otherwise, they abandon the transmission and resume the direct communication procedure.

(10) Charlie does error correction on Alice's and Bob's bit strings with the ECC code, and finally gets the secret messages from Alice and Bob, respectively.

Table 1 Relations of Alice's operation, Bob's operation, Trent's publication, and Charlie's measurement in the AMQSDC protocol 1

Trent's publication	*Charlie's measurement*	*Alice's operation*	*Bob's operation*
$\|+\rangle_T$	$\|P^+\rangle_{ABC}$	I_A (00)	I_B (0)
$\|+\rangle_T$	$\|S^+\rangle_{ABC}$	σ_{xA} (01)	I_B (0)
$\|+\rangle_T$	$\|S^-\rangle_{ABC}$	$i\sigma_{yA}$ (10)	I_B (0)
$\|+\rangle_T$	$\|P^-\rangle_{ABC}$	σ_{zA} (11)	I_B (0)
$\|-\rangle_T$	$\|P^-\rangle_{ABC}$	I_A (00)	I_B (0)
$\|-\rangle_T$	$\|S^-\rangle_{ABC}$	σ_{xA} (01)	I_B (0)

$\|-\rangle_T$	$\|S^+\rangle_{ABC}$	$i\sigma_{yA}$ (10)	I_B (0)
$\|-\rangle_T$	$\|P^+\rangle_{ABC}$	σ_{zA} (11)	I_B (0)
$\|+\rangle_T$	$\|R^+\rangle_{ABC}$	I_A (00)	σ_{xB} (1)
$\|+\rangle_T$	$\|Q^+\rangle_{ABC}$	σ_{xA} (01)	σ_{xB} (1)
$\|+\rangle_T$	$\|Q^-\rangle_{ABC}$	$i\sigma_{yA}$ (10)	σ_{xB} (1)
$\|+\rangle_T$	$\|R^-\rangle_{ABC}$	σ_{zA} (11)	σ_{xB} (1)
$\|-\rangle_T$	$\|R^-\rangle_{ABC}$	I_A (00)	σ_{xB} (1)
$\|-\rangle_T$	$\|Q^-\rangle_{ABC}$	σ_{xA} (01)	σ_{xB} (1)
$\|-\rangle_T$	$\|Q^+\rangle_{ABC}$	$i\sigma_{yA}$ (10)	σ_{xB} (1)
$\|-\rangle_T$	$\|R^+\rangle_{ABC}$	σ_{zA} (11)	σ_{xB} (1)

2.4 The efficient MAQSDC protocol 2

The scenario of the second efficient MAQSDC protocol is as follows: Alice and Bob want to send secret messages to the distant receiver Charlie simultaneously, but there is no quantum channel between Alice and Charlie (or Bob and Charlie). To finish the task, Alice encodes message on her particles and transmits them to Trent, instead of Charlie. Bob encodes message and still transmits to Charlie.

The second MAQSDC protocol is composed of the authentication process (Sec. 2.2) and the direct communication process. The direct communication process is similar to the first protocol, except that Alice transmits her particles to Trent and Bob transmits his particles to Charlie in step (4), Charlie make Bell measurement in step (5), and Trent performs Bell measurement in step (6). And the procedure is shown in Figure 3.

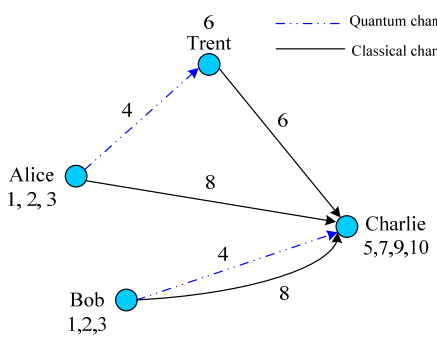


Figure 3 The direct communication process of the MAQSDC protocol 2

2.5 The efficient MAQSDC protocol 3

The scenario of the third efficient MAQSDC protocol is as follows: Alice and Bob want to send secret messages to the distant receiver Charlie simultaneously, but there is no quantum channel between each sender (Alice, Bob) and Charlie. To accomplish the task, Alice and Bob encode messages on her particles and transmit them to Trent, instead of Charlie.

The third MAQSDC protocol is consisted of the authentication process (Sec. 2.2) and the direct communica-tion process. The direct communication process is similar to the first protocol, except that Alice and Bob transmit their particles to Trent in step (4), Charlie makes X basis measurement in step (5), and Trent performs GHZ measurement in step (6). And the procedure is shown in Figure 4.

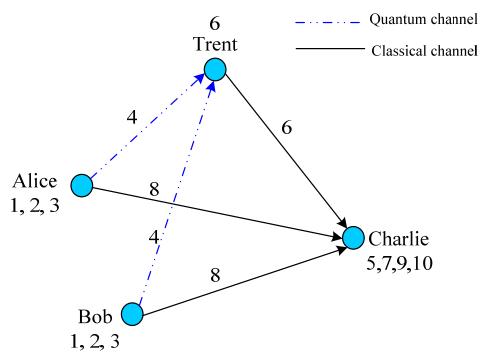


Figure 4 The direct communication process of the MAQSDC protocol 3

3 Security Analysis

The security of the MAQSDC protocols is based on the security of the authentication process and the direct communication process. Because of the similarity of these three protocols, we just take the first MAQSDC protocols into account.

(1) The authentication process is secure against Eve's attacks.

The authentication process is similar to Ref.[21, 23], and the proof of security was given. As is proved in the

two works, the process is secure against not only passive attack (such as the intercept-resending attack, the entanglement attack, etc.), but also active attack (such as the impersonation attack, the man-in-the-middle attack etc.).

Note that, there is no information to be leaked out before Eve was checked out. And if the process is confirmed to be safe, the four-particle GHZ states are safely shared among legitimate users and Trent.

(2) The direct communication process is secure against Eve's attacks.

During the process, the senders Alice and Bob send their particles to the receiver Charlie through quantum channels. Eve may take the intercept-resending attack, the entanglement attack to steal information about the secret message.

If Eve intercepts the qubits from Alice, impersonates Alice and resend spurious qubits to Charlie, then she will be detected when Charlie checks security by measuring the GHZ states, because Eve destroys the entanglement among Alice's, Bob's and Charlie's qubits. For example, if Alice's message is "00", Bob's message is "0" and Eve's spurious qubit is $|E\rangle = 1/\sqrt{2}(|0\rangle + |1\rangle)$, then the final state of the system composed of Eve's, Bob's and Charlie's qubits is as follows where the subscript E corresponds to Eve's spurious qubit. Obviously, Charlie will detect Eve's existence and Eve can't get any information about the message by measuring his intercepting qubits.

$$\begin{aligned}
|\Psi\rangle_{\mathrm{EABTC}} &= |E\rangle_E \otimes (I_A I_B |\varphi\rangle_{\mathrm{ABTC}}) \\
&= \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)_E \otimes \frac{1}{\sqrt{2}}(|0\rangle_A |000\rangle_{BTC} + \\
&\quad |1\rangle_A |111\rangle_{BTC}) \\
&= \frac{1}{2}|0\rangle_A \otimes (|000\rangle_{EBC} + |100\rangle_{EBC}) \otimes |0\rangle_T + \\
&\quad \frac{1}{2}|1\rangle_A \otimes (|011\rangle_{EBC} + |111\rangle)_{EBC}) \otimes |1\rangle_T \\
&= \frac{1}{2}|0\rangle_A \otimes (|P^+\rangle + |P^-\rangle + |S^+\rangle - |S^-\rangle)_{EBC} \otimes \\
&\quad |0\rangle_T + \frac{1}{2}|1\rangle_A \otimes (|S^+\rangle + |S^-\rangle + |P^+\rangle - \\
&\quad |P^-\rangle)_{EBC} \otimes |1\rangle_T
\end{aligned} \tag{10}$$

If Eve prepares some qubits and entangle them with Alice's qubits, she would make unitary operation U_{AE} on the pair of Alice's and her qubit $|E\rangle$:

$$U_{EA}|E0\rangle_{EA} = \alpha|e_{00}\rangle_E|0\rangle_A + \beta|e_{01}\rangle_E|1\rangle_A \tag{11}$$

$$U_{EA}|E1\rangle_{EA} = \alpha'|e_{10}\rangle_E|0\rangle_A + \beta'|e_{11}\rangle_E|1\rangle_A \tag{12}$$

where $|\alpha|^2 + |\beta|^2 = 1$, $|\alpha'|^2 + |\beta'|^2 = 1$, and $\alpha\beta^* + \alpha'\beta'^* = 0$. Suppose Alice's message is "00" and Bob's message is "0", then the state after Alice's and Bob's message encoding is

$$\begin{aligned}
|\Psi_1\rangle_{\mathrm{EABTC}} &= |E\rangle_E \otimes (I_A I_B |\varphi\rangle_{\mathrm{ABTC}}) \\
&= |E\rangle_E \otimes \frac{1}{\sqrt{2}}(|000\rangle_{ABC}|0\rangle_T + |111\rangle_{ABC}|1\rangle_T)
\end{aligned} \tag{13}$$

and the final state after Eve's entanglement operation with Alice's qubit is shown in Eq.(14).

$$\begin{aligned}
|\Psi_1'\rangle_{\mathrm{EABTC}} &= \mathrm{U_{EA}}|\Psi_1'\rangle_{\mathrm{EABTC}} \\
&= \frac{1}{\sqrt{2}}\mathrm{U_{EA}}|E\rangle_E \otimes (|000\rangle_{ABC}|0\rangle_T + \\
&\quad |111\rangle_{ABC}|1\rangle_T) \\
&= \frac{1}{2\sqrt{2}}\{|P^+\rangle_{ABC}[|+\rangle_T(\alpha|e_{00}\rangle_E + \beta'|e_{11}\rangle_E) + \\
&\quad |-\rangle_T(\alpha|e_{00}\rangle_E - \beta'|e_{11}\rangle_E)] + |P^-\rangle_{ABC} \\
&\quad [|+\rangle_T(\alpha|e_{00}\rangle_E - \beta'|e_{11}\rangle_E) + |-\rangle_T(\alpha|e_{00}\rangle_E + \\
&\quad \beta'|e_{11}\rangle_E)] + |S^+\rangle_{ABC}[|+\rangle_T(\alpha'|e_{10}\rangle_E + \\
&\quad \beta|e_{01}\rangle_E) - |-\rangle_T(\alpha'|e_{10}\rangle_E - \beta|e_{01}\rangle_E)] + \\
&\quad |S^-\rangle_{ABC}(|+\rangle_T(\alpha'|e_{10}\rangle_E - -\beta|e_{01}\rangle_E) - \\
&\quad |-\rangle_T(\alpha'|e_{10}\rangle_E + \beta|e_{01}\rangle_E)]\}
\end{aligned} \tag{14}$$

As shown in Eq. (13-14), Eve introduces error in the check bits with the probability of 1/2 regardless of the order of measurement by Bob, Trent, and Eve. Moreover, Eve cannot get any information from this kind of attack since she cannot distinguish which operation Alice had performed.

(3) The MAQSDC protocols are secure against Trent's and other participators' attacks.

We start with Trent's attack on the first protocol. If Trent attempts to gain Alice's message and takes the intercept-resending attack, suppose the procedure is as follows: At first, Trent intercepts Alice's qubits heading to Charlie, measures Alice's and his qubits separately in

the Z basis $\{|0\rangle,|1\rangle\}$, and tries to deduce Alice's operations according to the above measurement outcomes. Because Trent disturbs the entanglement of the GHZ states composed of Alice's, Bob's, Charlie's and his qubits, he will be detected as soon as Charlie checks the security. In addition, Trent cannot get any information about Alice's message. For example, if the outcomes of Alice's and Trent's are different, Trent doesn't know Alice had performed the σ_x operation or the $i\sigma_y$ operation (see Eq. (15-16)). Therefore, Trent knows nothing about Alice's message.

$$\sigma_{xA} I_B |\varphi\rangle_4 = \frac{1}{\sqrt{2}}(|1\rangle_A |0\rangle_T |00\rangle_{BC} + |0\rangle_A |1\rangle_T |11\rangle_{BC}) \quad (15)$$

$$i\sigma_{yA} I_B |\varphi\rangle_4 = \frac{1}{\sqrt{2}}(|0\rangle_A |1\rangle_T |11\rangle_{BC} - |1\rangle_A |0\rangle_T |00\rangle_{BC}) \quad (16)$$

If Trent takes the entanglement attack and entangles Alice's qubits in order to gain Alice's message, the case is as same as Eve's entanglement attack discussed in the above section, and Trent just acts on Eve's role. So, this attack won't success under our protocol, either.

In our protocol, there are two senders Alice and Bob. if Alice (Bob) wants to steal Bob's (Alice's) message, her (his) attacks is just the same as Trent's attacks. Then, we can say, our protocols can protect against other participators' attacks too.

In summary, our first MAQSDC protocol is secure against not only Eve's attacks, but also inner participators' attacks (including Trent's attacks), and the same security is shown in the protocol 2 and the protocol 3.

4 Conclusions

In this paper, we introduce three correspond efficient MAQSDC protocols according to different scenarios. These protocols are proved to be secure against outer Eve's attacks and inner anticipators' attacks (also including Trent's attacks). They show more powerful security than the protocols in Ref.[21].

In addition, these MAQSDC protocols show more efficiency, the key points are as follows: (1) The secrete identity key K_{TU} can be used repeatedly, just as their identity card, and the encryption string R_{TU} can refresh each communication round by changing r_{TU}. (2) These protocols show the same efficiency on message transmission as Ref. [23], and transmit more messages per GHZ state than Ref. [21].

Acknowlegement

This work is supported by the National Natural Science Foundation of China (60572071), and also partly supported by Jiangsu Provincial Natural Science Foundation of China (BM2006504, BK2007104) and the Natural Science Foundation of College of Jiangsu Province, China (06KJB520137).

References

[1] S. Wiesner, "Conjugate Coding", Sigact News, 15, 1983, pp.78-88

[2] C. H. Bennett, G. Brassard, S. Breidbart and S. Wiesner, "Quantum cryptography, or unforgeable subway tokens", Advances in Cryptology: Proceedings of Crypto '82, August 1982, Plenum Press, pp.267-275

[3] C. H. Bennett and G. Brassard, "Quantum cryptography: Public-key distribution and tossing", in Proceedings of IEEE International Conference on Computers, Systems and Signal Processing, Bangalore, India, IEEE Press, 1984, pp.175-179

[4] A. K. Ekert, "Quantum cryptography based on Bell's theorem", Phys. Rev. Lett. vol., 67, 1991, pp.661-663

[5] C. H. Bennett, "Quantum cryptography using any two nonorthogonal states", Phys. Rev. Lett., vol. 68, 1992, pp.3121-3124

[6] D. Bruß, "Optimal Eavesdropping in Quantum Cryptography with Six States", Phys. Rev. Lett., vol. 81, 1998, pp.3018-3021

[7] H. K. Lo and H. F. Chau, "Unconditional security of quantum key distribution over arbitrarily long distances", Science, vol. 283, 1999, pp. 2050-2056

[8] P. W. Shor, J. Preskill, "Simple Proof of Security of the

BB84 Quantum Key Distribution Protocol", Phys. Rev. Lett., vol. 85, 2000, pp.441-444

[9] A. Beige et al., "Secure communication with single-photon two-qubit states", Acta Phys. Pol. A 101, 2002, pp.357-361

[10] K. Bostro¨m and T. Felbinger, "Deterministic Secure Direct Communication Using Entanglement", Phys. Rev. Lett., vol. 89, 2002, 187902

[11] Q. Y. Cai, B. W. Li, "Improving the capacity of the Boström-Felbinger protocol", Phys. Rev. A, vol. 69, 2004, 05430

[12] F. G. Deng, G. L. Long and X. S. Liu, "Two-step quantum direct communication protocol using the Einstein-Podolsky-Rosen pair block", Phys. Rev. A, vol. 68, 2003, 042317

[13] F. G. Deng and G. L. Long, "Secure direct communication with a quantum one-time pad", Phys. Rev. A, vol. 69, 2004, 052319

[14] T. Gao, F. L. Yan, Z. X. Wang, "Deterministic secure direct communication using GHZ states and swapping quantum entanglement", J. Phys. A: Math. Gen., vol. 38, 2005, pp.5761–5770

[15] C. Wang et al., "Multi-step quantum secure direct communication using multi-particle Green-Horne-Zeilinger state", Opt. Commun., vol. 253, 2005, pp.15-20

[16] Q. Y. Cai, B. W. Li, "Deterministic secure communication without using entanglement", Chin. Phys. Lett. Vol. 21, 2004, 601

[17] Z. X. Man, Z. J. Zhang and Y. Li, "Deterministic secure direct communication by using swapping quantum entanglement and local unitary operations", Chin. Phys. Lett. 22, 2005, pp.18-21

[18] C. Wang, F. G. Deng, Y. S. Li, et al., "Quantum secure direct communication with high-dimension quantum superdense coding", Phys. Rev. A, vol. 71, 2005, 044305

[19] G. Y. Wang, X. M. Fang, X. H. Tan, "Quantum Secure Direct Communication with Cluster State", Chinese Phys. Lett., vol. 23, 2006, pp:2658-2661

[20] X. Lu, Z. Ma, D. G. Feng, "Quantum Secure Direct Communication Using Quantum Calderbank-Shor- Steane Error Correcting Codes", Journal of Software, vol. 17, No.3, 2006, pp.509-515

[21] H. Lee, J. Lim and H. Yang, "Quantum direct communication with authentication", Phys. Rev. A, vol. 73, 2006, 042305

[22] Z. J. Zhang, J. Liu, D. Wang, et al., "Comment on 'Quantum direct communication with authentication'", Phys. Rev. A, vol. 75, 2007, 026301

[23] W. J. Liu, H. W. Chen, Z. Q Li, et al., "Efficient quantum secure direct communication with authentication", quant-ph/0711.3502, 2007

The Effect of Mobility on Epidemic Spreading

Luosheng Wen[1,2] Jiang Zhong[3]

1 College of Mathematics and Physics, Chongqing University, Chongqing, 400044, China

2 College of Computer Science, Chongqing University, Chongqing, 400044, China
Email: wenluosheng@yahoo.cn;

Abstract

As mobile networks become more common and mobile phones are attacked more frequently, many researchers have paid their attentions to the effect of mobility on epidemic spreading. We build two models to describe the effect of mobility on epidemic spreading on regular lattice and scale-free network. Theoretical results and simulations show that the mobility leads to the epidemic threshold deceasing and the infected density increasing on both networks. The conclusion suggests that more efforts should be made for controlling mobile phone's viruses spreading.

Keywords: mobile phone virus, regular lattice, scale-free network, epidemic threshold, SIS model

1 Introduction

For most of people, although those troubles about mobile phone viruses are not faced yet, almost all evidences show that the damages resulted in them will reach to the extent of computer viruses. On the one hand, that smart phones popularize rapidly because of the development of electrical techniques gives mobile phone viruses a hardware platform and broader attacked objectives. Nowadays, the CPU and hard disk of smart phone are close to the basic configuration of personal computer in 2000. Market research firm IDC predicts that by 2008, vendors will sell more than 130 million smart phones, representing 15 percent of all mobile phones. ARC Group, another market research firm, said 27 million smart phones were sold worldwide in 2004, accounting for about 3 percent of the total global handset market [1]. On the other hand, the increase of sorts and sum of mobile phone viruses is more quickly than that of computer viruses because those mobile phone virus makers have richer experience. On the contrary, users of mobile phone have little knowledge on phone viruses. Since the virus appears, especially after 2005, the sum of viruses reported is remarkably increasing. In the 18 months from June 2004 to December 2005, about 130 sorts of phone viruses had been reported. Furthermore, the rapid data transmission by 3G network facilitates virus invasion. Although the sum of mobile phone viruses isn't large, according to the survey data from Kaspersky, the number in 2006 is 1.5 times as large as that in 2005 [2,3].

Since mobile phone viruses are similar with biology viruses and computer viruses, researching on mobile phone virus spreading naturally can borrow from them. Kermack and Mckendrick[4][5] presented the SIS model and analyzed epidemic threshold in their classical papers. On modeling computer virus, Kephart and White [6,7] presented the directed graph model in 1991.

With the increasing of mobile phone viruses, a few researchers model them spreading. Most of their models are based on those viruses by Bluetooth transmission. Recently, paper [8] shows an epidemic model in which the mobile phone with variable velocity is discussed. Paper [9] compares to the required condition of virus spreading in computer and gives the corresponding required condition of virus spreading in MANETs by simulation. Zheng et al. [10,11] introduced a model that contains movement velocity, distribution density and signal coverage radius. But the models didn't consider the fact that movement of phones result in infection rate decreasing. Paper [12] showed the effect of a combination of spatial and temporal correlations on the

threshold behavior.

On the other hand, the influence of network's topology on epidemic spreading is very important. Many networks, such as the Internet network, World Wide Web and Email network, were demonstrated that their degree distributions satisfy scale-free property, namely these networks are scale-free networks [11]. Similarly, human social relations network is also scale-free network. So we suggest that the topology of contacts network should be considered when modeling mobile phone virus on the network.

In this paper, we analyze the limitation of the KW model in mobile network. By giving some parameters new definitions, we present two SIS models, one of which deals with the regular lattice and another deals with the scale-free network. By comparing a static model with a dynamic model, we show mobility's effect on phone virus spreading. In the SIS model on random graph, we find that both the epidemic threshold and the infection density in steady state are larger than those in static model and less than those in KW model. Furthermore, we also show that the movement of phones decreases the epidemic threshold in the second model.

The materials are organized in this fashion: Section 2 we show the limitations of some existing models and analyze important factors in mobile environment. Section 3 presents two models for the random graph and the scale-free network, and then we give theoretical and numerical results. Finally, some conclusions are drawn in Section 4.

2 Analysis on Virus Spreading By Bluetooth Function

Up to date, the Cabir virus is the most representative virus by Bluetooth transmission. It spreads using the Bluetooth function in the phones and shows itself as "cabir.sis", which contains the virus. If the user chooses to install the application in curiosity the worm instantly starts to look for other units to infect via Bluetooth. Can we use the KW model to modeling the spreading?

Firstly, we recall the KW model. The Internet network is a graph, in which computers are viewed as nodes and physical connection between them are viewed as edges. Computers have two discrete statuses: susceptible (S) and infected (I). The KW model described the virus propagation is

$$\frac{dI}{dt} = \beta < k > I(1-I) - \delta I \tag{1}$$

where I denotes the proportion of infected computers，$<k>$ is the average connectivity, β is infection rate of virus, δ denotes cure rate of infected computer.

In fact, the KW model potentially assumes that the connectivity of every node is approximately equal and the infectious and susceptible nodes are well mixing. For mobile phone virus spreading, the fixed connection between phones and the network similar to the Internet are nonexistent.

According to above analysis, we can find that the model doesn't well treat the following factors:

1) The probability that a mobile phone is infected is relation to the time when it stays around the infected phone. The longer they stay in the place, the higher probability the susceptible phone to be in an infected state. Obviously, in mobile environment, we can't assume that two mobile phone stay the same place always.

2) What effect on virus spread is induced by phone's mobility?

3) Will the topology of phone contacts network impact virus spreading?

In next section, we present our models to deal with these tasks.

3 Models, Analysis and Simulations

Model A: the model of mobile phone virus spreading on random graph

The network is static, namely all nodes don't move in KW model. In mobile environment, we define β as the probability that a susceptible node is infected by the contact with an infectious node in a time step. We define $<k>$ as

$$< k >= \frac{1}{N\Delta T}\sum_j\sum_i T_{ij} \quad (2)$$

where N denotes the sum of nodes, ΔT is a period and T_{ij} is the time when the node j stays around the node i in the period. Our model A has the same form with eq.(1).

Although the node i in mobile environment has more neighbors than those in fixed environment, the connection time T_{ij} is obviously less than the time ΔT. In a fixed region where phones stochastically walk, the density of phones is a constant in both environments.

In fact, the KW model can't well describe epidemic spread in static networks. Especially, when the network has small average degree, the theoretical result remarkably deviates from the results through numerical simulations as depicted Figure 1.

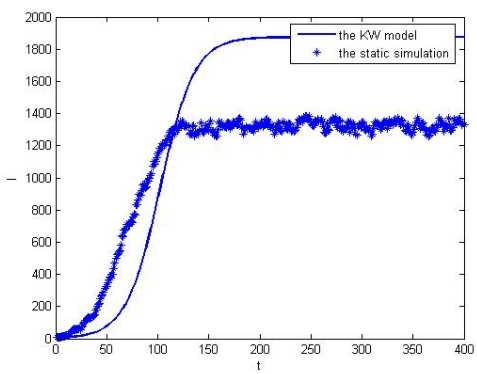


Figure 1 The numerical simulation in static network. The network is a plane grid with 50*50. The parameters $k = 8$, $\beta = 0.2$, and $\delta = 0.2$

Figure 1 shows the compare numerical simulation on static network with theoretical result on the KW model. The static network is a plane grid composed of 2,500 nodes. β = 0.2, δ = 0.2, and four nodes are infectious initially. Obviously, in steady state, the infection density in KW model is larger than that in static network.

Figure 2 shows the compare numerical simulation on the dynamic network with the result in the model A. The parameters β = 0.1 and δ = 0.1. Four nodes are infectious initially.

A large number of simulations show that the model A well follows those results by these simulations. Furthermore, we also find that the epidemic threshold of the model A is consistent with that by simulating.

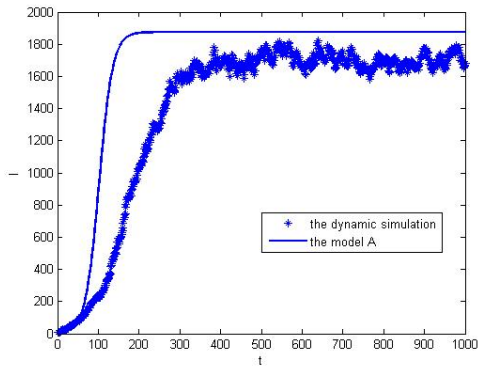


Figure 2 The numerical simulation on dynamic network. The moving space is a square 500*500 m^2. 2,500 phones stochastically move in the region. The parameters $\beta = 0.05$ and $\delta = 0.1$

Model B: the model of mobile phone virus spreading on SF network

Considering the topology of mobile phone contacts network is very essential. In our daily lives, we find that some people, for instance doctors, vendors and students and so on, contact vast amounts of people, on the contrary, most of people contact a few people daily. The social contacts network holds the small world and scale-free properties [13,14]. We can formulate the equation on epidemic spread on the type of network:

$$\dot{I}_k = -\delta I_k + \beta k(1 - I_k)\Theta \quad (3)$$

where I_k denotes the density of infection of the nodes around where k neighbors exist. δ and β are the cure rate and the infection rate, respectively. Θ is the probability that any given link points to an infected node. We also assume that Θ dose not depend on the connectivity of the emanating node and is only a function the total density of infected nodes pointed by the link. Namely,

$$\Theta = \frac{1}{< k >}\sum_k kI_k P(k) \quad (4)$$

where connectivity degree distribution is

$$P(k) \sim k^{-\gamma},\ k = m, m+1, \ldots, K \quad (5)$$

with $2 < \gamma < 3$. By straightforward computations, we get the equilibrium of eq.(3):

$$I_k = \frac{\beta k\Theta}{\delta + \beta k\Theta} \quad (6)$$

This set of equations show the following facts:

The higher the node connectivity (k), the higher probability to be in an infected state.

The higher infected rate, the higher probability to be in an infected state.

Applying eq.(4) and eq.(6), we have self-consistency equation

$$\Theta = \frac{1}{< k >} \sum_{k} \frac{\beta k^2 P(k) \Theta}{\delta + \beta k \Theta} \tag{7}$$

When $\frac{\beta}{\delta} \le \frac{< k >}{< k^2 >}$, eq.(7) has zero solution only. When $\frac{\beta}{\delta} > \frac{< k >}{< k^2 >}$, the unique positive solution exists.

Let's consider the SIS model on the scale-free networks with N nodes. The maximal connectivity K can be expressed by N as

$$K \approx m N^{\frac{1}{\gamma - 1}}. \tag{8}$$

The average connectivity is given by

$$< k > \approx \frac{\gamma - 1}{\gamma - 2} m \tag{9}$$

and the second moment of the connectivity distribution is

$$< k^2 > = \sum_{k=m}^{K} k^2 P(k) \approx \frac{\gamma - 1}{3 - \gamma} m^2 \left(\frac{K}{m}\right)^{3-\gamma}. \tag{10}$$

Combine to eq.(9) and eq.(10), we have

$$\beta_c(N) \approx \frac{(3-\gamma)\delta}{(\gamma - 2) m N^{\frac{3-\gamma}{\gamma - 1}}} \tag{11}$$

The movement of nodes increases the degree of all nodes and decreases the time when the connection between them keeps. So the movement brings about higher degrees and shorter connection time. In our model, we assume the degrees of all nodes on dynamic network are a times as large as that on static network. Notice that $E(\beta k)$ keeps a constant, the connection time between nodes is $1/a$ times as long as the time on static network. Thereby we assume infection rate is changed as β/a. The degree distribution of network in the change follows the power-law distribution $P(k) \sim k^{-\gamma}$, $k = m, m+1, \ldots, K$. Similarly, we can obtain the average connectivity of the mobile network

$$< k >_M \approx \frac{\gamma - 1}{\gamma - 2} am \tag{12}$$

and the second moment of the connectivity distribution

$$\begin{aligned} < k^2 > &= \sum_{k=am}^{aK} k^2 P(k) \\ &\approx \int_{am}^{aK} k^2 P(k) dk \approx \frac{\gamma - 1}{3 - \gamma} (am)^2 \left(\frac{K}{m}\right)^{3-\gamma}. \end{aligned} \tag{13}$$

So we can the epidemic threshold on mobile network

$$\beta_c^M(N) \approx \frac{(3-\gamma)\delta}{(\gamma - 2) am N^{\frac{3-\gamma}{\gamma - 1}}} \tag{14}$$

Comparing eq.(11) and eq.(14), we find the epidemic threshold in mobile environment is less than that in static environment.

The Figure 3 depicts the simulation in dynamic and static states. We find that the infection density in dynamic state is a little more than that in static state.

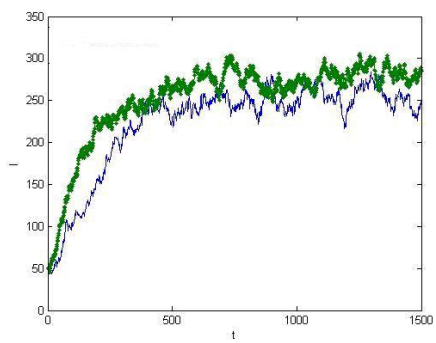


Figure 3 The compare viruses spreading on static and mobile SF network. The upper and bottom curve give the simulation in dynamic and static network, respectively. The network has 2,000 nodes and it's degree distribution follows $P(k) \sim k^{-3}$. Fifty nodes are infected initially. Infected rate $\beta = 0.1$, cure rate $\delta = 0.3$.

4 Conclusions

In this paper, we analyze the limitation of the KW model in mobile network. By giving some parameters new definitions, we present two SIS models, one of which deals with regular lattice and another deals with finite size scale-free network. By comparing the static model with the dynamic model, we show mobility's effect in phone virus spread. In the model A, we find that both the epidemic threshold and the infection

density in steady state are larger than those in static model and less than those in KW model. Furthermore, we also show that the movement of phones decreases the epidemic threshold in the second model.

To farther research the mobile phone virus spreading, we think, it is important to understand the mode of human actions. Researches show that the human contacts network takes on the small-world property besides the scale-free property. So it is necessary to build a model which contains the small-world and scale-free properties even connection weighted.

References

[1] http:/www.umrnet.com, The report on the remark trend of smart phone in 2007 (in Chinese)

[2] Liu Jin, Zhang Liang. Mobile phone virus will spread. *Computer Fans*. Vol. 6, 2007, pp. 15(in Chinese)

[3] Li Zhi, Wang Yanwei, and Zhu Ling. *Today and future of mobile phone virus*. Vol. 3, 2006, pp. 87-90(in Chinese)

[4] W.O. Kermack, A.G. McKendrick, Contributions to the mathematical theory of epidemics, *Proceedings of the Royal Society*, Vol. 115, 1927, pp. 700-721

[5] W.O. Kermack, A.G. McKendrick, Contributions to the mathematical theory of epidemics, *Proceedings of the Royal Society,* Vol. 138, 1932, pp. 55-83

[6] J. O. Kephart, S. R.White, Director-graph epidemiological models of computer viruses, *Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy*, Oakland, CA, 1991

[7] J. O. Kephart, S. R.White, Measuring and modeling computer virus prevalence, *the IEEE Computer Security Symposium on Research in Security and Privacy*, Oakland, CA, USA, 1993

[8] J.W. Mickens, and B.D. Noble, Modeling Epidemic Spreading in Mobile Environments, *Proceeding of the 2005 ACM Workshop Wireless Security*, 2005, pp. 77-86

[9] R. G. Cole, N.Phamdo, M. A. Rajab, A. Terzis, Requirements on Worm Mitigation Technologies in MANETs, *Proceedings of the Workshop on Principles of Advanced and Distributed Simulation* (PADS'05)

[10] Hui Zheng, Dong Li, Zhuo Gao, An Epidemic Model of Mobile Phone Virus. *1st international Symposium on Pervasive Computing and Applications*, Aug. 2006, pp. 1-5

[11] Xia Wei , Li Zhao-hui, Chen Zeng-qiang, Yuan Zhu-zhi，The Influence of Smart Phone's Mobility on Bluetooth. *WiCom 2007, International Conference on Worm Propagation, Wirelss Communications, Network and Mobile Computing*, 2007, pp. 2218-2221

[12] M. Nekovee, Worm Epidemics in Wireless Ad hoc Networks, 2007, arXiv:0707.2293v.

[13] R. Albert, A. L. Barabasi, Statistical mechanics of complex networks. *Review of Modern Physics*. Vol. 74, 2002, pp.47-97

[14] R. Pastor-Satorras, and A. Vespignani, *Physical Review E*, Vol. 63, 2001, pp.6117-6121

The Research and Realization of Clustering Algorithm Based On FPGA

Jun Feng Wenbo Xu Zhilei Chai

College of Information, Jiangnan University, Wuxi, 214122, China
E-mail: fengjun009@hotmail.com

Abstract

Clustering algorithm is an important method of researching neural network. This paper introduces a method to implement this algorithm, under the FPGA, implement K-means clustering algorithm, accelerating the arithmetic by hardware, more rapidness, more effective. The result shows that implement this algorithm take advantage of hardware better than software. K-means clustering algorithm is a kind of clustering algorithm which can be dealt with FPGA, because it is a parallel algorithm, it will be completely get faster by FPGA than it used to be. So that is all the reason why we implement K-means clustering algorithm by FPGA.

KeyWords: FPGA, Neural Network, K-means Clustering Algorithm

1 Introduction

These years, along with rapid progressing of the artificial intelligence and intelligence control, neural network gets more comprehensive application at many domains of the manipulative region. At first the neural network successfully apply to the region of signal processing, including image processing, machine vision, fault diagnosis, target detecting, adaptive filter theory and signal compacting and so on. Successful application at these aspects, it makes the application of the neural network keep expending. So many problems to hard resolve by the normal methods, people are going to get the method to resolve them through the neural network [1].

For researching the neural network, from last century, people get so many clustering algorithms, approximately there are two kinds. Based on compartmentalizing and based on layer, meanwhile the mixed clustering algorithm by synthesizing both of all gets appear. K-means clustering algorithm is the most famous algorithm of the clustering algorithm based on compartmentalizing, this is an algorithm by consecutive iteration to modify the center of cluster through studying, it has high precision, it is not enough rapid. FPGA (Field Programmable Gate Array) got comprehensive application at accelerating software, such as adaptive LMS algorithm [2], so accelerating it by FPGA.

At present, it is so normal to research K-means clustering algorithm, some improve the algorithm, such as at the aspect of the great searching capability, clustering result accuracy etc [3][4], some others research the problem about TSP(traveling salesman problem)、 sensor networks、 the classic genetic algorithm、text data processing[5][6][7][8]. And there are nothing results about the aspect of hardware. This paper introduce K-means clustering algorithm at first, then introduce how to carry it out by FPGA[9], it has six parts to expatiate on the details about the effect after implementing it[10].

2 Algorithm

2.1 The idea of k-means clustering algorithm

On the assumption that $X=\{X_1, X_2, \cdots, X_n\}$ is the set of n object. X_i = (xi, 1, xi, 2, …, xi, m) is an object who has m dimensional variable. K-means clustering algorithm assembles object set X to K clusters

on the process of clustering, to get the target function by least, P is the sum of the distance of every spot of all clusters to the center of clustering.

$$P(U,Z)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}u_{i,l}\ d(x_{i,j},z_{l,j})$$

In it, $\sum_{l=1}^{k}u_{i,l}=1$, $1\le i\le n$, at here:

①U is a n ×k subject matrix, $u_{i,j}$ is a bi-level variable, when $u_{i,j}=1$, it means that object i is distributed in cluster 1, when $u_{i,l}=0$, it means that object1 is not distributed in cluster 1; ②Z = { z_1, z_2, …, z_k } is the set of K more vectors, delegating the center of K more clustering clusters; ③d ($x_{i,j}$, $z_{l,j}$) is the distance or similarity at j dimension of the center of object i and cluster 1. Adopting Euclidean distance to express at here: d ($x_{i,j}$, $z_{l,j}$) $=(x_{i,j}-z_{l,j})^2$ 。

The algorithm process of K-means clustering algorithm:

Inputting condition: clustering number K, and the swatch set include n more data;

Outputting condition: K more clustering meets the least variance.

Disposal flow:

(1) Choosing K more object as the initial center of the clustering from n more data object;

(2)Cycling the following flow ① to ② till every clustering static; ① by the mean of any object of every clustering (the center object), countering the distance of every object and these center object of the swatch set, and compartmentalizing the homologous object by the least distance, videlicet, distributing the object into the proximate clustering with the center; ②re-countering every mean (center object) of the clustering (changing).

3 The Hardware Processing Flow and the Module

3.1 The design idea

Through the algorithm idea above of all, the idea of the paper is:

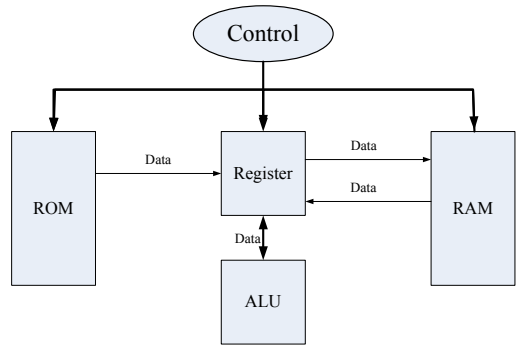


Figure 1 The chart of module

Registering the initial data in ROM, getting one data one time, delivering one port of the comparator, the other port of comparator putting the center of mass array data, start-up comparing, the data compare with every center of mass, getting the closest center of mass, putting in the homologous array, the homologous array in RAM, when all of the data is done, then getting K more cluster of N more data, afterward getting every data one by one deliver to the two ports of adder. the result registers in the Acc, getting the result in one port of divider, putting the data of the counter in the other port of divider, the data of counter is the number of all of the data, registering the result of divider in the register, cycling it K times, then we get the update center of mass array. From updating the clustering, to update the center of mass, until the center of mass doesn't change any more.

3.2 Data access graph

The date access graph as Figure 2.

4 The Design of Control Unit

The control unit design by the way of micro-program, it includes 62 bits, every bit expresses one micro-manipulative control signal, when set it equals to 1, implementing homologous manipulation, when set it equals to 0, no implementing homologous manipulation.

The control process is: at first read the data which is registered in ROM by address register, then put it in the register, then read the data of clustering center array where in the register by the address of address register,

put the data in the subtracter one by one subtract from the data of center arrays, to get the closest center of mass by the data, make them one cluster, put the first data in the register by compared, then make the data subtract from the other center of mass array, registering it in the closest center of mass in the RAM, repeating it by the same way, finally finish it.

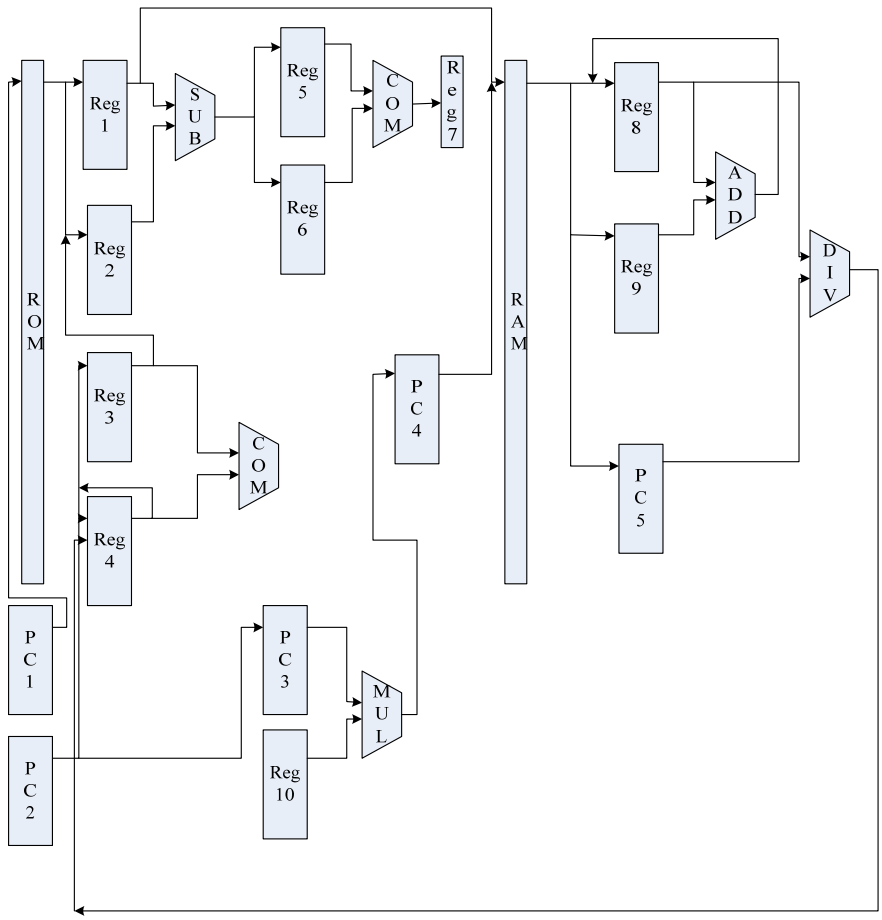


Figure 2 Figure of the data access

After clustering, get every array of RAM, add them each other, getting the result, make a program counter add 1 to itself every time, getting the number of all of the data, division the sum and the number, get the new center of mass array. Updating every center of mass, till it's done.

To sum up: basis on the two big modules, cycle implement it till the data of the center of mass unchanged, then get it.

5 Conclusion and Analysis

Using Xilinx ISE 7.1i program the programmer by the above design, firstly testing algorithm by VC, when input 9 data to get 3 clusters, 9 data: 4、6、8、2、12、14、16、20、10. The result by software:

```
Input data number:  9
Input cluster number:  3
Input 9 data:  4 6 8 2 12 18 16 20 10
------------------------------------ array 0: center of mass (10.000000)  8.00
0000 12.000000 10.000000 array 1: center of mass (4.000000)  4.000000 6.000000 2
.000000 array 2: center of mass (18.000000)  18.000000 16.000000 20.000000
```

Figure 3 The result of the test by software

Compiling program by the Xilinx ISE 7.1i，then testing it by Modsim, we can get the result by the software which is modsim:

Figure 4 The result of the test by hardware

The result we get by the software and hardware, we find out they have the same result, then we use a particular software to get the time which the program implement in software,`program spends 1.5e-008 seconds`, countering the exactly clock which the program need, using the time divide the frequency of computer, we can get the clock, approximate equals to 44 clocks, obviously we used 3 clocks by the hardware, so we know the program use the clock in the hardware less than in the software, if we use the hardware to process a great deal of data, the predominance of the hardware is more visible.

6 The Future Work

This paper actualizes one more clustering algorithm by FPGA. The algorithm has given the optimization disposal. The future work is going to get the algorithm more optimization by FPGA. More data is going to be processed by it, to get more effective utilization.

Reference

[1] WANG Xudong, SHAO Huihe, The Theory of RBF Neural Network And Its Application in control. Information and Control. Vol 26, No. 4 272~284

[2] Hu Zheng-wei, Xie Zhi-yuan, Realization of Adaptive LMS Based on FPGA. Journal of North China Electric Power University. Vol.30, No14 74~77

[3] LIU Jing-ming, HAN Li-chuan, HOU Li-wen, Cluster Analysis Based on Particle Swarm Optimization Algorithm. Systems Engineering-theory& Practice

[4] WANG Yuan-mei, On The Improvement of K-means Clustering Algorithm. Journal of Yangtze University(Nat Sei Edit). Vol. 3, No. 4 76~77

[5] Huang Ying, TSP Evolvement Algorithm Based on K-means Clustering. Journal of Henan Radio & TV UniversityVol.19, No.4 61~62 65

[6] LUO Ying-ying, CHEN Chuan, MAO Yun-fang, Research on k-means clustering arithmetic based on sensor networks. Computer Engineering and Design. Vol. 28, No. 6 1349~1351

[7] WANG Chang, CHEN Zeng-Qiang, YUAN Zhu-Zhi, K-Means Clustering Based on Genetic Algorithm. Computer Science. Vol. 30, No. 2 163~164

[8] YANG Xin-hua, YU Kuan, K-means Clustering Algorithm Based on Self-Adoptively Selected Density Radius. Journal of DaLian JiaoTong University. Vol 28, No.1 41~44

[9] Hao Zhiquan, Wang Zhensong, Liu Bo, Research on Real2Time Real izing PGA Algorithm in FPGA. Journal of Computer Research and Development. 45 (2) : 342~347

[10] L.Charaabi, E.Monmasson, M-A Nassani, I.Slama-Belkhodja, FPGA-based implementation of DTSFC and DTRFC algorithms. 2005 IEEE: 245~250

A Multiobjective Heuristic for ICs Test Suite Reduction

Yue Huang Wenbo Xu

School of Information Technology Southern Yangtze University, Wuxi, 214000, China
Email: huangyuewx@163.com

Abstract

It spends more and more time for testing in the process of designing ICs with the increasing complexity of ICs. Test efficiency is becoming increasingly important. Large numbers of redundant test cases in ICs test suite result in reducing ICs test efficiency greatly. This paper has analyzed Greedy algorithm, GE and GRE proposed by Chen and a heuristic base on the importance of test cases proposed by Harrold which are used in test suite reduction. At the same time a new heuristic (PrioritySelected) for ICs test suite reduction base on the cases 'essentialness' and number of tested ICs faults is presented . An optimal representative suite of an example selected by the heuristic (PrioritySelected) is presented in this paper. It is proved that the heuristic (PrioritySelected) is effectual.

Keywords: heuristic; ICs test; test suite reduction;

1 Introduction

ICs complexity is rapidly increasing with development of IC technology. As a result not only ICs test suite generation is more difficult but also test suite increasing more quick. It spends a great deal of time for testing in the process of designing ICs with the increasing complexity of ICs. Because of including lots of redundant cases, ICs testing time is prolonged greatly. Therefore, ICs test suite reduction becomes more and more important.

ICs test suite reduction algorithms is distributed into two classes. One is static, that is to say reducing test cases on the precondition of keeping the fault coverage which primary test suite finished. Another is dynamic, this mean that a smaller test suite is created by using test suite reduction strategy during test suite created. The algorithm in this paper is static.

the problem of estimating the size of a minimum single stuck-at fault test set for a given irredundant combinational circuit is proven to be NP-hard [1,2], several test suite reduction algorithms based on different heuristics are proposed in the literature[3], e.g. independent and compatible fault sets based test generation [4,5,6,7], reverse order fault simulation [8], maximal compaction [6], rotating backtrace [6], ROTCO [9], high-level test generation [10],double detection [4,11].

2 Preliminaries

Given: $|A|$ is used to denote the cardinality of set A. R Denote all fault suites; $R=\{r0,r1,\ldots,ri,\ldots,r_M\}$ M denote the number of all faults. T denote all test suites, $T=\{t0,t1,\ldots,ti,\ldots,t_N\}$, N denote the number of all test cases. Runselected denote untested fault suite; Selected denote selected test suite. Unselected denote unselected test suite. $S(T,R)$ denote a binary relation from T to R; $S(R,T)=\{(t,r)\in T\times R;\ t\in T\ ;\ r\in R\ ;$ test case t can detect fault r$\}$. Whenever there is no ambiguity, we use S instead of $S(T,R)$; Ri denote a test suite of satisfying ri; Ti denote a fault suite of detected by ti.

There are two different kinds of test cases in a test suite, namely the essential test cases and the redundant test cases. A test case is said to be an essential test case of T if $|Ri|=1$. Contrary to the concept of essential test cases is the concept of redundant test cases. A test case ti is said to be a redundant test case of T if $Ti\text{-}Runselected=\phi$. There is a special kind of redundant test cases known as the l-to-l redundant test

cases. A test case $ti \in T$ is said to be l-to-l redundant if there exists a test case $ti \in T$ such that $Ti=Tj$

3 Summary of Pertinent Algorithms

It was proved that finding an optimal representative suite is NP-complete [1, 2]. For this reason, people have proposed many approximation algorithms. These algorithms mentioned in the following text are all approximation algorithms of an optimal representative suite.

Greedy heuristic [12] selected the test case which could detect maximum faults, then this detected fault was deleted from R, and selected test case was deleted from T. If a tie situation occurs among multiple test cases, an arbitrary choice is made. Since G selects one test case at a time and the test case satisfies at least one unsatisfied requirement, G will loop at most min (m, n) times. Within each loop, the time complexity for selecting the test case that satisfies a maximum number of unsatisfied requirements is at most O (n) so repeatedly, until all faults is deleted from R. The worst case time complexity of the greedy heuristic G is $O(mn.min(m,n))$ [13].

The heuristic GE[14,15] is based on two strategies: the greedy strategy (the strategy of selecting a test case which satisfies the maximum number of not yet satisfied requirements) and, the essentials strategy (the strategy of selecting all essential test cases). The greedy heuristic G is, in fact, the repeated application of the greedy strategy. The essentials strategy should be applied as early as possible because all essential test cases must appear in any representative set. The heuristic GE applies the essential strategy first and then repeatedly applies the greedy strategy to find a representative set. It should be noted that after a test case is selected by the greedy strategy and those satisfied requirements are removed, there will be no new essential test cases for the reduced problem. The worst case time complexity of GE is $O(mn+min(m,n)nk)$ [15].

The heuristic GRE proposed by Chen and Lau [14,15] is based on three strategies: the greedy strategy, the 1-to-1 redundancy strategy (the strategy of removing all 1-to-1 redundant test cases), and the essentials strategy.

When 1-to-1 redundant test cases are removed, the size of the problem decreases. Moreover, an optimal representative set of the reduced problem is also optimal with respect to the original problem. Hence, as many 1-to-1 redundant test cases should be removed as early as possible. Furthermore, some remaining test cases may then become essential after the removal of 1-to-1 redundant test cases [15]. This is because Test(r) may become a singleton set. These essential test cases should then be selected because any representative set with respect to the reduced problem must include them. On the other hand, after the essentials strategy is applied, some test cases may then become 1-to-1 redundant. Hence, the essentials and 1-to-1 redundancy strategies can be alternately applied. The worst case time complexity of GRE is $O(mn+nk + min(m,n>(m + n2k))$ [16].

Harrold L had presented a heuristic base on the importance of test cases [17]. It distributes all test faults $r1,r2,\cdots,r_m$ into $R1,R2,\cdots Rd$. Ri $(i=1,2,\cdots,d)$ represents fault set tested by i test cases. If $i<j$, it considers the test cases satisfy fault set Ri is more important than the test cases satisfy fault set Rj. So the heuristic first selects all test cases satisfy fault set R1, deletes faults set R1 from R. And then it selects test case in R2 which can satisfy a maximum number of unsatisfied faults in R2. Whenever there is a tie situation among several test cases, it then selects the test case which satisfies a maximum number of unsatisfied requirements in R3 amongst those test cases which cause the tie situation in R2 If a tie situation occur again in R3; it selects the test case which satisfies the maximum number of unsatisfied requirements in R4. If a decision cannot be made, it continues on the next group and so on. Eventually, if a decision cannot be made when the process reaches Rd an arbitrary choice is then made. The above process continues until all unsatisfied requirements in R2 are satisfied. The heuristic then considers the group of unsatisfied faults in R3, Rd (one at a time) and selects test cases in a similar manner. The worst case time complexity of H is O (m (m + n).d) [7].

Greedy heuristic give preference to test case which satisfies a maximum number of unsatisfied faults. It doesn't consider the importance of test cases and redundancy. GE and GRE heuristic presented by Chen is still dominated by Greedy heuristic, but they have considered essential test cases and l-to-l redundancy. The heuristic presented by Harrold give preference to importance of test case.

4 A Multiobjective Heuristic for ICs Test Suite Reduction

This paper proposes a heuristic 'PrioritySelected' for ICs test suite reduction which integrates maximum faults strategy and test cases importance strategy. The heuristic is an approximation algorithm of an optimal representative suite. The heuristic 'PrioritySelected' is based on following strategies:

1) Essential test cases are selected using essential strategy proposed by Chen from Unselected. At the same time, faults detected by essential test cases are deleted from Runselected.

2) A test case in Unselected is added 1/i weight if it detects certain fault which this faults detected by i test cases in Runselected. It is added 0 weights if it doesn't detect this fault. Every test case is calculated sum of weight for all faults. If the sum of weight equals 0, the test case is redundant. It is deleted from Unselected.

3) A test case with maximum weight is selected in Unselected randomly. Then faults detected by this test case are deleted from Runselected. The test case is deleted from Unselected.

Repeat 2) ，3) until Runselected is empty.

Algorithm PrioritySelected

INPUT

R: faults suite，$\{r_0, r_1, ..., r_i, ..., r_M\}$;

T: test suite，$\{t_0, t_1, ..., t_i, ..., t_N\}$;

S(T,R): S (R,T) = {(t, r) ∈ T × R:test case t can detect fault r}

Selected:result of test suite reduction;

DECLARE

Ri: a test suite of satisfying ri;

Ti: a fault suite of detected by ti;

Runselected:fault suite of undetection;

Unselected:test suite of unselection;

Tselected:test suite of temporary selection;

Piror:all faults tested number set; $\{pr_0, pr_1, ..., pr_i, ..., pr_M\}$

Pirot:all test cases weight set; $\{pt_0, pt_1, ..., pt_i, ..., pt_N\}$

```
BEGIN
/*initialization*/
 Runselected=R ;  Unselected=T ;
 Selected=ϕ ; Tselected= ϕ ;
For each   ti ∈ T   { Ti= ϕ;pti=0;};
For each  ri ∈ R   {Ri= ∈.pri=0;}  ;
/*search Ri and Ti */
   For each       ti ∈ T
         For each    rj ∈ R
              If  S(ti, rj)   then
  {ti+ {Rj};Rj+ {ti};}
/*select essential test cases*/
For each  ri ∈ R
    If  |Ri|=1=1 then
 Tselected= Tselected+ Ri;
 Selected= Tselected ;
 Runselected= Runselected- {∪ Ti, ti ∈ Tselected};
 Unselected= Unselected- Tselected;
Loop
/*calculate test cases weight*/
For each  ti ∈ U  unselected
{ Pti=∑(1/prj, rj ∈ Ti)  ;
If pti=0      then
   Unselected= Unselected-{ti}  ;
   }
/*select a test case with maximum weight */
 Unselected= Runselected-{Ti, pti=
 max {pt0, pt1,..., ptI,..., ptN}}
Until Runselected=ϕ ;
END
```

5 Example

There is a minimum test suite {t1, t2, t3} or {t1, t2, t8} in following example. We illustrate

heuristic PrioritySelected with this example.

Suppose test suite T= {t2, t2, t3, t4, t5, t6, t7, t8} , fault suite R= {r1, r2, r3, r4, r5, r6, r7, r8, r9, r10} . Satisfiability relation S(T,R) is given in Table 1.

Table 1 The satisfiability relation S(T,R))

	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	Pirot
t1	1	0	1	0	1	0	1	0	1	0	2
t2	0	1	0	1	0	1	1	0	0	0	1.416667
t3	1	0	1	0	0	1	0	1	0	1	1.5
t4	0	1	0	1	0	0	0	0	1	0	1.333333
t5	0	0	0	0	1	1	0	1	0	0	1.083333
t6	1	0	0	0	0	0	1	0	0	1	0.916667
t7	0	1	0	0	1	0	0	0	0	1	1.083333
t8	0	0	1	0	0	1	0	1	0	1	1.166667
Piror	3	3	3	2	2	4	3	2	2	4	

Initialization:

Runselected={r1,r2,r3,r4,r,r6,r,r8,rr9,r10};

Unselected= {t1, t2, t3, t4, t5, t6, t7, t8} ;

Selected=ϕ ;

Because components value in Piror isn't '1', there are none essential test cases in this example.

First loop:

As show in the table 1, pt1=1/pr1+1/pr3+1/pr5+ 1/pr7+1/pr9=1/3+1/3+1/2+1/3+1/2=2.

Other components in Pirot can be calculated analogously. According value of components in Pirot, t1 is selected.

Selected= {t1};

Unselected= {t2, t3, t4, t5, t6, t7, t8};

Runselected= {r2, r4, r6, r8, r10};

Second loop:

As show in the table 2, Pirot is recalculated, and redundancy is deleted if pti equal '0'.

Table 2 first recalculate Pirot and delete redundancy

	r2	r4	r6	r8	r10	Pirot
t2	1	1	1	0	0	1.08333333
t3	0	0	1	1	1	0.83333333
t4	1	1	0	0	0	0.83333333
t5	0	0	1	1	0	0.58333333
t6	0	0	0	0	1	0.25
t7	1	0	0	0	1	0.58333333
t8	0	0	1	1	1	0.83333333
Piror	3	2	4	3	4	

According value of components in Pirot t2 is selected. Selected= {t1, t2};

Unselected = {t3, t4, t5, t6, t7, t8} ;

Runselected= {r8, r10};

Third loop

As show in the table 3, Piror is recalculated, and redundancy is deleted if pti equal '0' again.

Table 3 Second recalculate Pirot and delete redundancy

	r8	r10	Pirot
t3	1	1	0.58333333
t5	1	0	0.33333333
t6	0	1	0.25
t7	0	1	0.25
t8	1	1	0.58333333
Piror	3	4	

According value of components in Pirot t3 is selected. Selected= {t1, t2, t3};

Unselected=t {t5, t6, t7, t8} ;

Runselected=Runselected= {}.

Because Runselected is empty, loop is broken. Result of test suite reduction is {t1, t2, t3} .

6 Conclusion

The heuristic 'PrioritySelected' considers not only the number of tested faults but also the importance of test cases when it selects test cases. Therefore its selection strategy is more reasonable than Greedy heuristic and the heuristic presented by Harrold. All redundancy is deleted by calculating test cases sum of weight in Heuristic 'PrioritySelected'. It is more effective and simpler than the GE and GRE presented by Chen which could barely deleted 1-1 redundancy.

References

[1] B. Krishnamurthy and S. B. Akers, "On the complexity of estimating the size of a test set," IEEE Trans. Computers, vol. C-33, no. 8, Aug. 1984, pp. 750–753

[2] M.J. Harrold, R. Gupta, M.L. Soffa, "A methodology for controllingthe size of a test suite," ACM Transactions on Software Engineeringand Methodology 1993,2 (3): 270-285

[3] Ilker Hamzaoglu and Janak H. Patel, "Test Set Compaction

Algorithms for Combinational Circuits," IEEE Trans. Computer-Aided Design, vol.19, no.8, August 2000, pp. 957-963

[4] S. B. Akers, C. Joseph, and B. Krishnamurthy, "On the role of independent fault sets in the generation of minimal test sets," in Proc. Int. Test Conf., Aug. 1987, pp. 1100–1107

[5] K. Kinoshita, and S.M. Reddy, "Cost effective generation of minimal test sets for stuck-at faults in combinational logic circuits," IEEE Trans. Computer-Aided Design, vol. 14, Dec. 1995, pp. 1496–1504

[6] I. Pomeranz, L. Reddy, and S. M. Reddy, "Compactest: A method to generate compact test sets for combinational circuits," in Proc. Int. Test Conf., Oct. 1991, pp. 194–203

[7] G.-J. Tromp, "Minimal test sets for combinational circuits," in Proc. Int. Test Conf., Oct. 1991, pp. 204–209

[8] M. H. Schulz, E. Trischler, and T. M. Sarfert, "SOCRATES: A highly efficient automatic test pattern generation system," IEEE Trans. Computer-Aided Design, vol. 7, Jan. 1988, pp. 126–137

[9] L. N. Reddy, I. Pomeranz, and S. M. Reddy, "ROTCO: A reverse order test compaction technique," in Proc. IEEE EURO-ASIC Conf., Sept. 1992, pp. 189–194

[10] M. C. Hansen and J. P. Hayes, "High-level test generation using physically- induced faults," in Proc. IEEE VLSI Test Symp., Apr. 1995, pp. 20–28

[11] S. Kajihara, I. Pomeranz, K. Kinoshita, and S. M. Reddy, "Cost effective generation of minimal test sets for stuck-at faults in combinational logic circuits," in Proc. Design Automation Conf., June 1993, pp. 102–106

[12] V. Chvatal," A Greedy Heuristic for the Set-Covering Problem， Mathematics of Operations Research. 4(3), August 1979, pp.233-235

[13] T. H. Cormen, C. E. Leiserson, R. L. Rivest, "Introduction to Algorithms," MIT Press, Cambridge, MA, 1990

[14] Chen T Y, Lau M F．"A new heuristic for test suite reduction," Information and Software Technology, vol. 40, No. 5, 1998, pp. 347--354

[15] Chen T Y，Lau M F，"A simulation study on some heuristics for test suite reduction," Information and Software Technology, vol. 40, No. 13, 1998, pp 777-787

[16] Chen T Y, Lau M F, "Dividing strategies for the optimization of a test suite," Information Processing Letters , vol.60, no.3, Nov.1996, pp.135-141

[17] Harreld M J, Gupta R, Sofia M L , "A methodology for controlling the size of a test suite," ACM Transactions on Software Engineering and Methodology, vol. 2, no. 3, July 1993, pp. 270-285

Java For Embedded Real-time Systems

Yuan Shen Wenbo Xu

Dept of Computer, School of Information Technology, Jiangnan University, Wuxi 214122, China
Email: shenyuan82@yahoo.com.cn

Abstract

Though Java was first used in an embedded system, due to some technical reasons Java language is not widely applied to embedded systems. To change the situation, Sun defines some new Java specifications for embedded application development. With the Real-Time Specification for Java (RTSJ) proposed, more and more programmers begin to devote to real-time Java applications. This paper gives an overview of Java-based embedded real-time systems and analyzes future work about Java-based embedded real-time technology.

Keywords: Real-time Java, Embedded Java, Embedded real-time systems, Real-Time Specification for Java (RTSJ)

1 Introduction

Embedded Systems are designed to perform a few specific tasks in the most efficient way. Most embedded systems are time critical applications meaning that the embedded system is working in an environment where timing is very important, So embedded systems are also known as real-time systems which are respond an input or event and produce the result within a guaranteed time period.

Java is a modern object-oriented programming language. It is used in many different areas of software development. However, due to the lack of acceptable real-time performance use of the Java language in real-time systems isn't widespread. To point to the need for a common, high-level, fully-supported, correct, advanced, Java-based, real-time application development platform [1], Greg Bollella et al propose the Real-Time Specification for Java (RTSJ) [2]. As thus, many implementations, research papers and real-time or embedded projects using the RTSJ.

This paper will present the status of embedded Java and Real-time Java, and then introduce some Java platforms for embedded real-time system. At last this paper will discuss future work about embedded real-time systems.

2 Java for Embedded Systems

Java is not just an object-oriented programming language. It's a complete dynamic platform that extra infrastructure to run on embedded devices. The primary features of Java that make it attractive for embedded systems are platform independence, automatic memory management and safe pointers. However, Java is not suitable for embedded systems naturally due to the strict requirements of embedded systems.

1) Embedded systems requirements:
2) Demanding Robustness
3) Tight Integration with the Environment
4) Limited Functionality Kernels
5) Limited Resources (CPU, memory, etc)

To overcome the technical difficulties, some Java specifications for embedded systems were released.

2.1 Java standard edition (java SE) for embedded

Java SE for Embedded [3] is part of Java SE. It can not only support the same platforms and functionality as Java SE, but also provides specific features and support for the higher-end embedded application. These embedded-specific features and support currently include additional platforms, small footprint Java

Runtime Environment (JRE), headless configurations and memory optimizations.

2.2 Java platform, micro edition (java ME)

Java ME [4] was designed by Sun Microsystems to provide specific support for small, resource-constrained devices such as cell phones, personal digital assistants (PDAs) and set-top boxes.

Java ME platform technology has three components:

1) A Configuration is a combination of a Java virtual machine (JVM) and a set of application support application programming interfaces (APIs) that are shared across a class of devices.

2) A Profile is a set of APIs (designed for a specific configuration) that address the needs of a narrower device category.

3) An Optional Package is a set of technology-specific APIs that extends the capabilities of a Java application environment.

There are currently two configurations: the Connected Limited Device Configuration (CLDC) [5] and the Connected Device Configuration (CDC) [6].

The CLDC: is a specification of a framework for J2ME applications targeted at devices with very limited resources such as cellular phones, pagers, low-end personal organizers, and machine-to-machine equipment. The requirements of devices that support CLDC include at least 192 KB of total memory and a 16-bit or 32-bit processor. The Java Specification Requests (JSRs) for CLDC include CLDC Version 1.0 (JSR 30) and CLDC Version 1.1 (JSR 139).Some enhancements of CLDC 1.1 is the support for floating points and weak references.

The CDC: is a specification of a framework for J2ME applications targeted at on embedded devices such as smart communicators, high-end PDAs, and set-top boxes. Devices that support CDC typically include a 32-bit microprocessor/controller and make about 2 MB of RAM and 2.5 MB of ROM available to the Java application environment. The JSRs for CDC include CDC Version 1.0 (JSR 36) and CDC Version 1.1 (JSR 218).

3 Java for Real-Time Systems

This section will give an overview of real-time Java.

3.1 Real-Time systems

Generally Real-Time systems can be divided into two categories according to their time constraints:

Hard real-time: Hard real-time systems must satisfy the deadlines on every occasion. In these systems, the completion of an operation after its deadline is considered useless. Most embedded systems are hard real-time systems.

Soft real-time: Soft real-time systems can tolerate a very short delay, but service quality may be decreased.

A hard real-time system obviously has much more constraint than a soft real-time system on the performance of the system.

3.2 Difficulty in real-time java applications

Standard Java applications running on a general-purpose JVM on a general-purpose operating system can only hope to meet soft Real-Time requirements at the level of hundreds of milliseconds [7]. The following problem issues of Java language specification make Java difficult to meet hard Real-Time requirements:

Thread scheduling: The JVM relies on the scheduler that host operating systems or specific hardware support, but not all such schedulers are capable of real-time scheduling.

Garbage collection: Dynamic memory management is convenient for common Java application designers, but for real-time Java application garbage collection is a intractable problem because traditional garbage collectors can interrupt the execution of applications for unpredictable intervals of time.

Class loading: Usually the Java Class loader is a part

of the JRE that dynamically loads Java classes into the JVM. Classes are only loaded on demand. The dynamic class loading process is fairly complicated, including loading, linking and initializing classes and interfaces. This loading does not occur until the class is actually used by the program. When Java program runs, this loading may occur repeatedly. Such non-deterministic behavior of class loading may cause an unexpected delay. So the dynamic class loading is not suitable for hard real-time systems, it's may be suitable for some soft real-time systems.

Compilation: Most modern JVMs initially interpret Java bytecode and then only compile the bytecode executed frequently to native code, the other still remains as interpreted bytecode. This compilation results in fast start-up and reduces the amount of compilation performed when a program is running. But this compilation makes it impossible to predict when the compilation will occur. For hard real-time systems, such nondeterminism can't be tolerated.

3.3 The real-time specification for java

To support both hard and soft real-time Java applications, Greg Bollella et al proposed the Real-Time Specification for Java. The RTSJ is targeted toward real-time systems by providing real-time capabilities. It enhances Java in the following areas:

Memory Management: Java garbage collection has always been a problem to real-time programming due to its unpredictability. RTSJ defines immortal and scoped memory areas to supplement the standard Java heap memory. Objects allocated in scoped memory have a well-defined life time. Immortal memory is shared among all threads. Objects created in immortal memory are freed only when the program terminates. Scoped memory and immortal memory are areas of memory which are logically outside of the heap and, therefore, are not subject to garbage collection. To avoid dangling reference, strict assignment rules shown in Table 1 must be checked by the implementation.

Table 1 Memory Assignment Rules

From memory area	To Heap Memory	To Immortal Memory	To Scoped Memory
Heap Memory	allowed	allowed	forbidden
Immortal Memory	allowed	allowed	forbidden
Scoped Memory	allowed	allowed	allowed for same scope or outer scope
Local Variable	allowed	allowed	generally allowed

Thread Scheduling: RTSJ introduces the concept of a schedulable object, adding the RealtimeThread class for thread management. A schedulable object is any object which implements the Schedulable interface. The default scheduler is a preemptive, priority-based scheduler with 28 priorities.

Synchronization: The RTSJ includes priority-inheritance support to manage synchronization when it occurs. Threads waiting to enter a synchronized block are priority and FIFO within priority ordered.

Asynchronous Transfer of Control (ATC): The RTSJ's approach to ATC is based on the class Asynchronously Interrupted Exception (AIE), a subclass of the checked exception class Interrupted Exception. Asynchronously Interruptible (AI) and ATC-deferred section are another two terms used. AI method is asynchronously interruptible if it includes AIE in its throws clause. ATC-deferred section is a synchronized statement, a synchronized statement method, or any another method and constructor which lacks a throws AIE clause.

4 Java Platform for Embedded Real-time System

Java platform is comprised of JVM and Java API. Though every JVM must be capable of executing Java bytecode, but the way of execution can be chosen. The specification of JVM is flexible. It allows that JVM can be implemented in pure software's way or most part of JVM is implemented in hardware's way. Different JVMs have different execution engines. The following execution engines are in the JVM implemented in

software's way:

Interpreter: Interpreter is the commonest execution engine. It needs the support of operation system. The defect of Interpreter is low efficiency. Currently several Real-Java platforms including RJVM [8-10], OVM [11-13], JTIME [14], RTSJ RI [14] and Mackinac [15] use Interpreter.

Just-In-Time (JIT) Compiler: JIT compiler converts bytecode at runtime into native machine code prior to executing it. If the executable bytecode is repeatedly reused, the time's gap between compile and execution will be shortened. JIT compiler may consume more memory and cause a slight delay in initial of an application due to the time taken to compile the bytecode. Therefore JIT compiler is not suitable for embedded real-time system.

Offline Bytecode Compiler: There are two types of offline bytecode compilers: native and non-native. Native compilers produce code that is directly executable, while non-native ones produce code in an intermediate language. Native compiler is faster and more efficient, but it is not portable and generation of efficient code requires an extended knowledge of the features of the target processor. Non-native compiler is more flexible and has competitive performance. The JVM using offline bytecode compiler can be applied to embedded real-time system, for example [16].

Besides the above-mentioned techniques, execution engine can be supported by hardware. Java processor is a dedicated processor which implements the JVM and directly executes the Java bytecode. The advantages of Java processor are low power, high performance, less memory consumption, so it is more suitable for embedded system. Though such processors are not widely available, the application in embedded Real-Time system of Java processor has good prospects for the future. Currently there are a few Java processors for embedded real-time systems such as JOP [17-20], Komodo [21, 22], aJ-100 [23] and FemtoJava [24, 25].

5 Conclusion and Future Work

There are some restrictions of Java for embedded real-time systems, but many organizations and scholars are devoting to mitigating these restrictions. Different implementations and papers about embedded Java and real-time Java are proposed. To promote the development of the domain further, the future work can focus on the following aspects.

Hardware support for embedded real-time system: According to Section 4, at the present time most embedded real-time systems are based on the Java platform implemented in software's way. For embedded systems the limits of resource, cost, memory and power consumption are fairly strict. Java processors are easier to meet the requirements. Besides, Java processors can not only avoid the overhead of translation of the bytecode to another processor's native language, but also provide support for some mechanisms of the RTSJ in hardware. Implementing RTSJ-compliant processors for embedded real-time system is worth while to research.

Class loading for embedded real-time system: As section 3.2 analyzes, dynamic class loading is more suitable for some soft real-time systems such as PDAs and mobile phones. An option of the class loading for hard real-time system is that all classes should be loaded, linked and initialized before the Java program runs. As thus, the influence of dynamic loading on real-time performance can be eliminated. This class loading is suitable for deep embedded systems.

Garbage collector for embedded real-time system: Though RTSJ defines scoped memory, immortal memory and NoHeapRealtimeThread, but there's also innovation in the use of real-time garbage collection. And there are two reasons that heap memory should not be discarded by programmers:

1) Immortal memory is never collected, so it is a limited resource that must be used carefully. If immortal memory is abused, it may result in memory leak.

2) It is not always easy to use scoped memory, for example sharing scoped memory between multiple threads is difficulty due to the single-parent rule for scoped memory [26].

To implement garbage collectors for embedded real-time systems, the following options can be considered:

1) Support Real-time garbage collection in hardware.

2) Improve algorithmic and compression techniques to eliminate overhead as much as possible.

3) Determine an application allocation rate in advance. Tobias Mann et al present a static analysis to bound conservatively an application's allocation rate. [27] This kind of analysis is done at compile time and is crucial to the correct operation of a real-time collector.

References

[1] G. Bollella, "The RTSJ and the SUN JAVA REAL-TIME SYSTEM," http://fi.sun.com/sunnews/events/2006/technical_breakfasts/presentations/HelsinkiWorkshopMay2006.pdf

[2] G. Bollella, J. Gosling, B. Brosgol, P. Dibble, S. Furr, D. Hardin, and M. Trunbull, "The Real-Time Specification for Java," Addison Wesley, vol. 1st edition, 2000

[3] http://java.sun.com/javase/embedded

[4] http://java.sun.com/javame/reference/apis.jsp

[5] http://java.sun.com/products/cldc/overview.html

[6] http://java.sun.com/products/cdc/overview.html

[7] M. Stoodley, M. Fulton, M. Dawson, R. Sciampacone, and J. Kacur, "Real-time Java, Part 1: Using the Java language for real-time systems," http://www.ibm.com/developerworks/java/library/j-rtj1/index.htm

[8] A. J. Wellings, G. Bollella, P. Dibble, and D. Holmes, "Cost Enforcement and Deadline Monitoring in The Real-Time Specification for Java," presented at Proceedings of the 7th IEEE International Symposium on Object-Oriented Real-Time Distributed Computing ISORC-2004, 2004

[9] H. Cai and A. Wellings, "Towards a high integrity real-time Java virtual machine," ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS 2003: OTM 2003 WORKSHOPS, vol. 2889, pp. 319-334, 2003

[10] H. Cai and A. Wellings, "Supporting mixed criticality applications in a ravenscar-java environment," ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS 2004: OTM 2004 WORKSHOPS, PROCEEDINGS, vol. 3292, pp. 278-291, 2004

[11] J. Manson, J. Baker, T. Cunei, S. Jagannathan, M. Prochazka, B. Xin, and J. Vitek, "Preemptible Atomic Regions for Real-time Java," presented at Proceedings of the 26th IEEE International Real-Time Systems Symposium (RTSS05), 2005

[12] http://www.cs.purdue.edu/homes/baker29/ovm/Overview.html

[13] Jason Baker, Antonio Cunei, Chapman Flack, Filip Pizlo, Marek Prochazka, Jan Vitek, Austin Armbuster, Edward Pla, and D. Holmes, "Real-time Java in Avionics Applications," presented at Proceedings of the 12th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), 2006

[14] http://www.timesys.com

[15] G. Bollella, B. Delsart, R. Guider, C. Lizzi, and F. Parain, "Mackinac: making HotSpot/spl trade/ real-time," presented at Eighth IEEE International Symposium on Object-Oriented Real-Time Distributed Computing, 2005. ISORC 2005., 2005

[16] A. Nilsson and S. G. Robertz, "On real-time performance of ahead-of-time compiled Java," presented at Eighth IEEE International Symposium on Object-Oriented Real-Time Distributed Computing, 2005. ISORC 2005., 2005

[17] M. Schoeberl, "JOP: A Java optimized processor," ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS 2003: OTM 2003 WORKSHOPS, vol. 2889, pp. 346-359, 2003

[18] M. Schoeberl., "JOP: A Java Optimized Processor for Embedded Real-Time Systems," Phd dissertation, http://www.jopdesign.com, 2005

[19] M. Schoeberl, "Restricitons of Java for Embedded Real-Time Systems," presented at Proceedings of the 7th IEEE International Symposium on Object-Oriented Real-Time Distributed Computing,Austria, Vienna, May 2004.

[20] M. Schoeberl, "Java Technology in an FPGA," presented at Proceedings of the International Conference on Field-Programmable Logic and its applications, Antwerp, Belgium, August 2004

[21] U. Brinkschulte, C. Krakowski, J. Kreuzinger, R. Marston, and T. Ungerer., "The Komodo Project: Thread-Based Event Handling Supported by a Multithreaded Java Microcontroller," presented at 25th EUROMICRO Conference, Milano, 1999

[22] U. Brinkschulte, C. Krakowski, J. Kreuzinger, and T. Ungerer, "A multithreaded Java microcontroller for thread-oriented real-time event-handling," presented at International Conference on Parallel Architectures and Compilation Techniques, 1999

[23] http://www.ajile.com/downloads/aj100.pdf

[24] S. A. Ito, L. Carro, and R. P. Jacobi, "Making Java work for microcontroller applications," Design & Test of Computers, IEEE, vol. 18, pp. 100 - 110, 2001

[25] M. Wehrmeister, L. Becker, and C. Pereira, "Optimizing real-time embedded systems development using a RTSJ-based API," ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS 2004: OTM 2004 WORKSHOPS, PROCEEDINGS, vol. 3292, pp. 292-302, 2004

[26] C. Gough, A. Hall, H. Masters, and A. Stevens, "Real-time Java, Part 5: Writing and deploying real-time Java applications," http://www.ibm.com/developerworks/java/library/j-rtj5/index.html?S_TACT=105AGX02&S_CMP=ART

[27] T. Mann, M. Deters, R. LeGrand, and R. K. Cytron, “Static Determination of Allocation Rates to Support Real-Time Garbage Collection,” http://www.cs.wustl.edu/~mdeters/doc/papers/static_determination_of_alloc_rates.pdf

Computed of Bridge-Type NEMS Series Contact Switch

Guozhu He[1] Jiankang Liu[1*] Lan Di[2]

1 Sichuan Agricultural University, Dujiangyan, 611830, China
Email: hguozhu@scfc.edu ljiankang@scfc.edu.

2 School of Information Technology, Jiangnan University, Wuxi, 214122, China
Email: Dilan@company.com

Abstract

A dielectric-bridge-type MEMS series contact switch is designed, fabricated and measured, which has a moveable membrane with the upper gold electrodes, as well as the contact metal bar, underneath the SiON dielectric bridge. This structure results in electrically isolating bias from the RF signal, simplifying the fabrication process, and decreasing the pull-in voltage to a certain extent. The measured data show the pull-in voltage of 23.3V and the good RF performance of the insertion loss of beyond -1.2dB and the isolation of below -53dB at 0-10GHz, indicating that the witch is suitable for the 0-10GHz applications.

Keywords: Bridge-Type, contact Switch, metal bar, gold electrodes, RF performance

1 Introduction

As the key device of radio frequency or high frequency transmission systems, RF-switch is widely used in civil and military applications of RF, microwave and millimeter-wave circuits and systems. Previously, RF-switching is implemented by using p-i-n diodes and GaAs MESFETs in the form of junction field-effect transistor (JFET)-based semiconductor switches [1][2]. Recently, RF micro-electro-mechanical system (MEMS) switches have received significant attention over the traditional ones, due to their low insertion loss, high isolation, and low power consumption.

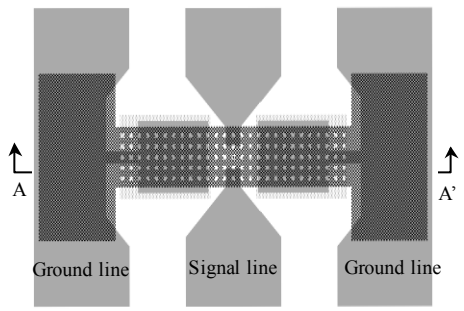
MEMS switches are mainly categorized as contact and capacitive switches. In contrast with the capacitive one, such as the switch reported by Xiao-Feng Lei [3], the contact switch has a wider band down to the DC frequency. Lei L. Mercado, etc. reported a SiON cantilever series MEMS switch [4], but the cantilever structure is too sensitive to the stress of the dielectrics, adding to the fabrication difficulties. The University of Michigan developed an all-metal broadside-series switch [5]. However, this structure does not avoid RF signal isolation, which results in rapidly degrading performance on high frequency band. Another dielectric membrane broadside-series switch [6] has solved the problem of signal isolation. But the upper electrodes and contact metal are on top of and beneath the membrane respectively, which increasing the fabrication complexities.

This paper presents a dielectric-bridge-type MEMS switch, including contact metal bar and upper electrodes both underneath the SiON bridge. It can solve all the problems described above, such as applications in low frequency, isolation of the RF signal from bias, insensitivity to the membrane stress and simplification of the fabrication process. Additionally, the closer distance between the upper and lower electrodes reduces the pull-in voltage. According to the measured data, the designed and fabricated switch exhibits low insertion loss (>-1.2dB@0-10GHz), high isolation (<-53dB @0-10GHz), and low pull-in voltage (Vp=23.3V).

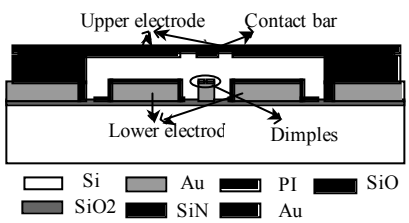
* Authors for correspondence: Jian-Kang Liu

2 Structure Design

In this section, the main structure of the designed switch is described. The basis is a metal contact, dielectric bridge series switch based on the CPW transmission line designed to present the 50 Ω characteristic impedance, as shown in Figure 1. The whole switch is fabricated on the silicon substrate with a silicon oxide layer on top of it. The central part of the signal line is separated to make a gap, and is narrowed to get better isolation performance on the "off" state. A SiON bridge of 500um long, 100um wide and 0.6um thick is suspended across the gap, and is fixed over the ground lines via a polyimide layer on the both sides. Arrays of holes, as shown in Figuie 1(a), are made in the bridge to help releasing the sacrificial layer easily. A gold bar clings to the bottom surface of the bridge center, making two overlapped areas together with the separated signal line. Several dimples are fabricated on the overlapped areas of the signal line, to make a good metal-metal contact between the signal line and the gold bar, as shown in Figure 1(b).



(a) Top view of the switch



(b) Cross view of the switch along A-A' derection

Figure 1 Main structure of the switch

In particular, the two side upper electrodes also cling to the bottom surface of the bridge, and are positioned on the same plane of the contact bar. The RF signal can be isolated due to the separation of the bar and the upper electrodes, and the fabrication process can be simplified due to the same plane of them. The two side lower electrodes are connected to the ground lines, which are DC voltage insulated by the dielectric polyimide layer. When the DC voltage is applied between the upper and lower electrodes, the bridge is pulled down toward the substrate to obtain a metal-metal contact between the contact bar and the signal line, resulting in the switch "on" state; otherwise, the switch keeps the "off" state.

3 Fabrication

The main fabrication process of the switch is shown in Figure 2. The process is started with a high-resistivity 4-inch silicon substrate, including a 0.8um thick silicon oxide layer on top of it as the buffer layer by oxidation process. The silicon oxide layer is patterned and etched by BHF solutions on some region to make a series of 0.3um high dimples, as shown in Figure 2(a). After that, a gold seed layer is sputtered, followed by electroplating using the photoresist AZ4620 as the plating mould, to form the 2um thick gold CPW line and the lower electrodes. Then, the excessive seed layer is removed by wet-etching process, and the silicon nitride layer is deposited on the lower electrodes by PECVD process, as shown in Figure 2(b). At the same time, the dimples on the surface of the signal line are obtained naturally. In the next step, a polyimide layer is spun as the sacrificial layer, part of which as the isolation layer between the upper and lower electrodes. Then, a gold layer is sputtered and a SiON layer is deposited using PECVD apparatus. The two layers are patterned and etched to form the shapes of contact bar, electrodes and bridge, as shown in Figure 2(c). It should be noted that arrays of holes are etched in the bridge to sufficiently release the sacrificial layer. Finally, the

sacrificial layer is released in the atmosphere of oxygen plasma to complete the whole fabrication process, as shown in Figure 2(d).

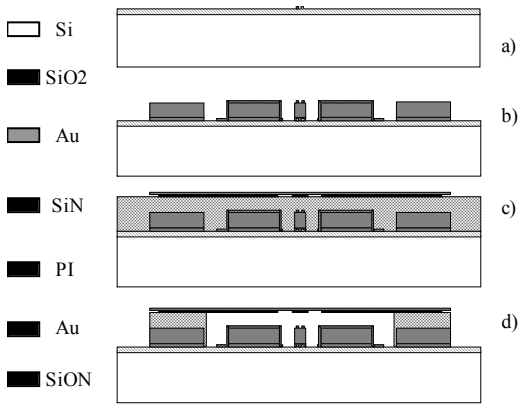


Figure 2 Fabrication process of the switch

4 Results and Discussion

Figure 3 exhibits the scanning electron microscope (SEM) photograph of the achieved switch. Seen from the photograph, the SiON bridge is suspended over the gap of the two ground lines, together with the upper electrodes and contact bar underneath the bridge. The initial vertical distance between the signal line and contact bar is 1.8um.

After destroying the bridge, the four dimples on each signal line end are also observed, as shown in Figure4, to ensure a good metal-metal contact and avoid the "stick" problem between the signal line and contact bar. The contact resistance between the contact bar and signal line on the "on" state is 0.8-1.4Ω, which. achieved a good level according to the previous works, profit from fabricating a series of dimples.

By testing accompanying samples, the low tensile residual stress 76.8MPa of the PECVD SiON membrane is obtained. The low stress process ensures the device reliability and reduces the pull-in voltage. Calculated by the equation provided by Jei Cai [7], the pull-in voltage is 28V. The measured voltage is 23.3V, 4.7V below the calculated data. The diverseness of the results is mainly because of the module simplification, such as regarding the mixed two layer bridge as a uniform SiON layer. The cripples of the bridge due to the discontinuity of the lower gold layer, as shown in Figure 3, can also reduce the bridge stiffness, and reduce the pull-in voltage.

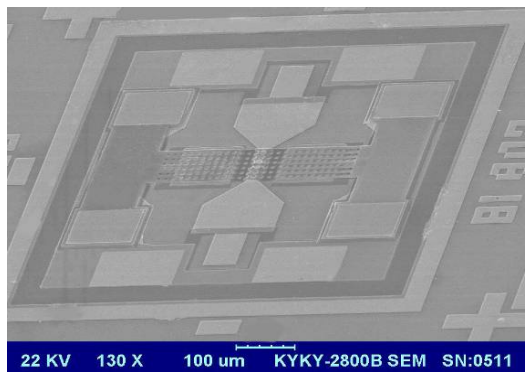


Figure 3 SEM photograph of the switch

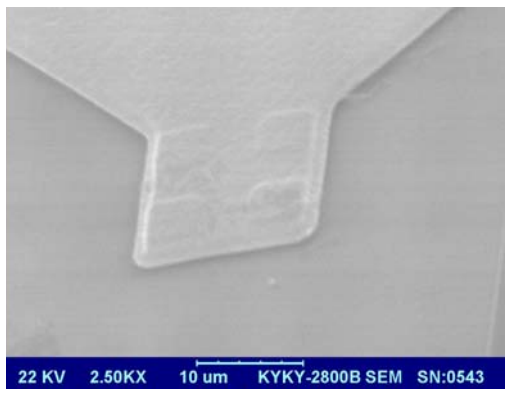


Figure 4 SEM photograph of the dimples

Additionally, if the upper electrodes are on the top of the bridge, which increases the initial vertical distance between the upper and lower electrodes, the pull-in voltage increases to 29.8V, according to the equation described above. So, the structure of the upper electrodes underneath the bridge can reduce the pull-in voltage to a certain extent.

Figure 5 and Figure 6 show the RF performance of the switch tested on the HP8722ES Network Analyzer and the SUSS PM5 Probe Station. The isolation on the "off" state is below -37dB at 0-40GHz, and -53dB at 10GHz, as shown in Figure 5. The insertion loss on the "on" state is beyond -12dB at 0-40GHz, and -1.2dB at 10GHz, as shown in Figure 6. In accordance with the measured results, the designed and achieved switch is suitable for the 0-10GHz frequency applications, such as

wireless personal communication systems, wireless local area networks, etc.

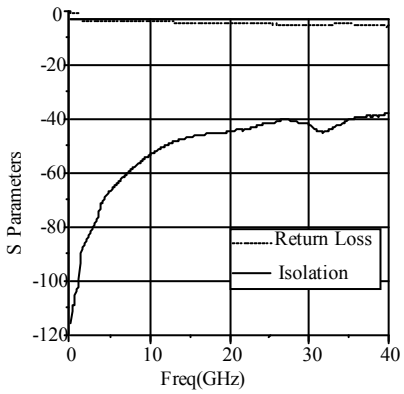


Figure 5 Measured RF performance on the "off"

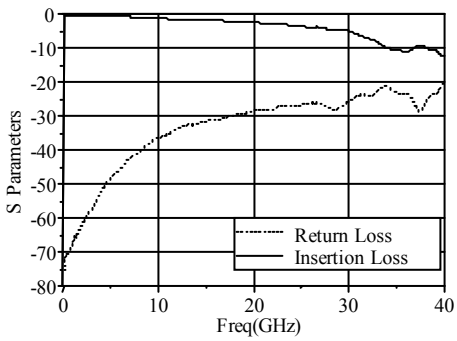


Figure 6. Measured RF performance on the "on" state

5 Conclusion

In this works, a dielectric-bridge-type MEMS series contact switch is designed and achieved by the MEMS fabrication process. With the structure of both the upper electrodes and contact bar fabricated underneath the SiON dielectric bridge, the switch can isolate the RF signal from bias, overcome the sensitivity to the dielectric stress, reduce the pull-in voltage and simplified the fabrication process. The measured results show the contact resistance of 0.8-1.4Ω, the pull-in voltage of 23.3V, the isolation of below -53dB at 0-10GHz on the "off" state, and the insertion loss of beyond -1.2dB at 0-10GHz. It is suitable for the 0-10GHz frequency band.

Acknowledgements

1.These authors contributed equally to the work.

2.*Authors for correspondence:Ljiankang@ scfc.edu .

3.Fund: Sichuan Agricultural University in Dujiangyan at the Technology Fund funded projects.

References

[1] L. W. Ke, Y. J. Chan and Y. C. Chiang, Microwave Opt. Technol. Lett., vol. 13, no.1, p.47-49 (1996)

[2] K. Kawakyu, Y. Ikeda, M. Nagaoka and N. Uchitomi, IEEE MIT-S Int. Microwave Symp. Dig, San Francisco, CA, p.647-650 (1996)

[3] Xiao-Feng Lei, Ze-Wen Liu, and Yun Xuan, etc, Measurement and Control Technology, p.7-9 (2004)

[4] Lei L. Mercado, Shun-Meen Kuo, Tien-Yu Tom Lee, and Lianjun Liu, IEEE Transactions on Components and Packaging Technologies, Vol. 27, No.3 (2004)

[5] J.B. Muldavin and G.M. Rebeiz, IEEE Microwave Wireless Comp. Lett. Vol. 11, p.373-375 (2001)

[6] Gabreil M. Rebeiz and Jeremy B. Muldavin, IEEE Microwave Magazine 59 (2001)

[7] Clark P J, Evans F C. Distance to nearest neighbour as a measure of spatial Relationships in populations[J]. Ecology, 1954, 35:445~453

[8] Donnelly K P. Simulation to determine the variance and edge-effects of total nearest Neighbor distance[A]. In: Hodder, I.R.(ed.). Simulation methods in archaeology[M]. Cambridge University Press, London, United Kingdom, 1978, 91~95

[9] Smaltschinski T. Charakterisierung von Baumverteilungen [J]. Forstwiss. Cent. bl., 11, 1998,7:355~363

[10] Jie Cai, Xiao-Ping Liao and Jian Zhu, MEMS Device and Technology. Vol 8 (2005)

A Mechanism to Improve the Implementation of Synchronization in RI

Xiao Cheng[1] Wenbo Xu[2]

1 School of Information Technology, Jiangnan University, Wuxi, 214122, China
Email: chengxiaono1@163.com

2 School of Information Technology, Jiangnan University, Wuxi, 214122, China
Email: xwb@sytu.edu.cn

Abstract

This paper proposes an implementation of wait-free synchronization for the Real-Time Java Specification. The proposed implementation can avoid priority inversion problem and support multiple threads (real-time java threads or regular java threads) to access the shared data in a wait-free manner. Through the analyze and compare both the reference implementation (RI) and our implementation on the mechanism of synchronization, our implementation outperforms the reference implementation. Therefore, it can support the Real-Time Java Specification more effectively on the mechanism of the synchronization.

Keywords: Real-Time Specification for Java (RTSJ), Synchronization, Implementation, RI

1 Introduction

Using Java as the programming language for real-time and embedded systems has attracted certain academic and industry interests in recent years. To make up deficiency about the real-time problem in the Java Language Specification and the Java Virtual Machine Specification, the Sun, IBM and other corporations have organized the requirements working group for real-time extensions for the Java Platform, and have constituted the Real-Time Specification for Java (RTSJ) [1], which provides the guideline for the implementation of Real-Time Java platform.

Following the publication of the RTSJ, there is more and more research about Real-Time Java [3,4,5,6,7,8]. To make Java more suitable for real-time programming, RTSJ enhances Java in several areas with better determinism and multithreading. The enhancements include thread scheduling and dispatching, memory management, synchronization and resource sharing and others.

Among the above enhancements, we are more interested in the synchronization. Through the research of the RTSJ and implementation of RI, in this paper, we propose an implementation about synchronization between real-time threads and non-real-time threads in RTSJ. Our implementation can avoid priority inversion problem and support multiple threads to access the shared data in a wait-free manner.

The remainder of this paper is structured as follows. The next section describes the synchronization in the RTSJ. We present our implementation in section 3. In section 4, we discuss our implementation, and compare the RI with our implementation. Section 5 concludes the paper.

2 Synchronization in the Rtsj

The RTSJ is designed for multithreading priority-based uniprocessor systems. The application program must see the minimum 28 priorities as unique; for example, it must know that a thread with a lower priority will never execute if a thread with a higher priority is ready. If threads with the same priority are eligible to run, they will execute in FIFO order.

The RTSJ [1] strengthens the semantics of Java

synchronization for use in real-time systems by mandating monitor execution eligibility control, commonly referred to as priority inversion control. A *MonitorControl* class is defined as the superclass of all such execution eligibility control algorithms. *PriorityInheritance* is the default monitor control policy; the specification also defines a *PriorityCeilingEmulaton* option.

The wait-free queue classes provide protected, concurrent access to data shared between instances of regular Java threads and *NoHeapRealtimeThreads*(NHRT).

2.1 Monitors

Java monitors, and especially the *synchronized* keyword, provide a very elegant means for mutual exclusion synchronization. Thus, rather than invent a new real-time synchronization mechanism, the RTSJ strengthens the semantics of Java synchronization to allow its use in real-time systems. In particular, the specification mandates priority inversion control. Priority inheritance and priority ceiling emulation are both popular priority inversion control mechanisms; however, priority inheritance is more widely implemented in real-time operating systems and so is the default mechanism in the RTSJ.

By design the only mechanism required by the specification which can enforce synchronized. Nothing that mutual exclusion in the traditional sense is the keyword synchronized by both instances of the specification allows the use of *java.lang.Thread*, Realtime Thread, and *NoHeapRealtimeThread* and that such flexibility precludes the correct implementation of any known priority inversion *java.lang.Thread* and algorithm when locked objects are accessed by instances of *NoHeapRealtimeThread*, it is incumbent on the specification to provide alternate means for protected, concurrent data access by both types of threads (protected means access to data without the possibility of corruption). The three wait-free queue classes provide such access.

2. 2 Wait-Free queues

Basically, there exist two different new queue classes in RTSJ: the *WaitFreeWriteQueue* class and the *WaitFreeReadQueue* class. The wait-free queues have two main methods, write and read. In *WaitFreeWriteQueue* the write method is wait-free and the read method is blocking, and in *WaitFreeReadQueue* the read method is wait-free and the write method is blocking.

The wait-free queue classes in RTSJ are used to solve a dilemma caused by NHRT, garbage collection and mechanisms for solving priority inversion in the specification. Garbage collection is an important language feature of Java and is kept in RTSJ. In RTSJ, regular Java threads and Real-time Thread cooperate with garbage collectors. The NHRT is introduced for threads which need to run without the intervention of garbage collection. When lock-based synchronization is used between threads, the priority inversion problem must be prevented by either priority ceiling emulation protocol or priority inheritance protocol, as required in the specification. Synchronization between NHRT and regular Java threads causes a dilemma if the synchronization is lock-based: regular Java threads may preempt NHRT to avoid priority inversion; in the mean time, garbage collection may preempt the regular Java threads and intervene in the execution of NHRT. For example, let us assume a NHRT TN shared information through a shared data object SO with a regular Java thread TR. The specification wants (a) the thread TN running without the interference of the garbage collection process and requires (b) that the thread TR cannot block the garbage collection process. The priority of TN is higher than that of thread TR. To protect the consistency, the shared object is guarded by a monitor; to prevent the priority inversion problem, PCP or PIP is used at the same time. If TN preempts TR while it accesses the shared object SO, the priority of TR will be prompted to be higher than TN. Now, if the garbage collection process starts, shall it preempt TR? By the requirement (a), it cannot preempt TR which blocks TN; if it preempted TR, the action renders the introduction of NHRT meaningless. By requirement (b), it has to preempt TR to satisfy the consistency of JVM. To avoid the dilemma, RTSJ introduce wait-free synchronization between NHRT and regular Java threads.

3 The Implementation of Synch ronizations for Rtsj

Our implementation is designed to take advantage of the fact that the RTSJ only requires one side of the queue to be wait-free. It is implementation as a linked list where reads are made at the head of the list and writes are made at the tail of the list. The objects in the queue are stored in the nodes that make up the queue. The nodes are reused with the help of a memory manager when they are removed from the queue since we can't allocate new memory every time we write.

In the wait-free write queue a head variable is used to keep track of the node that is currently at the beginning of the queue. Since it is only used in the blocking read it can not be changed by a preempting thread. Using a variable to keep track of the tail of the list however would put it very much at risk of being subjected to the enabled-late-write problem. To avoid this, the algorithm gives each priority its own tail variable. Since a thread must have a strictly higher priority than another thread to preempt it, a thread may safely writ to the tail variable at its own priority.

All nodes except the node at the end of the queue has a pointer to the next node in the list, so to find the correct tail node, the algorithm has to go through all tail variables looking for the one with a null pointer. To simplify this, the tail variables are stored in an array where the indices correspond to the available priorities. The same method is used for the wait-free read queue but there it's the head that needs to be stored in an array.

Instead of checking from the start if another thread needs help, we announce our operation first and then check if any other thread needs help. In the worst case we might need to help all other priorities. The announcement is made trough an array of nodes the size of all priorities in the system. If we are dong a write we store the new node in the announcement array and if we are dong a read we store a special *dumbCell* node in the array.

Wait-Free Write

1) Store the object that is to be written in a new node that we get from the memory manager.

2) Place the new node in the announcement array to signal to any preempting thread that we are writing an object.

3) Go through the announcement array and help all lower priority threads that need help.

a. Find the correct tail by gong through the tail array looking for nodes that doesn't have a pointer to another node.

b. Make the tail point to the new node.

c. Store the new node in the tail array using the priority of the thread we are helping as index.

d. Clear the current threads' entry in the announcement array and help the next thread.

4) Return true

Wait-Free Read

1) Store a *dumbCell* in the announcement array using the priority as index to signal to any preempting thread that we are reading an object.

2) Go through the announcement array and help all lower priority threads that need help.

a. Find the correct head by going through the head array looking for nodes that have a pointer to another node.

b. If we don't find a head the queue is empty and we return false.

c. The head node is placed in the announcement array using the priority as index.

d. We store the new head, the node that head is pointing to, in the head array.

e. The old head nodes' pointer is cleared.

Return the value in the node and put it in the announcement array.

4 Discussion

We implemented *WaitFreeWriteQueue* (The implementations of the two wait-free queue classes are quite similar. So, we only present the implementation of the *WaitFreeWriteQueue* class to illustrate the ideas.) and test it under the RTSJ reference implementation from Timesys. We compare the proposed

implementation with the one in the reference implementation from Timesys.

For the Timesys implementation, we don't have the source code. To our knowledge, their implementation only has support for wait-free access from one thread at a time. For example, if two real-time threads would like to do wait-free writes to a queue at the same time, lock-based synchronization needs to be applied between the two threads. Therefore, when several real-time threads want to access the wait-free implementation from Timesys, we have to choose some kinds of protocol to avoid the priority inversion problem.

For our implementation, we use a simple announce-and-help scheme for the operation to wait-free queue. When a task will run and there is a task which can preempt it, the higher priority task will help the lower priority talk. It only announces itself. Because the priorities are bounded, there always exists a task which will not be preempted by another renqueue task. Therefore, all tasks that announced their operations will be helped (either by themselves or by higher priority tasks). So, our implementation strictly and effectively avoids the priority inversion problem. Meanwhile, both these queue classes are unidirectional. The information flow for the *WaitFreeWriteQueue* is from the real-time side to the non-real-time one, as shown in Figure 1. The information flow for the *WaitFreeReadQueue* is from the non-real-time side to the real-time one, as shown in Figure 2. When a NHRT wants to send data to a regular Java thread, it uses the write (real-time) operation of *WaitFreeWriteQueue* class. Regular threads use the read (non-real-time) operation of the same class to read information. The write side is non-blocking and wait-free, so that NHRT will not experience delays from the garbage collection. The read operation, on the other hand, is blocking. The *WaitFreeReadQueue* class, which is unidirectional from non-real-time to real-time, works in the converse manner. Because of the announce-and-help scheme and the fact that wait-free queues are unidirectional, our implementation not only can avoid the priority inversion problem but also support multiple threads to access the shared data in a wait-free manner at a time, which is the RI can not support.

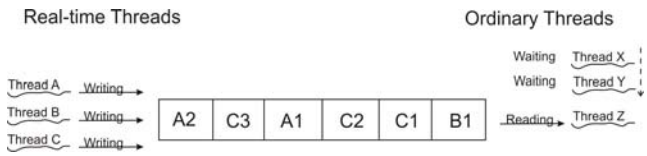


Figure 1 WaitFreeWriteQueue

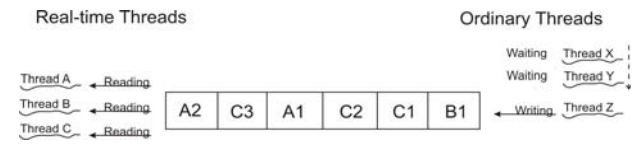


Figure 2 WaitFreeReadQueue

Compare the reference implementation and our implementation on the mechanism of the synchronization, we can find that our implementation outperforms the reference implementation.

5 Conclusions

In this paper, an implementation of synchronization for the real-time Java Specification is presented. The implementation can avoid priority inversion problem and support multiple threads to access the shared data in a wait-free manner. Therefore, it can support Real-time Java Specification effectively on the mechanism of synchronization.

Meanwhile, through the analyze the synchronization mechanism of the two implementations, our implementation outperforms the reference implementation.

References

[1] G. Bollela, J. Gosling, B. Brosgol et al "The Real-Time Specification for Java", Addison Wesley, 1st edition, 2000

[2] M.Herlihy, "Wait-Free Synchronization," ACM TransactIons on Programming Languages and Systems, Vol. 11, No 1, January 1991, Pages 124-149

[3] H. Cai and A. Wellings, "Towards a high integrity real-time Java virtual machine," ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS: vol. 2889, pp. 319-334, 2003

[4] http:/www.cs.purdue.edu/homes/jv/soft/ovm/documents. htm

[5] "Java Reference Implementation (RI) and Technology

Compatibility Kit (TCK)," http://www.timesys.com

[6] D. S. Hardin, "aJile Systems: Low-Power Direct-Execution JavaTM Microprocessors for Real-Time and Networked Embedded Applications," Available at http://www.ajile.com/downloads/aJile-white-paper.pdf

[7] S. A. Ito, L. Carro, and R. P. Jacobi, "Making Java work for microcontroller applications," Design & Test of Computers, IEEE, vol. 18, pp. 100 - 110, 2001

[8] M. Schoeberl, "JOP: A Java optimized processor," ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS 2003: OTM 2003 WORKSHOPS, vol. 2889,2003, pp. 346-359

[9] P. Dibble and A. Wellings. The real-time specification for java: current status and future work. In Proceedings of the Seventh IEEE International Symposium on Object-Oriented Real-Time Distributed Computing, pages 71-77. IEEE Computer Society Press, 2004

[10] M. Higuera-Toledano. Illegal references in a real-time java concurrent environment. In Proceedings of the Seventh IEEE International Symposium on Object-Oriented Real-Time Distributed Computing, pages 321–324. IEEE Computer Society Press, 2004

[11] Jason Baker, Antonio Cunei, Chapman Flack, Filip Pizlo, Marek Prochazka, Jan Vitek, Austin Armbuster, Edward Pla, and D. Holmes, "Real-time Java in Avionics Applications," presented at Proceedings of the 12th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), 2006

[12] T. Lindholm and F. Yellin, "The Java Virtual Machine Specification, 2nd edition," Addison Wesley, 1999

[13] J. Gosling, B. Joy, G. Steele, and G. Bracha, "The Java Language Specification Second Edition," 2000

An Instruction Reconfigurable Framework for RTSJ-optimized Java Processor

Xiaolong Ren[1 2] Zhilei Chai[1]

1 School of Information Technology, Jiangnan University, Wuxi, Jiangsu 214122, China

2 Department of Continuing Education, Shanxi Architectural Technical College, Taiyuan, Shanxi 030006, China
Email: xiaogou04@163.com

Abstract

Due to the preeminent work of the real-time specification for Java (RTSJ), Java is increasingly expected to become the leading programming language in embedded real-time systems. In order to provide an efficient real-time Java platform suitable for embedded applications, we designed a FPGA-centric Java processor optimized for RTSJ. Because the software for most of the embedded systems is application-specific and hardly changed, some bytecodes are probably never used in a special Java-based embedded application. In this paper, an instruction reconfigurable framework for the Java processor is introduced. Based on this framework, the processor can be customized according to the application requirement. Only the bytecodes really used in the application will be implemented, making the Java processor resource-saving and organization simple.

Keywords: Real-time Systems, Real-time Specification for Java, Java Processor, Worst Case Execution Time, Instruction Reconfigurable

1 Introduction

Real-time specification for Java [1] is a real-time extension for the Java language specification [2] and the Java virtual machine specification [3] under the requirements for real-time extensions for the Java platform [4]. It provides an application programming interface that enables the creation, execution, and management of Java threads with predictable temporal behavior. With the advantages as an object-oriented and concurrent programming language, and with the real-time performance guaranteed by the RTSJ, Java is increasingly expected to become the leading programming language in embedded real-time systems.

Currently, to provide an efficient Java platform suitable for embedded real-time applications, many different implementations are proposed. These implementations can be generally classified as Interpreter (such as RJVM [5] and Mackinac [6]), Ahead-of-Time Compiler (Anders Nilsson et al [7] and OVM [8]) and Java Processor (aJile-80/100 [9], FemtoJava [10] and JOP [11]). Comparing with other implementing techniques, a Java processor that can execute Java bytecode directly in silicon avoids the slow execution model of an interpreter and the memory requirements of a compiler, making it an appealing execution platform for Java in embedded systems. Nevertheless, few of the current Java processors provide special support for mechanisms of the RTSJ. In [12], we propose a Java processor optimized for RTSJ (called JPOR for short) for embedded real-time systems. This processor provides special support in hardware for mechanisms of the RTSJ such as asynchronous transfer of control (ATC), thread, synchronization, memory management, and offers a simpler programming model through ameliorating the scoped memory of the RTSJ.

Generally, the software for most of the embedded systems is special purpose and hardly changed. There is an enormous diversity on the use frequency of a special bytecode between different applications. Some bytecodes are probably never used in a special Java-based embedded application. On the other hand,

the reconfigurability of the FPGA technology makes the fine-grain reconfiguration feasible.

In this paper, we introduce the instruction reconfigurable framework for JPOR processor. Based on this framework, the bytecodes never used or too complex to implement can be eliminated from the processor, making the processor resource-saving and organization simple. The complex bytecodes can be replaced by a sequence of simple bytecodes, which will improve the pipelining throughput of the processor.

2 JPOR Architecture Overview

2.1 The java platform based on JPOR

A Java processor alone is not a complete Java platform. In our implementation, the complete Java platform is composed of the CConverter (class loader), APIs (class library) and the JPOR processor (including execution engine, memory and I/O), which is shown in Figure 1.

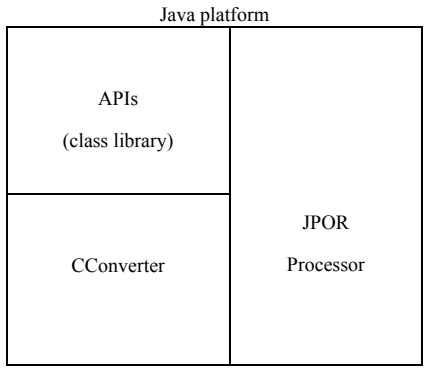


Figure 1 Java platform based on JPOR

The APIs provide a profile based on the RTSJ for Java application programmers. JPOR is the processor proposed in [12] to execute Java bytecode directly and provide support optimized for the RTSJ. The CConverter is the software we designed to preprocess the Java class file before being executed on top of the processor.

Similar to other real-time Java platforms, the execution of Java applications on this platform is also divided into two phases: initialization phase (non real-time) and mission phase (real-time). During the initialization phase, all of the class files including application code and class libraries referred by the program are loaded, verified, linked and transformed into a binary representation. This transformation is performed by the CConverter and not executed on top of JPOR. During the mission phase, the binary representation is downloaded and executed on JPOR with predictable WCET.

2.2 Architecture of JPOR processor

As shown in Figure 2, the JPOR processor core is simply divided into three pipeline stages: Fetch Instruction, Decode and Execution.

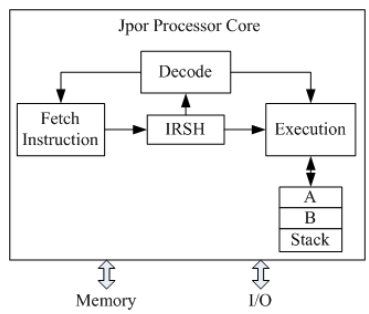


Figure 2 architecture of JPOR processor

Fetch Instruction: To build a self-contained Java processor, direct access to the memory and I/O devices is necessary. However, there is no bytecode defined for low-level access in conventional JVM. Some extended instructions should be defined to solve this problem. In JPOR, the bytecodes from 0xcb to 0xe4 that are used by _quick bytecodes in the conventional JVM are selected as extended instructions because _quick instructions are never used in JPOR. Take M2R(0xce), reg1, reg2 for an example, this extended instruction reads data from memory or I/O according to the address denoted by register reg2, and writes it into register reg1. Extending instructions in this way abides by the uniform format with other bytecodes. So, the fetch unit of JPOR can process them conveniently as a single instruction set. To reduce memory access frequency, a register IRSH is used as a FIFO in JPOR to fetch multiple instructions at a time. The fetched instructions are located into register IRSH for being used in decode or execution stages.

As shown in Figure 3, the MPC can be updated by the interrupt, next MPC or IRSH from the fetch unit. It

is used as the entry of decode unit to find the proper decoded control signals.

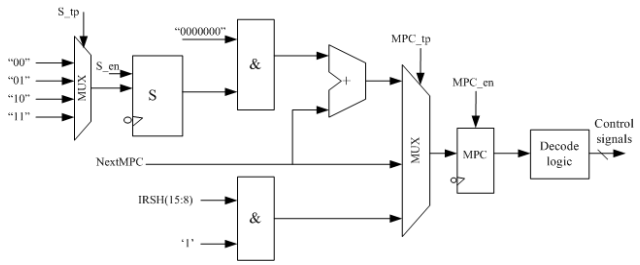


Figure 3 the logic diagram of the decode unit

Decode: The logic diagram of the decode unit is shown in Figure 3. It is implemented with a special hardwired model similar to the micro-programmed structure. Thus, the instructions can be added into or eliminated from the processor conveniently. The details of the decode unit will be introduced in section 3. The decode unit always fetches the highest 8-bit of IRSH as the entry to find the proper decoded control signals.

Execution: The JPOR processor is implemented as a stack-oriented machine to fit the JVM behavior. For the stack is accessed frequently for operands and Locals. It is placed into the same chip with the processor core. This stage performs ALU operations, load, store, and stack operations. To avoid extra write-back stage and data forwarding, similar to reference [11], two explicit registers A and B providing operands for ALU are also used as the two topmost stack elements. The execution unit can get operands from IRSH, stack, memory and I/O.

Memory and I/O: the binary representation produced by the CConverter is downloaded into memory. All of the string, static fields and other data can be accessed by their addresses directly. The processor core can access I/O through interrupt or loop from a uniform addressing with the memory.

3 Instruction Reconfigurable Framework for JPOR

As shown in Figure 3, the decode logic is the key part of the decode unit. In JPOR, the decode logic is implemented with an instruction reconfigurable framework shown as Figure 4. The MPC is the input signals for decoding and the data is the decoded output signals.

```
process(MPC) begin
case MPC is
-- fetch:
when "000000000"=>data<="101111111111……1100000000011";
when "000000001"=>data<="111111111111 …… 000000000101";
when "000000010"=>data<="111111111111 …… 000000000101";
when "000000011"=>data<="111111111111 …… 000000000101";
when "000000100"=>data<="111111111111 …… 000000000101";
--decode:
when "000000101"=>data<="111111111111 …… 110000000000";
--other icr instructins area
when "000001000"=>data<="111101011100 … 010000000000";
--nop 0x00
when "100000000"=>data<="111111111111 …… 010000000000";
--aconst_null 0x01
when "100000001"=>data<="111111111111 …… 000000001000";
--arithmatic and logic operations
--iadd(0x60)
when "101100000"=>data<="111111001100 … 010000000000";
--isub(0x64)  B - A
when "101100100"=>data<="111111001100 … 010000000000";
……
when others => data <= "111111111111 …… 1100000000000";
end case;
end process;
```

Figure 4 the instruction reconfigurable framework described in VHDL

Some details are omitted for the clearance. One instruction can be implemented by inserting a sequence of 'when' clauses.

As shown in Figure 4, with the increment of MPC entries, the complexity of the decode logic increases dramatically. Through this framework, an instruction can be inserted into or deleted from the decode logic conveniently by inserting or deleting a sequence of "when "000000101"=>data<="111111111111 … 0000000000" clauses shown in Figure 4. After synthesizing, the JPOR processor will only include the bytecodes actually used by the application.

4 Discussion and the Status

With the development of the FPGA technology, today it is even possible to build an entire system on a single FPGA chip. It leads to a dramatic decrease in design time and cost compared to ASIC-based systems, making FPGA-centric systems increasingly useful in embedded systems. The advantage of the FPGA-centric system is that it can be reconfigured according to the actual application. Some research focused on reconfiguration of the FPGA-centric Java processor is proposed in [11] and [13]. JOP is a micro-programmed Java processor. Its micro code is the native code defined by JOP itself. When bytecode being executed, it should

be translated into micro codes first. One bytecode is translated into one or more micro codes. The reconfiguration is accomplished by updating the micro code memory. The ref. [13] is focused on increasing the overall performance of the FPGA-centric Java processor through adding a hardware accelerator. The reconfiguration is based on the Java processor core, not the processor itself.

The JPOR processor designed by our group is a hardwired FPGA-centric Java processor. In JPOR, the bytecodes from 0xcb to 0xe4 that are used by _quick bytecodes in the conventional JVM are selected as extended instructions because _quick instructions are never used in JPOR. Extending instructions in this way abides by the uniform format with other bytecodes. So, the fetch unit of JPOR can process them conveniently as a single instruction set, avoiding the cycles to translated bytecodes into the native instruction. The primary goal of reconfiguration for JPOR is saving the hardware resources and enhancing the execution frequency through simplifying the combinational logic. Furthermore, the complex bytecodes can be replaced by a sequence of simple bytecodes, which will improve the pipelining throughput of the processor.

Currently, the reconfiguration is accomplished item by item manually. It will be accomplished by the CConverter automatically at the next step.

5 Summary

In this paper, an instruction reconfigurable framework is introduced for the JPOR processor proposed in our previous work [12]. This framework utilizes the properties of FPGA-centric systems making the processor instruction reconfigurable according to the actual application. With this framework, only the bytecode actually used in the application will by implemented in hardware. It can decrease the hardware complexity and improve the execution frequency as well. Another advantage of the instruction reconfigurability is that it can implement complex instructions into a sequence of simple one, which will improve the pipelining throughput of the processor.

References

[1] G. Bollela, J. Gosling, B. Brosgol, P. Dibble, S. Furr, D. Hardin, and M. Trunbull, "The Real-Time Specification for Java," Addison Wesley, vol. 1st edition, 2000

[2] J. Gosling, B. Joy, G. Steele, and G. Bracha, "The Java Language Specification Second Edition," 2000

[3] T. Lindholm and F. Yellin, "The Java Virtual Machine Specification, 2nd edition," Addison Wesley, 1999

[4] L. Carnahan and Marcus Ruark, "Requirements For Real-time Extensions For the Java™ Platform," http://www.itl. nist. gov/div897/ctg/real-time/rtj-final-draft.pdf, September 1999

[5] H. Cai and A. Wellings, "Towards a high integrity real-time Java virtual machine," ON THE MOVE TO MEANINGFUL INTERNET SYSTEMS: vol. 2889, pp. 319-334, 2003

[6] G. Bollella, B. Delsart, R. Guider, C. Lizzi, and F. Parain, "Mackinac: making HotSpot/spl trade/ real-time," presented at Eighth IEEE International Symposium on Object-Oriented Real-Time Distributed Computing, ISORC 2005, 45 – 54

[7] A. Nilsson and S. G. Robertz, "On real-time performance of ahead-of-time compiled Java," presented at Eighth IEEE International Symposium on Object-Oriented Real-Time Distributed Computing, ISORC 2005, 372 – 381

[8] http:/www.cs.purdue.edu/homes/jv/soft/ovm/documents.htm

[9] D. S. Hardin, "aJile Systems: Low-Power Direct-Execution JavaTM Microprocessors for Real-Time and Networked Embedded Applications," Available at http://www.ajile.com/ Documents/aJile-white-paper.pdf

[10] S. A. Ito, L. Carro, and R. P. Jacobi, "Making Java work for microcontroller applications," Design & Test of Computers, IEEE, vol. 18, pp. 100 - 110, 2001

[11] M. Schoeberl, "JOP: A Java Optimized Processor for Embedded Real-Time Systems", http://www.jopdesign.com/ thesis/thesis.pdf, 2005

[12] Z. L. Chai, W. K. Zhao, and W. B. Xu, "Real-time Java Processor Optimized for RTSJ" Accepted by The 22nd Annual ACM Symposium on Applied Computing, SAC'2007

[13] F. Gruian, P. Andersson, K. Kuchcinski and M. Schoeberl, "Automatic Generation of Application–Specific Systems Based on a Microprogrammed Java Core", 2005 ACM Symposium on Applied Computing, pp.879-884

Design of Embedded UDP Protocol based on Foreground/Background System

Chunyan Zhang[1] Bo Xiao[2]

1 Department of Information, Wuxi Professional College of Science and Technology, Wuxi, 214028, China
Email: zcy0006@163.com

2 Department of Electronic Science and Engineering, Nanjing University, Nanjing, 210093, China
Email: eewolf@163.com

Abstract

This paper presents an implementation of high-reliable network card driver used in the 8-bit mcu, which enables the mcu with low performance to drive RTL8019 always correctly even in a high-load network. The method for designing a simplified communication protocol stack which includes UDP, IP, ARP and ICMP is introduced, as well as the program structure with the ability to let the embedded equipment run a long-time consumed background program without the support of a operating system by running the whole communication stack related code in the interrupt foreground program. At last this udp protocol is used in the OSD equipment in order to realize parallel control by arbitrary number of computers.

Keywords: Embedded UDP Communication, Rtl8019as, Ethernet Driver, MCU, OSD

1 Introduction

With the development of the Internet, the concept of "information" embedded equipment, which can be set an IP address just like a PC, appears. Through the ethernet, the ability of long-range communication and monitoring is realized. The embedded ethernet application based on arm processor is often seen [1, 2], whose design is relatively simple. But in some applications, the use of arm processor will greatly increase the costs, while the 8-bit MCU with a very low price also can realize the required ethernet communication functions. The operating speed and storage space of 8-bit MCU are very limited. The free and open source embedded ethernet communication protocol stack which can be used on MCU platform includes uip [3] and lwip [4]. But the realization structure of uip is single-threaded and the whole of codes runs in the main program. And when the control tasks need a long time($\geqslant$ 200ms), the ethernet communication task will have the overtime errors because of long-time un-running. Although lwip has a multi-threaded structure, the need of the support of an embedded operating system leads to a heavier running load. This paper introduces a novel design scheme of the embedded protocol stack: The MCU uses the T2 timer generate a continuous 19ms timer interruption to query whether there is a data packet for transmitting or receiving in the manner of time-triggered approach in the interruption foreground program; The control tasks are running in the background main program, doing corresponding reactions according to the control command transmitted by PC via ethernet.

2 Design of Hardware

In this design, the MCU is winbond w78058 based on the 51 core, which has the P4 port and thus has the 4 extended I/O lines because of the PLC44 package. The frequency of crystal is selected to be 36.864 MHz in order to set an accurate serial communication baud rate. It should be noted that the value of the start-oscillation capacitor should be 5pf, or the crystal will be worked in the harmonic frequency. The hardware connection diagram is shown in the Figure 1. Because the external

32K bytes ram is extended for the MCU and the memory mapped approach is used for the operation of the network card, the address line A15 is used as the ram chip-select pin; the network card chip is RTL8019AS [5, 6], whose IOCS16 pin is connected to ground to let the chip work in the 8-bit mode. The BD0~3 pins are network card base address selected pins, which are set to 0110 in order to stagger the address space of the network card and the external ram. The address space for ram is 0x0000~0x7fff and for the network card registers is 0x8040~0x804f. The reset pin RSTDRV is linked to the watchdog chip as well as the reset pin of the MCU. The x5045 is used as the watchdog chip, in which the serial EEPROM stores the MAC address, IP address, gateway address, subnet mask, UDP communication port, and other network configuration information.

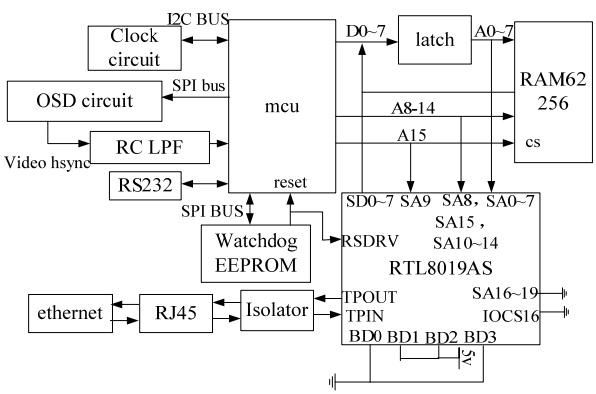


Figure 1 Block diagram of the hardware realization

3 Design of Software

When receiving data from the network card, there are two methods: query operation and interrupt request. Because of the limited running speed, the MCU can't realize real-time response to the interrupt request and only can operate the network card in the means of query.

The receive buffer of network card is a ring FIFO, which has 64 pages (256 bit per page). The CURR register indicates the next page of the current write page, which is the FIFO write pointer. After the network card completes the write operation, the CURR will plus one automatically; The BNRY register is the boundary register, which is the FIFO read pointer. The BNRY is controlled by the network card driver program. When the program read a page of data from the FIFO, the BNRY should be added one and updated by the driver program self. Therefore, the initial state is CURR = BNRY+1, which means the fifo is empty; So the usual method to query if there is a new package for receiving is to judge whether CURR equals to BNRY+1 [7, 8, 9, 10] directly. When there are many broadcast packets and the task of control takes a long time, the receive FIFO will overflow and the network card is unable to communicate because the program doesn't read data from network card for a long time. To deal with this situation, two schemes of driver program are given:

(1) When the FIFO overflows, the corresponding pins of the network card will output the IRQ signals. Although the pins are not linked to the MCU, the MCU can get the interrupt information from the OVW bit in the ISR register. If the overflow occurs, the operations done by the MCU are: If the network card is sending the data, wait for the completeness; Stop the DMA operation; Reset the counter registers; Reset the network card through the soft reset register; Re-initialize the network card.

(2) The reason that the network card can't receive data in the event of overflow is that when the ring FIFO is full, CURR equals to BNRY+1 again, which is identical to the initial state. The driver program will consider the FIFO is empty by mistake.

A relative simple solution is: Because the completeness of read operation is realized by the value update of BNRY, the full FIFO can be cleared by the method that rewriting BNRY with the value read from BNRY. Therefore, as long as the situation CURR= BNRY+1 occurs, rewrite BNRY. If the FIFO overflows, this operation will clear the FIFO; if the FIFO is empty, this operation has no effect.

The full TCP/IP protocol stack should be simplified according to the desired functions [11, 12]. The ARP protocol is implemented in the data link layer. The capacity of ARP table is ten, of which each item uses 2 bytes for storing the ttl (time to live) value. The

ttl value is set to be the maximum value 0xFFFF, which minuses one at each T2 timer interrupt until 0. The simplified IPv4 protocol is implemented in the network layer, which doesn't support IP option, IP fragment and so on. The ICMP protocol only supports the echo behavior, namely only can reply the ping data packets from the upper computers. Because the equipment is mainly used in the LAN, the functions of the TCP protocol are a bit redundant, and the load of the TCP protocol is relatively heavy for an 8-bit MCU. The more practical and efficient UDP protocol is used in the transport layer.

The reliability of the UDP protocol can be improved by the function expansion of the application layer protocol. The application data frame is added with a sequence ID. The length of ID is 2 bytes, which enables the ID doesn't repeat in a certain period of time. After receiving the data, the format of application layer is shown in the Figure 2.

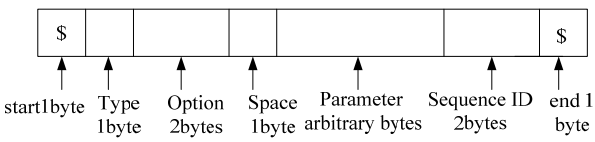


Figure 2 Data frame format of application layer

In this programming model, the programming skills are:

(1) Use the small model to compile the program to improve the running speed. The variables which are often used are stored in the internal ram, while others are stored in the external ram.

(2) The flow diagrams of program are depicted in the Figure 3 and the Figure 4. Through putting the communication process in the T2 timer interrupt service program and the control process in the main program, the purpose of asynchronous communication without the support of an embedded operating system is achieved. A FIFO realized by a two-dimensional array is used as a communication pipe. Because of the parallel running structure, T2 timer interrupt should be disabled when the main program changes the read and write pointers of the FIFO to avoid conflict and uncertainty. In addition, the serial port is used to download the network configuration information into the EEPROM. Because of the existence of T2 timer interrupt, the interrupt priority level of serial port should be promoted to enable normal communication.

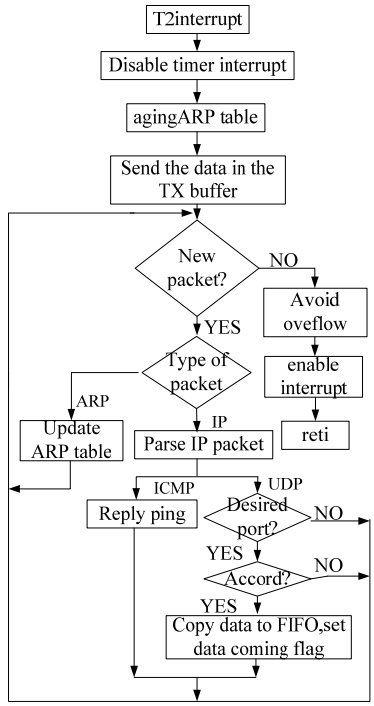


Figure 3 Flow diagram of T2 interrupt service program

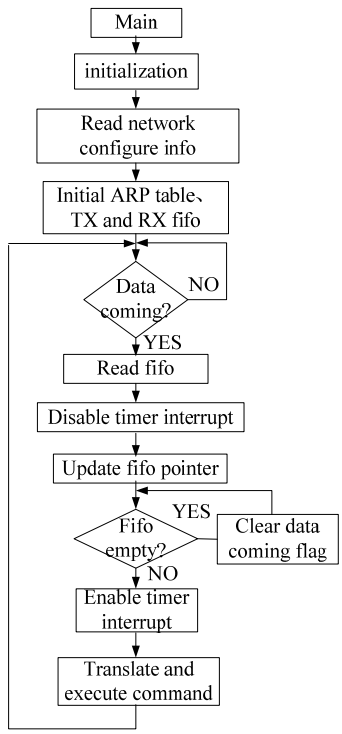


Figure 4 Flow diagram of main program

(3) In order to increase the running speed of the background control tasks, when the related functions are invoked, the global variables are used for information transmission instead of the arguments to avoid the time consumption of the operations of the stack.

(4) Two global arrays which are defined in the external ram are used as send and receive buffers instead of the dynamic linked list ram allocation scheme used in LwIP to increase speed. The receiving data type conversion diagram is depicted in the Figure 5.

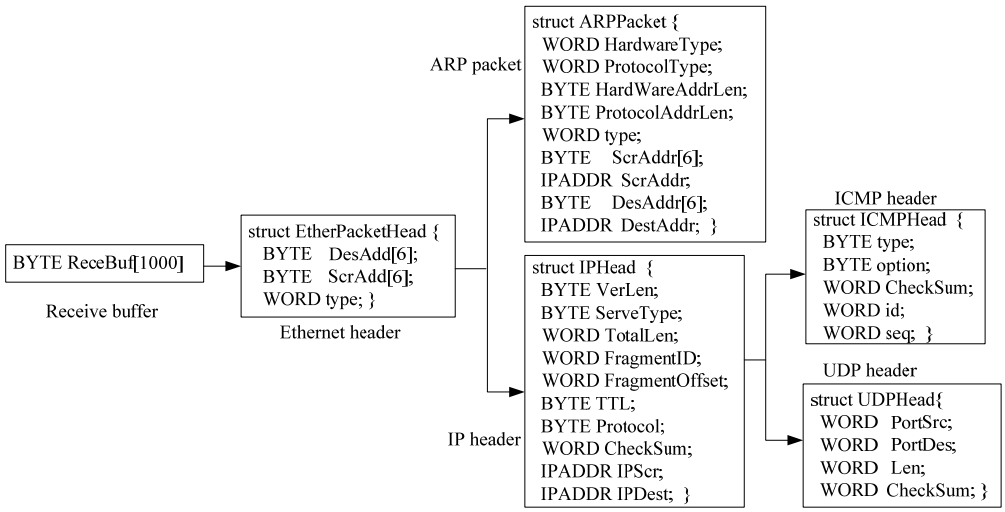


Figure 5 Flow diagram of the receiving data type conversion

(5) After the RT8019as chip has received data from the network, three data fields (4 bytes) which are receiving state, next page pointer and ethernet frame length, will be automatically added to the header of the data. The length of the order actually used normally will not exceed 120 bytes. Then, in order to increase the processing speed, when the ethernet frame length is in excess of 200 bytes (the length of IP and UDP headers is taken into account), the pointer of the next ethernet frame is assigned to the BNRY register to drop the current frame.

(6) In order to follow the protocol stack layered programming principles, the function nesting level is deep. So a big stack is needed to store the temporary variables. The default value of stack pointer SP should be changed to avoid stack overflow. The protocol stack diagram is showed in Figure 6.

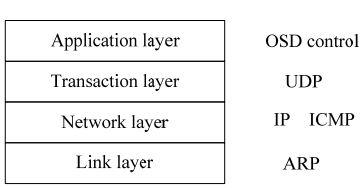


Figure 6 Hierarchy chart of the protocols

4 Results of Experiment

Technical parameters: code (byte): 9233; external ram (byte): 6720; internal ram (byte): 80.3; sending speed (byte / sec): 3200.

The tests to the performance of the protocol stack have two aspects:

(1) Ability of long time work: During 400 hours, PC communicates with the embedded equipment continuously using the command “ping -t”. Every ICMP packet is replied without a loss.

(2) Communication capacity: The equipment is connected to a large-scale LAN. While replying a lot of broadcast packets, the equipment handles the control command transmitted by 5 PCs at the time interval of 200ms, without command loss and system halted phenomenon.

The correctness of the receiving data is guaranteed by the double checksum fields defined in the UDP and IP protocol headers. The experiment result shows the designed protocol stack has a high speed and reliability.

5 Conclusion

This paper discusses the design scheme of the embedded ethernet UDP communication protocol stack based on the foreground/background system. The design of the high- reliable driver program is also introduced, which enables the MCU with a low speed to do ethernet communication and long time control without the support of an embedded operating system. The programming experiences and skills are summarized.

References

[1] H.T. Zhang, S.L. Nie, D.M. Tang, "Design of IPX Model for 32 Bits Microcontroller ARM", Computer Engineering, 33(7), 2007, pp.252-254

[2] X.J. Zhang, Y. Liu, "Ethernet driver design and implementation based on S3C44B0", Journal of Shananxi University of Technology, 22(4), 2005, pp.58-62

[3] Y.I. Su, R.Q. Deng, "An Embedded Network Access Module", Computer Applications, 26(6), 2007, pp.45-48

[4] H. Li, X.J. MA, "Implementation of lwip protocol in the uc/OS", Journal of Information Engineering University, 6(2), 2006, pp.81-84

[5] J.X. Su, Q.J. Yang, "The Detailed Configuration of Ethernet Controlling Chip RTL8019AS", .Modern Electronics Technique, 30(22), 2007, pp.1511-153

[6] X.L. Hu, G. Wu. "Application of RTL8019AS in Embedded Ethernet System", Electronic Measurement Technology, 5(3), 2005, pp.81-82

[7] D.Z. Deng, W.X. Zhang, D.M. Tang, "Application of Uip TCP/IP Stack in 51 mcu", Computer Information, 20(3), 2004, pp.88-90

[8] Y.T. Zhang, D. Liu, "Uip Protocol Analysis and its application", Journal of Information Engineering University, 7(42), 2006, pp.147-152

[9] X.P. Fan, Y.G. Niu, "Implementation of Ethernet Communication of the Electronic Belt Scale Instrument", Sci-tech Information Development, 17(8), 2007, pp.226-228

[10] G.E. Li, F.T. Wang, "Design and Realization of Embedded Ethernet Interface Based on ARM", Computer Information, 24(2), 2007, pp.40-41

[11] X.C. Huang, "TCP/IP Communication of MCU on RTL8019AS", Electronics and Computer, 22(3), 2005, pp.228-232

[12] Y.M. Dai, J.G Shi. "Temperature and Humidity Monitor System Based on uIp", Journal of Sichuan University of Science & Engineering, 20(2), 2007, pp.50-53

Combining Circulant Space-Time Coding with IFFT/FFT and Spreading

Xiaonan Chen Peilv Ding

Department of Information Engineering, Wuxi Professional College of Science and Technology
Wuxi, Jiangsu 214028, P.R.China

Email:cxn@wxstc.cn

Abstract

Circulant structures were among the first space-time coding techniques ever used for Multiple- Input Multiple-Output (MIMO) systems due to their simplicity and full rate. In this correspondence, the circulant structure is combined with an inverse fast Fourier transform (IFFT) at the transmitter and a fast Fourier transform (FFT) at the receiver due to the fact that a circulant matrix is diagonalized by the discrete Fourier transformation matrix. Using this method, the spatial mixing effect of the MIMO channel is decoupled but the diversity gain is lost. We propose to recover the diversity loss by spreading the transmitted symbol vector. According to the full diversity design criterion for our scheme, the precoding or constellation rotation matrices for the signal diversity design problems can be used as spreading matrix to achieve full diversity. The proposed scheme is also full rate and can be easily applied to any number of transmit antennas. Our simulation results show that the performance of our scheme is close to the performance of the ideal orthogonal space-time code and much better than the conventional circulant space-time code.

Keywords：MIMO, Space-time, Circulant, Diversity, Spreading

1 Introduction

Space-time transmit structures are very critical for multi-antenna systems and have attracted extensive research interest. The orthogonal space-time block code (OSTBC) was first introduced in [1] for two transmit antennas and was then extended to a general number of transmit antennas in [2]. The orthogonality in the code enables maximum likelihood(ML)detection based only on linear processing. However, it was shown that there is no complex orthogonal space-time design that provides full diversity and full rate for more than two antennas. A rate loss of one-fourth or more is needed to keep the orthogonality and full diversity.

Space-time structures have also been proposed in a variety of other prior works (see for example the discussion and references in[3]).Circulant structures were among the first space-time coding techniques ever used for Multiple-Input Multiple-Output(MIMO) systems due to their simplicity and full rate[4],[5]. And it can be easily applied to any number of transmit antennas without any design effort. Due to its special structure, a very important property of circulant matrix is that it can be diagonalized by the Fourier transformation matrix.

In this paper, we first combine the circulant structure with inverse fast Fourier Transform (IFFT) and fast Fourier transform(FFT)to utilize that special property of circulant matrix. By taking the IFFT at the transmitter, sending the circulant matrix through the multiple antenna channel, and taking the FFT at the receiver side, the spatial mixing effect of flat MIMO channel is eliminated. Similar to Orthogonal Frequency Division Multiplexing (OFDM),every symbol is affected by one element of the FFT vector of the channel. But there is a spatial diversity loss because of the same

reason that OFDM loses the multipath diversity. To recover full diversity gain, we propose to spread the symbols by well designed unitary matrices. According to the full diversity design criterion for our scheme, the precoding or constellation rotation matrices for the signal diversity design problems can be used as spreading matrix to achieve full diversity. The scheme is always full rate and can be easily extended to any number of transmit antennas by selecting the appropriate spreading matrix. Our simulation results on Quadrature Phase Shift Keying (QPSK) and 16 Quadrature Amplitude Modulation (16QAM) constellations show that the performance of our scheme is close to the performance of ideal orthogonal space-time coding. It was also compared with the diagonal algebraic space-time(DAST)code[6]and the conventional circulant space-time code.

The outline of the paper is as follows. Section II describes the signal model and some space-time block code background. The proposed space-time transmission scheme is discussed in Section III. The simulation results are presented in Section IV. Section V contains a concluding discussion.

2 Preliminaries

A. Signal Model

For notational simplicity we consider a Multiple-Input Single-Output (MISO) channel with N_t transmit antennas and one receive antenna; extending the results in this paper to multiple receive antennas is straight forward.

A flat fading MISO channel can be described by a $N_t \times 1$ channel vector

$$\mathbf{h} = [h_1, ..., h_{N_t}]^T, \quad (1)$$

whose nth entry h_n is the fading coefficient between the nth transmit antenna and the receive antenna. We further assume: (1) the channel is slowly time-varying so that h is constant during the transmission of one block, (2) the elements of h are independent, identically distributed (i.i.d.) complex Gaussian random variables, and (3) the receiver has perfect knowledge of the channel.

When a $T \times N_t$ space-time block code $\mathbf{X}$ is transmitted, we receive

$$y = \mathbf{X}h + w, \quad (2)$$

where T is the number of time slots, y is the $T \times 1$ received signal, and w is zero-mean, white, complex Gaussian noise with variance $N_0/2$ per real and imaginary dimension.

The total average transmitted power over a time slot is defined as

$$E_t = tr(X^H X)/N_t, \quad (3)$$

where tr(.) denotes the trace.

B. Space-Time Block Code

A space-time block code (STBC) is the mapping from a block of p symbols $\{s_1, ..., s_n\} \in A^p$ to a $T \times N_t$ space-time code matrix X where A is the constellation.

$$\{s_1, ..., s_n\} \rightarrow X, \quad (4)$$

The rate of this code is $R = p/T$ symbols per channel use.

When the mapping is linear in the symbols $\{s_1, ..., s_n\}$ we have linear STBC which is a very important subclass of STBC. For linear STBC, the transmit signal X is formed as follow

$$X = \sum_{n=1}^{p} (\mathrm{Re}(s_n) A_n + \mathrm{Im}(s_n) B_n), \quad (5)$$

where Re(.) and Im(.) denote the real part and imaginary part operator, $\{A_n, B_n\}$ is a set of fixed complex matrices.

Transmit diversity gain D_t is one of the most important features of a space time code. Under the assumptions given in Section II-A, D_t of a given STBC with maximum likelihood decoding is defined as

$$\mathrm{Dt} = \min_{s \neq v; s, v \in A^p} \mathrm{rank}(X(s) - X(v)), \quad (6)$$

The maximum D_t achievable for a MIMO system with N_t transmit antennas is N_t.

Also the coding gain of the code is optimized by the determinant criterion: maximizing the minimum of the determinant of $(X(s) - X(s))^H (X(s) - X(v))$ over all

possible pair of s and v.

3 The Space-Time Transmission SchemeBased on Circulant Matrix

A. Combining Circulant Structure with IFFT/FFT

A $N \times N$ circulant matrix is one having the form

$$\mathbf{C} = \begin{bmatrix} c_0 & c_{N-1} & \cdots & c_1 \\ c_1 & c_0 & \cdots & c_2 \\ \vdots & \ddots & \ddots & \vdots \\ c_{N-1} & \cdots & c_1 & c_0 \end{bmatrix} \qquad (7)$$

where each column is a cyclic shift of the previous column. It is obvious that C is completely specified by its first column. So by $C(c)$ we denote the circulant matrix whose first column is c.

It is known that a circulant matrix can be diagonalized by the Fourier transformation matrix. Therefore, matrix-vector multiplication can be written as

$$C(c)v = ifft(fft(c) \circ fft(v)), \qquad (8)$$

where $ifft(.)$ and $fft(.)$ denote the IFFT and FFT transform, $\circ$ is the component wise product of two vectors. Based on that property, we can decouple the spatial mixing effect of multiple transmit antennas channel by combining circulant structure with IFFT/FFT.

First the symbol vector s of length N_t is transformed into $f = ifft(s)$ by IFFT. Then a circulant matrix $X = C(f)$ is constructed from f and transmitted through the channel h .

$$y = Xh + w \qquad (9)$$

Due to the circulant structure of $\mathbf{X}$, we have

$$y = ifft(fft(f) \circ fft(h)) + w \qquad (10)$$

$$= ifft(s \circ fft(h)) + w, \qquad (11)$$

So after the FFT operation on y, we get

$$z = fft(y) = s \circ fft(h) + fft(w), \qquad (12)$$

Similar to OFDM, every symbol of s fades according to one corresponding element of the FFT vector of the channel.

B. The Spreading Matrix and the Design Criterion

Unfortunately, when using the circulant structure with IFFT/FFT for transmission, the transmit diversity is lost. That is because every symbol only experiences one component of the FFT vector of the channel. To recover the diversity gain, we propose to spread the symbols before the IFFT transform, that is multiply the symbol vector by a spreading matrix $\mathbf{Q}$ giving

$$d = \rho \mathbf{Q} s, \qquad (13)$$

where ρ is the normalization coefficient that makes the total transmitting power over a time s lot equal to Es i.e. , $Et = Es$, which is the symbol energy for a given constellation. In Figure 1 we show a diagram of the proposed space-time transmission scheme for $N_t = 4$.

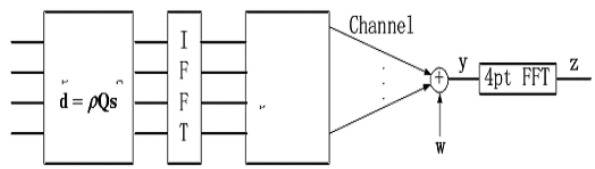


Figure 1 Block diagram of the proposed space-time transmission scheme for N_t=4

The transmission model between the symbol vector s and the received signal after the FFT can be expressed as follows:

$$z = \rho \, \mathrm{diag}(\tilde{h}) Q s + \tilde{w} \qquad (14)$$

$$= \rho \, \mathrm{diag}(Qs) \tilde{h} + \tilde{w} \qquad (15)$$

where $\tilde{h} = fft(h)$, $diag(\tilde{h})$ denotes the diagonal matrix with $\tilde{h}$ on its diagonal line and $\tilde{w} = fft(w)$.

The spreading matrix needs to be designed to recover the diversity loss. Due to the fact that $\tilde{h} = fft(h)$ is still an i.i.d. complex Gaussian random vector when h is i.i.d. complex Gaussian, and the Parseval's Theorem, the transmission model（14）will achieve full diversity when

$$\min_{s \neq v} \mathrm{rank}\{\mathrm{diag}(\mathbf{Q}(s - v))\} = N_t, \qquad (16)$$

Since the rank of a diagonal matrix is nothing but the number of nonzero entries along the diagonal, we have the following *full diversity design criterion* for the spreading matrix:

$$\left|\Delta_i^{(s,v)}\right| \neq 0, \forall i \in [1, \cdots, N_t], \forall s \neq v; s, v \in A^{N_t}, \qquad (17)$$

where $\Delta_i^{(s,v)}$ is the ith element of

$\Delta^{(s,v)} = \left[\Delta_1^{(s,v)},\cdots,\Delta_{N_t}^{(s,v)}\right]^T = Q(s-v)$.

In fact, this criterion is equivalent to the *nonzero minimum product distance criterion* in signal space diversity design problems [7], [8], [9] due to the fact that $\tilde{h}$ and h have the same statistical property. So precoding or constellation rotation matrix designs proposed for those problems can be applied as the spreading matrix achieving full diversity in our scheme.

Here, we use the unitary design in [] as the spreading matrix.

$$Q = \mathrm{F}_{N_t}^T diag(1,\alpha,\cdots,\alpha^{N_t-1}), \qquad (18)$$

where F_{N_t} is the FFT matrix with (m,n)st entry given by $N_t^{-1/2}\exp(j2\pi(m-1)(n-1)/N_t)$. When N_t is a power of 2, α is selected as $\alpha = \exp(j\pi/(2N_t))$. For any N_t which is not a power of 2, α is selected such that the minimum polynomial of α over the field $Q(j)$ has degree greater than or equal to N_t. For examples, when $N_t = 3$, α is selected to be $\exp(j\pi/9)$; when $N_t = 5$, α is selected to be $\exp(j\pi/25)$.

C. Symbol Detection at the Receiver

According to the transmission model Eq. 14 and assuming perfect channel knowledge, there are several methods for symbol detection at the receiver side. Maximum likelihood decoding can be employed to detect s optimally, but the detection complexity increases exponentially with N_t.

Alternatively, the sphere decoding algorithm [10] can be applied to achieve near-optimium performance. Due to the spreading nature of our scheme, the parallel interference cancellation (PIC) in CDMA [11] can also be used to do the detection. But because of error propogation, parallel interference cancellation has worse performance than sphere decoding in our simulations.

4 Simulation Results

In this section, we provide simulation results for the proposed scheme and compare it with the results for ideal OSTBC, the conventional circulant code [5], and the DAST code, which is also a full rate full diversity code. By ideal OSTBC, we mean the performance of a full rate full diversity complex OSTBC for a given N_t. Note that it is only ideal because practically no such design exists for $N_t > 2$.

In all simulations, we assume the channel to be quasistatic as previously mentioned and i.i.d. complex Gaussian with variance one. The total transmit power for every time slot is normalized to E_s which is the symbol energy for a given constellation.

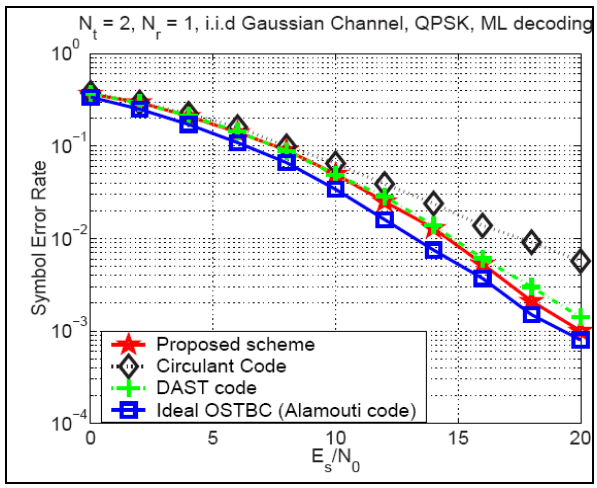


Figure 2 Symbol error rate vs E_s/N_0 for $N_t = 2$,QPSK

We first consider the case with $N_t = 2$ transmit antennas, one receive antenna, and QPSK constellation. The symbol error rate (SER) performances of the proposed scheme using ML decoding is shown in Figure 2 It was compared with ideal OSTBC which is just the Alamouti code in this setting. The performance of the DAST code and circulant space-time code is also plotted for comparison in Figure 2 .

The simulation result for $N_t = 4$; $N_r = 1$ and QPSK constellation is given in Figure 3 , where sphere decoding is used for detection. Figure 4 provides simulation results for $N_t = 3$; $N_r = 1$ and QPSK constellation. The results for $N_t = 4$ and 16QAM constellation is shown in Figure 5 .

In all these simulations our scheme is much better than the conventional circulant space-time codes and has a very close performance to ideal OSTBC. Our scheme also has similar performance as the DAST code.

5 Conclusion

In this paper, we proposed to combine the circulant matrix structure with IFFT/FFT and spreading for space-time transmission in multiple-antenna systems. The proposed scheme is rate one and can be easily applied to any number of transmit antenna. Our simulation results on QPSK and 16QAM constellations show that the performance of our scheme is close to the performance of ideal orthogonal space-time code and outperforms the conventional circulant space-time code.

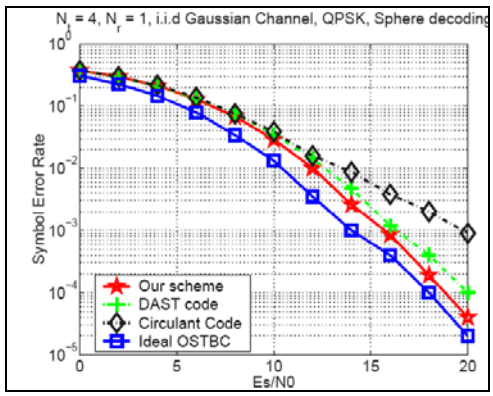


Figure 3 Symbol error rate vs E_s / N_0 for $N_t = 4$,QPSK

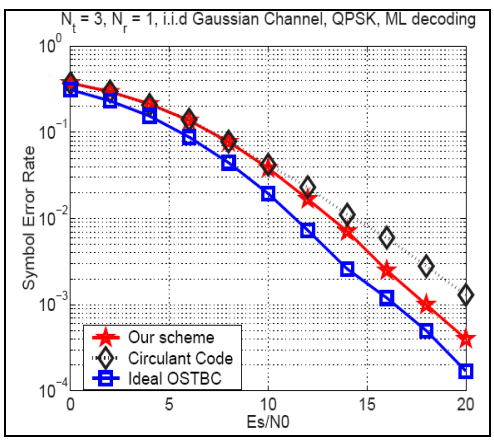


Figure 4 Symbol error rate vs E_s / N_0 for $N_t = 3$,QPSK

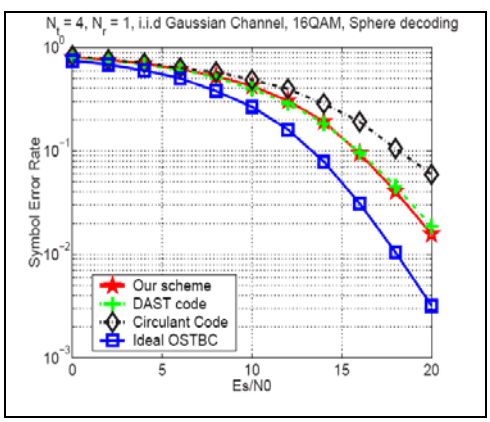


Figure 5 Symbol error rate vs E_s / N_0 for $N_t = 4$,16QAM

References

[1] S. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select Areas Commun.*, vol.16, no. 10, Oct. 1998,.pp. 1451–1458

[2] V. Tarokh, H. Jafarkhani, and A. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inform. Theory*, vol. 45, no. 10, Oct. 1999,pp. 1456–1467

[3] A. Hottinen, O. Tirkkonen, and R. WichmanGolub, *Multi-antenna Transceiver Techniques for 3G and Beyond*, John Wiley & Sons, New York, NY, 2003

[4] G. Raleigh and J. Cioffi, "Spatio-temporal coding for wireless communication," *IEEE Trans. Commun.*, vol. 46, no. 3, Mar. 1998, pp.357–366

[5] M. Zoltowski and M. Breinholt, "Space-time block codes using square hankel data blocks," in *Proc. IEEE VTC Fall*, 2001, vol. 1, pp. 372–374

[6] M. O. Damen, K. Abed-Meraim, and J. C. Belfiore, "Diagonal algebraic space-time block codes," *IEEE Trans. Inform. Theory*, vol. 48, no. 3, March 2002 , pp. 628–636

[7] X. Giraud, E. Boutillon, and J. C. Belfiore, "Algebraic tools to build modulation schemes for fading channels," *IEEE Trans. Inform. Theory*, vol. 43, no. 3, May 1997, pp. 938–952

[8] J. Boutros and E. Viterbo, "Signal space diversity: A power and bandwidth efficient diversity technique for the rayleigh fading channel," *IEEE Trans. Inform. Theory*, vol. 44, no. 4, July 1998,pp. 1453–1467

[9] Y. Xin, Z. Wang, and G. B. Giannakis, "Space-time diversity systems based on linear constellation precoding," *IEEE Trans. Wireless Commun.*, vol. 2, no. 3, March 2003, pp. 294–309

[10] B. Hochwald and S. Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Trans. on* Commun., vol. 51, no. 3, Mar. 2003,pp. 389–399

[11] M. Varanasi and B. Aazhang, "Multistage detection in synchronous code-division multiple-access communication," IEEE Trans. Commun., vol. 38, no. 4, Apr. 1990, pp. 509–519

Modeling of Glumatic Acid Fermentation Process Based on PSO-SVM

Xianfang Wang [1, 2] Zhiyong Du[3] Hua Wen[4] Feng Pan[2]

1 Henan Institute of Science and Technology, Xinxiang, Henan 453003, China

2 School of Information & Control Engineering, Jiangnan University, Wuxi, Jiangsu 214122, China

Email: wangxianfang@sina.com

3 Henan Mechanical and Electrical Engineering College, Xinxiang, Henan 453002,China

Email: zhiydu@126.com

4 Department of Electronic Science & Engineering, Huanghuai University, Zhumadian, Henan 463000, China

Abstract

In a fermentation process several variables, such as biomass concentration are conventionally determined by off-line laboratory analysis, i.e., the process control is unavailable to industrial production in time just because of time delay that often makes the analysis results inefficient. Utilizing the ability simple in application and quick in convergence of Particle Swarm Optimization (PSO) algorithm and the high generalization ability of Support Vector Machine(SVM), selecting the appropriate state variables, a dynamic time-varying model has been built. Using the model and algorithm to per-estimate some biochemical state variables which can not be measured on-line，and to optimize some operational variables .It is proved that the method is efficiency through the practical application of Glumatic Acid fermentation process.

Keywords： Particle Swarm Optimization, Support Vector Machine, State variables, fermentation process, Modeling

1 Introduction

Glumatic Acid fermentation process is complex and high non-linear, so single modeling method usually out of good modeling result by reason of lacking training data. This contribution addresses the application and comparison of model based estimation, optimization, and control methods for fed-batch bioprocesses. For the application of model based control, appropriate knowledge of the system's state is required[1,2].

Artificial Neural network (ANN)[3,4] is the most representative one in ferment process modeling methods. Because the theory of ANN is based on Empirical Risk Minimization (ERM), it has some disadvantages such as over fitting, local minimum, and some others questions. For solving this problem, a normal Support Vector Machine (SVM)is used[5,6], which based on Structure Risk Minimization (SRM), especially fits for classification, forecasting, estimation and fault diagnosis when samples are few, and then overcomes inherent disadvantages of ANN inherent disadvantages and greatly improves models' generalization.

In recent years, the Support Vector Machine (Support Vector Machine, SVM) for the non-linear process modeling are attracting increasing attention [7-9]. In theory, support vector machine is based on small samples, the overall situation will be the most advantage in the neural network solution can not be avoided in the local minimum value, from its topology support vector decision has avoided the traditional nerve network topology needs Copilot test determined, and it will also be able to arbitrary accuracy approaching arbitrary function. However, the realization of these advantages

depends on the correct choice of parameters, if the improper modeling results will be greatly affected. The choice of the majority of the current parameters of thumb is selected, are excessively dependent on the user's level, largely limiting its application.

Particle swarm optimization (PSO) is an evolutionary computation technique developed by Kennedy and Eberhart in 1995[10]. It is motivated by the behavior of organisms such as fishing schooling and bird flock. PSO is similar to a genetic algorithm in that the system is initialized with a population of random solutions. In PSO, instead of using more traditional genetic operators, each particle adjusts its "flying" according to its own flying experience and its companions "flying experience"[11-13]. The normal PSO algorithm is a validated evolutionary computation way of searching the extremum of function, which is simple in application and quick in convergence.

On the basis of an analysis of these parameters in SVM, Particle Swarm Optimization (PSO) algorithm is used to optimize them. Then selecting the appropriate state variables, a dynamic time-varying model has been built. It can be realized to per-estimate which some biochemical state variables can not be measured on-line. Through using in the practical application of Glumatic Acid fermentation process, the result shows that that the method is efficiency.

2 Support Vector Machine

2.1 The principles of SVM

SVM [14] is started from the simple case of two classes that is linearly separable. Its basic mathematical model can be expressed as the following formula:

$$y_i(w\cdot x_i + b) \geq 1 \qquad i = 1, 2, \cdots N \qquad (1)$$

where w is the normal direction of a separation plane, and b is a threshold.

If the sample data is not perfectly linear separable, some relaxation factors $(\xi_1, \cdots \xi_N)$ are introduced, where $\xi \geq 0$, and the set of the separation planes becomes:

$$q_k = \min\{p_{1k}, p_{2k}, p_{3k}, p_{4k}\} \qquad (2)$$

Therefore, the solution of the optimal separation plane in linear case that is not perfectly linear separable is to minimize the following formula:

$$\frac{1}{2}(w\cdot w) + C\sum_{i=1}^{N}\varepsilon_i \qquad (3)$$

where C is a penalty constant for those sample points mis-separated by the optimal separation plane. Its role is to strike a proper balance between the calculation complexity and the separating error.

If the sample set is non-linear, the optimal separation plane needs to be constructed anew. The basic idea is to use a non-linear transform to project the sample data into a high-dimensional feature space, and then to find the optimal separation plane in the high-dimensional feature space. The kernel functions are used for nonlinear transform.

The pattern recognition with SVM algorithm is briefly described as following, a more detailed description of SVM can be found in Ref. [15].

2.2 Parameter analysis in SVM

The parameters in SVM such as C, ε and the width coefficient γ in the kernel function K(x, xi) exert a considerable influence on the performance of SVM. Whether the value of C is too big or too small can reduce the generalization of SVM. The value of ε indicates the error expectation in the classification process of the sample data, and it affects the number of support vectors generated by the classifier, thereby affects the generalization error of the classifier. If the value of ε is too big, the separating error is high, the number of support vectors is small, and vice versa. The parameters in the kernel function reflect the characteristics of the training data, and they also affect the generalization of SVM. Therefore, only after the choice of all these parameters are correctly made, can the SVM achieve its best performance.

As to how to select these parameters correctly, there is not a most effective method up to now. Generally the cross verification trial or the gradient step-down operation is adopted, but these methods involve too many human factors or the function is required continuously differentiable, and the classifiers are susceptible to local minimum.

3 Optimized Parameters of SVM by PSO

3.1 Theory of particle swarm optimization (PSO)

In a standard PSO system, each particle flies in a D-dimensional space S. The velocity and location for the ith particle is represented as $v_i = (v_{i1}, v_{i2}, \cdots, v_{id})$ and $x_i = (x_{i1}, x_{i2}, \cdots, x_{id})$. The particles move according to the following equations:

$$v_{id} = w \cdot v_{id} + c_1 \cdot rand() \cdot (p_{id} - x_{id}) + c_2 \cdot Rand()(p_{gd} - x_{id}) \quad (4)$$

$$x_{id} = x_{id} + v_{id} \quad (5)$$

Where c_1 and c_2 are positive constants and rand () and Rand ( ) are two random functions in the range of [0, 1]. Parameter w is the inertia weight introduced to accelerate the convergence speed of the PSO[16]. Vector $p_i = (p_{i1}, p_{i2}, \cdots, p_{id})$ is the best previous position of particle i called pbest, 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 gbest.

3.2 Optimized parameters by PSO

Fitting sample data sets $\{x_i, y_i\}(i = 1, 2, \cdots, n)$; $x_i \in R^d; y_i \in R)$ by SVM could obtain the regression function:

$$\begin{aligned} f(x) &= w \cdot \phi(x) + b \\ &= \sum_{i=1}^{n} (\hat{a} - a_i)(\varphi(x_i) \cdot \varphi(x)) + b^* \\ &= \sum_{i=1}^{n} (\hat{a} - a_i) k(x_i, x) + b^* \end{aligned} \quad (6)$$

Where $\alpha_i, \hat{\alpha}_i$ are Lagrange operators; b^* is threshold;

$k(x_i, y_j) = \exp(-|x_i - x_j|^2 / 2\gamma^2)$ is the kernel function.

For PSO-SVM, the location and speed of every particle are decided by 3D parametric ε, c, γ, The MSE(Mean Square Error) is selected as fitness function which could reflect directly the performance of SVM.

$$MSE = (\sum_{i=1}^{n} (\hat{y} - y_i) / n)^{1/2} \quad (7)$$

The detailed steps of the optimization are as follows:

Step1 Initialized ε, c, γ of the PSO, determined population size m, the maximum of the algorithm $iter_{\max}$, set the minimum and the maximum values of the weights $\omega_{\max}, \omega_{\min}$.

Note: The approximate range of these parameters is offered by Ref. [12] to avoid the blindness of initializing parameters: $\varepsilon = [0, 0.2], c = [1, 10^8], \gamma = [0.01, 2.0]$

Step2 Settings each individual particle extreme for the current position, using the function（6）,（7）to calculate the adaptation of each particle, adopting the best adaptation of the individual particles as an initial value of the extreme overall situation g_{best};

Step3 According to formula （4）~（5）, calculating iteratively and update the position, speed of the particle.

Step4 Comparing the value adaptation of each particle with the corresponding value of p_{ibest}, if gifted, updating p_{ibest}, otherwise retain actual value.

Step 5 Comparing the updated p_{ibest} with the value of the extreme overall situation p_{best}, if gifted, updating p_{best}, otherwise retain actual value;

Step6 Determining whether the conditions of termination is fit, if the maximum number of iterations is arrived, or no change in the iteration, then the process is end, otherwise return to Step 3.

4 Fermentation Modeling Based on PSO–SVM

When we don't know any prior knowledge, support vector machine (SVM) modeling, which has a good predicting result, it is a black-box method. The traditional kinetics reflects the mechanism of the process [13], while it has some error between the traditional kinetics model's predicting result and the real output. Fermentation process is complex and high non-linear, so single modeling method usually out of good modeling result by reason of lacking training data. It is a good idea to model with improved SVM and kinetics model

for fermentation process. Here, PSO-SVM is adopted to model the process.

There are 10 batches of production data, each batch express a complete fermentative process. Taking 7 batches of data takes as the training data, each of them is divided 18 samples. For the reflection fermentative process's misalignment relations, use the radial direction base nuclear function, Therefore the SVM parameter mainly has the insensitive coefficient ε, penalty coefficient C and the breadth factor σ . Uses various sample points relative error's average value (the mean error ratio) the computation model training error and the prediction error.

These parameters are selected by PSO, after the training. Optimized parameters are as follows:

σ =1、C =100、ε =0.4

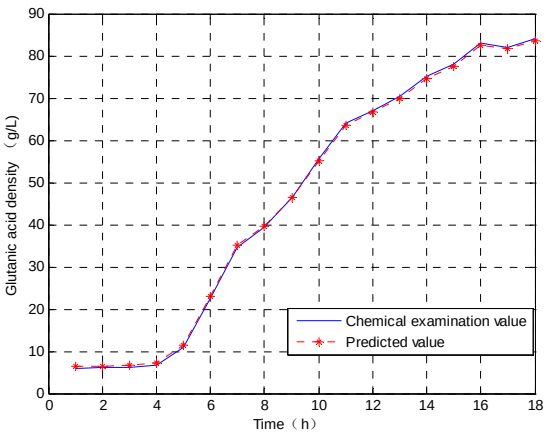


Figure 1 The result of the training

One of these batches of data training's result like Figure 1, the training relative error's average value is 0.1508. With trains the good machine to estimate that 2nd batch of data's simulation result like Figure 2 , the forecast relative error's average value is 0.1498. May see the modeling effect from Figure 2 to be good, the PSOSVM method establishes the model is higher than in the training data in test data performance the performance.

5 Comparison with an Modeling Method

An artificial neural network (ANN), often just called

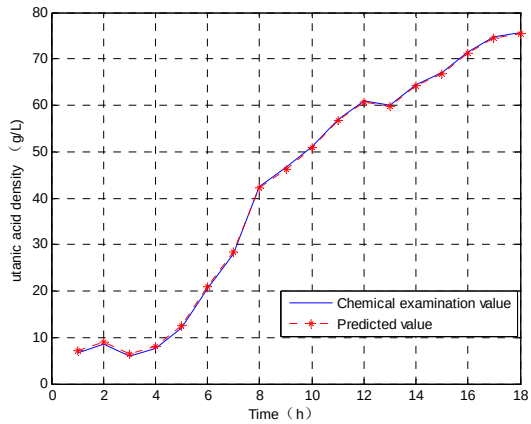


Figure 2 The result of the test

a "neural network" (NN), is a mathematical model or computational model based on biological neural networks. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. RBF network isone class of ANN, which has the advantage of not suffering from local minima in the same way as Multi-Layer Perceptions. RBF networks have the disadvantage of requiring good coverage of the input space by radial basis functions. Associating each input datum with an RBF leads naturally to kernel methods such as Support Vector Machines and Gaussian Processes (the RBF is the kernel function). All three approaches use a non-linear kernel function to project the input data into a space where the learning problem can be solved using a linear model. Like Gaussian Processes, and unlike SVMs, RBF networks are typically trained in a Maximum Likelihood framework by maximizing the probability (minimizing the error) of the data under the model. SVMs take a different approach to avoiding overfitting by maximizing instead a margin. RBF networks are outperformed in most classification applications by SVMs. In regression applications they can be competitive when the dimensionality of the input space is relatively small.

In order to confirm the validity that is based on PSO-SVM, Comparison with ANN modeling method. We built two models:

The RBF nuclear function is adopted in SVM, Optimized parameters by PSO are as follows:

σ=1、C=100、ε=0.4;

ANN neural network error target e=0.01, breadth factor σ=0.5.

From the Table 1, it is clearly show that two kind of modeling method's training error reduce gradually along with sample number's increase, but the prediction error reduces gradually besides PSO-SVM, the phenomenon which after other two methods presented reduces first, increases, explained that PSO-SVM the pan-ability is better than the ANN neural network. This is as a result of the neural network method based on the experience risk minimum principle, the application error anti-propagation method causes its experience risk to be smallest, but the experience risk is smallest cannot guarantee that the real risk is smallest, therefore the neural network will present "the fitting" the phenomenon, namely may approach the sample collection by the random precision, but the pan-ability will not be ideal. But PSO-SVM based on the structure risk minimum principle, its objective function contains the experience risk and the confidence interval two targets, the two decide SVM together the real risk. Therefore, the PSO-SVM pan-ability is better than the neural network method. Is obvious from the sample level of dependency analysis, when sample number change, the SVM pan-error change scope is smaller than the neural network, explained that PSO-SVM is smaller than to sampled data's level of dependency the neural network.

Table 1 Comparison of errors between two methods

Trainning number	Testing number	PSO-SVM		ANN	
		Training error	Predicting error	Training error	Predicting error
13	56	0.031357	0.070653	0.041839	0.050
19	56	0.042315	0.056318	0.038654	0.043
28	56	0.021508	0.040128	0.032958	0.038
37	56	0.020316	0.033516	0.030614	0.029
55	56	0.019809	0.027605	0.028915	0.028

6 Conclusions

SVM is a valuable pattern-recognition method in theory and in application. In order to improve the performance of SVM and reduce the influence of human factors in real application, PSO algorithm is used to optimize the parameters in SVM in this paper on the basis of performance analysis. Utilizing the ability simple in application and quick in convergence of Particle Swarm Optimization (PSO) algorithm and the high generalization ability of Support Vector Machine (SVM), selecting the appropriate state variables, a dynamic time-varying model has been built. Using the model and algorithm to per-estimate some biochemical state variables which can not be measured on-line，and to optimize some operational variables .It is proved that the method is efficiency through the practical application of Glumatic Acid fermentation process.

References

[1] Bailey J, Ollis D. Biochemical fundamental. New York: McGraw-Hill Inc[M],1986:1-30

[2] Shi Zhongping, PAN Feng, Fermentation Process resolution, control and detection[M], Beijing: Chemical Industry press, 2005:1-10(in Chinese)

[3] Glassey, J., Montague, G. A., Ward, A. C., and Kara, B. V.: Artificial neural network-based experimental design procedures for enhancing fermentation development. Biotechnol Bioeng., 44, 397-405 (1994)

[4] Warnes M R, Glassey J, Montague G A.Application of radial basis function and feedforward artificial neural networks to the Escherichia coli fermentation process [J]. Neurocomputing, 1998, 20（1）: 67-82

[5] Feng Rui ZHANG Yue-Jie ZHANG Yan-Zhu SHAO Hui-He. Drifting Modeling Method Using Weighted Support Vector Machines With Application to Soft Sensor [J]. ACTA AUTOMATICA SINCA, 2004, 3(30):436-441

[6] Vapnik V, Golowich S, Smola A. Support vector machine for function approximation, regression estimation, and signal processing [A]. In: Mozer M, Petsche T(eds). Neural Information Processing Systems [M], MIT Press, 1997, 9:1-200

[7] GAO Xue-jin, WANG Pu, SUN Chong-zheng, etl. Modeling for Penicillin Fermentation Process Based on Support Vector Machine Journal of System Simulation[J] 2006.07:2052-2055(in Chinese)

[8] Feng Rui ZHANG Yue-Jie ZHANG Yan-Zhu SHAO Hui-He. Drifting Modeling Method Using Weighted

Support Vector Machines With Application to Soft Sensor [J]. ACTA AUTOMATICA SINCA, 2004, 3(30):436-441

[9] LIU Yi and WANG Hai-Qing. Modeling a Penicillin Fed-batch Fermentation Using Least Squares Support Vector Machines [J]). Chinese Journal of Biotechnology. 2006,(01):144-148(in Chinese)

[10] JKennedy, R. C. Eberhart. Particle Swarm Optimization[C]. Proc. IEEE International Conference on Neural Networks, Piscataway, NJ: IEEE Service Center, 1995, IV: 1942-1948

[11]]R. Eberhart, Y. Shi. Particle Swarm optimization: development, applications and resource [J]. IEEE Int conf on evolutionary Computation ,2001:81-86

[12] Angeline, P. Evolutionary optimization versus particle swarm Philosophy and Performance Differences[C]. In Evolutionary Programming VII, 1998:601-610

[13] F.Van den Bergh, A.P.Engelbrecht. A new locally convergent particle swarm optimizer[C]. IEEE International conference on systems, man and cybernetics, 2002:96-101

[14] Vapnik V, Golowich S, Smola A. Support vector machine for function approximation, regression estimation, and signal processing [A]. In: Mozer M, Petsche T(eds). Neural Information Processing Systems [M], MIT Press, 1997, 9:1-200

[15] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1999

[16] Y. Shi, R. C. Eberhart. A Modified Particle Swarm optimizer. Proc. 1998 IEEE International Conference on Evolutionary Computation, 1998: 1945-1950

A Novel Method to Generate UWB Shape Impulse

Bo Hu[1] Hongxin Yang[2]

1 Department of Physics and Electronics, Bin Zhou University, Binzhou, Shandong, 256603, China

Email: hubobz@163.com

2 E&E, Beijing University of Posts and Telecommunications, 100876, China

Abstract

The impulse waveform design is an essential factor to the UWB system performance. A new method to generate orthogonal UWB shape impulses is proposed based on the eigenvector of Hermite matrix and Chirp signal, and also the desired shape impulses have been achieved. The simulation results show that the power spectral density distribution of the short time impulses generated by this method is consistent to the FCC spectral mask of UWB. The UWB shape impulses given here have many characteristics such as high spectral occupation rate, high autocorrelation and zero cross-correlation, which is beneficial to the multi-access detection.

Keywords: UWB (Ultra Wide Band); shape impulse; PSD (Power Spectral Density); Chirp signal

1 Introduction

The proposition to the concept of Wireless Internet, wireless LAN and wireless personal area network requires wireless data transmission speeds up to 100 Mbps in the indoor environment to build personal information network system. To achieve such a high rate of wireless transmission in the indoor environment, UWB (Ultra Wide Band) technology is a very competitive way and in recent years it has attracted extensive research and attention in academia and industrial circles [1-5]. In order to promote and regulate the development of UWB technology, in April 2002 the United States Federal Communications Commission (FCC) issued a preliminary rule of UWB wireless devices, and redefined the UWB, namely UWB signal bandwidth should be more than or equivalent to 500 MHz or the relative bandwidth is more than 0.2. As can be seen, UWB has not only confined to the initial pulse communications, but includes all forms of communication using wide spectrum. FCC regulates the actual use of the spectrum of UWB in indoor communication is from 3.1 to 10.6 GHz, and in this range, the effective isotropic radiated power is not more than 41.3 dBm / MHz [6]. The main form of UWB signal is base band narrow pulse, using the width of nanoseconds, sub-nanoseconds and base-band narrow pulse sequence to communicate. It usually carries the information via pulse position modulation (PPM), the pulse polarity or pulse amplitude modulation (PAM). Narrow pulse can use a variety of different waveforms, such as Gaussian waveform or cosine wave form. In narrow-band pulse UWB communication, because of the narrow pulse width and smaller duty cycle under normal circumstances, it has good multipath channel resolution and anti-multipath performance. And because it doesn't need modulate carrier wave, the transceiver machine has simple structure and lower cost which make the system low power consumption [2]. However, baseband narrow pulse contains more low-frequency components, so under the provisions of the FCC on UWB communications power spectrum, the spectrum utilization is not effective, which can be improved through the pulse waveform optimized design [7, 8]. The pulse waveforms should be smooth in the time domain and the power spectral density should be easily controlled by the time domain parameter [7]. Traditional Gaussian waveform or cosine waveform cannot guarantee the orthogonality between waveforms, thereby

there is a serious interference between users at the receive end, resulting in increased bit error rate and the impact on system performance. UWB was initially applied to radar and other military communications, yet Chirp (LFM) signals are widely used to the radar systems due to its characteristics of bigger time-bandwidth product, convenient filter-matching test and easy generation. Because of its excellent multi-site detection performance, the research on Chirp signal as UWB shaping pulse has been continuously carried out [9, 10]. On the basis of Chirp signal and Hermite matrix eigenvector method, this paper depicts the orthogonal UWB waveform, which can fundamentally reduce error rate and interference between users. Also, this paper [10] notes that the eigenvector Waveform Design has more advantages than Hermite polynomial structure waveform [12] such as the short pulse duration, the high data rate and the low side lobe spectrum.

2 UWB Impulse Forming Algorithm

In the design of UWB impulse waveform, the impulse frequency spectrum is ranged from 3.1 to 10.6 GHz band according to the FCC regulations, and its radiation should make the biggest spectrum utilization as far as possible under the circumstances of meeting FCC spectrum constraints. Supposing a shaped impulse signal is ψ (t), the spectrum should be distributed in the spectrum limit range of FCC regulations after it passes the following system: the impact response is h (t) and the frequency response is H (f). Defining the impulse width of shaped impulse signal is Tm, then

$$\psi(t)=\begin{cases} p(t), t<\left|T_m/2\right| \\ 0, t\geq\left|T_m/2\right| \end{cases} \quad (1)$$

If we make $H(f)\phi(f)=\lambda\phi(f),$ (2)

After attenuation, the frequency spectrum of impulse signal can be limited to the ranges of the FCC [10], of which λ is an attenuation factor. Supposing pulse sampling point is N, the discrete time domain convolution form corresponding formula (2) will be as follows:

$$\sum_{m=-N/2}^{N/2}\phi[m]h[n-m]=\lambda\phi[n], n=-N/2,\cdots,N/2 \quad (3)$$

It is shown by matrix form:

$$\begin{bmatrix} h[0] & h[-1] & \cdots & h[-N] \\ h[1] & h[0] & \cdots & h[-N+1] \\ \vdots & \vdots & & \vdots \\ h[\frac{N}{2}] & h[\frac{N}{2}-1] & \cdots & h[-\frac{N}{2}] \\ \vdots & \vdots & & \vdots \\ h[N] & h[N-1] & \cdots & h[0] \end{bmatrix} \begin{bmatrix} \phi[-\frac{N}{2}] \\ \phi[-\frac{N}{2}+1] \\ \vdots \\ \phi[0] \\ \vdots \\ \phi[\frac{N}{2}] \end{bmatrix} = \lambda \begin{bmatrix} \phi[-\frac{N}{2}] \\ \phi[-\frac{N}{2}+1] \\ \vdots \\ \phi[0] \\ \vdots \\ \phi[\frac{N}{2}] \end{bmatrix} \quad (4)$$

Namely:

$$H\phi=\lambda\phi \quad (5)$$

As can be seen, matrix H is Hermite matrix, its required shaped impulse ψ and attenuation factor λ can be obtained by the eigenvector and eigenvalue of matrix H [10]. As H is Hermite matrix, the obtained eigenvector group is non-linear orthogonal vector group and the eigenvalues are real numbers, therefore the shaped impulses are irrelevant. Obviously, such shaped impulse can reduce the error rate and channel interference between users at the receiving end, which improves the system performance. Due to Chirp signal's characteristics of big time-bandwidth product and low sidelobe [9,10], we can consider it as the impulse response h(t) and suppose the corresponding bandwidth is from 3.1 to 10.6 GHz in order to make it a function of the shaped impulse, then its time domain can be expressed in the following form:

$$h(t)=\sin(2\pi f_0 t+kt^2), \quad (6)$$

$$k=\pi(f_u-f_l)/T_m, f_u=10.6\text{GHz}, f_l=3.1\text{GHz}.$$

3 Simulation Results

As is shown above, Chirp signal is firstly generated by formula （6）and then acted as the impulse response of the system; next, Hermite matrix H is constructed on the basis of formula （4）and formula （5）; thus, the corresponding eigenvector ψ can be calculated and the required shape impulse can be obtained by eigenvector. Supposing impulse sampling points N = 64, impulse duration Tm = 1 ns, shape impulse simulation results can be shown in Figure 1. Figure 1 (a) and (b) respectively represent the shape impulses ψ1 (t) and ψ2 (t), corresponding to the two largest eigenvalue; Figure 2 (a) and (b) respectively represent the power spectral density distribution, corresponding to the two shape impulses. From Figure 2, we can see that the power spectrum of UWB shape impulse is within the range of the FCC, and it has higher utilization of the spectrum.

Figure 3(a) and (b) respectively represent the autocorrelation function of ψ1 (t) and the cross-correlation function of ψ1 (t) and ψ2 (t). It is shown that where the auto-correlation function gets the maximum value, cross-correlation function is zero, which is very favorable to eliminate interference between users and improve multi-user detection.

4 Conclusion

This paper investigated a novel method to generate orthogonal UWB shape impulses based on the eigenvector of Hermite matrix and Chirp signal, by which the orthogonal instantaneous shape impulse sequence can be achieved and its spectral density is consistent with the FCC regulations. And the UWB shape impulses given here have the characteristics of high spectral occupation rate, high autocorrelation, zero cross-correlation and small interference between users which is beneficial to the multi-access detection. The simulation results have proved the feasibility of the new method and it can be directly used to UWB signal transmission system due to its simple generation process.

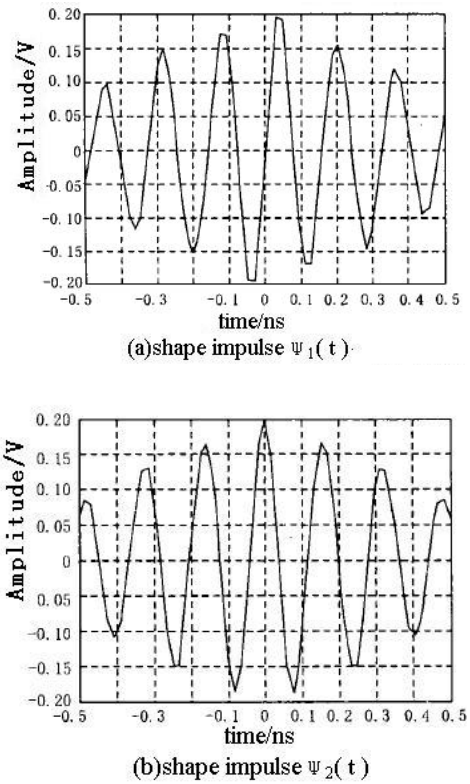


Figure 1 UWB shape impulse

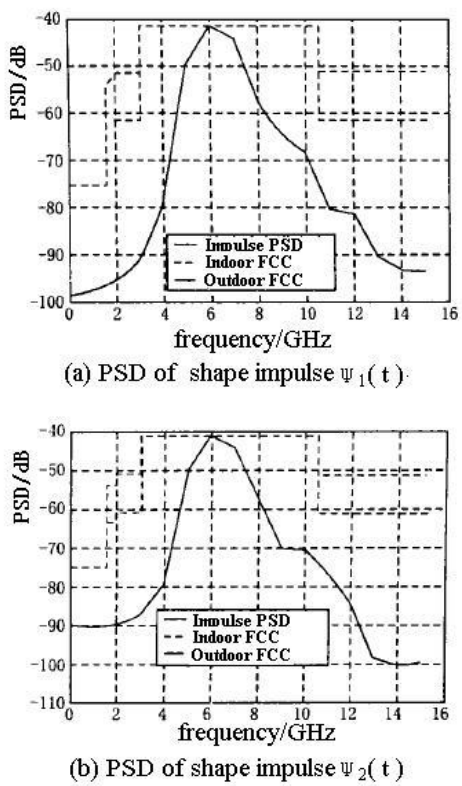


Figure 2 Power Spectral Density of UWB shape impulse

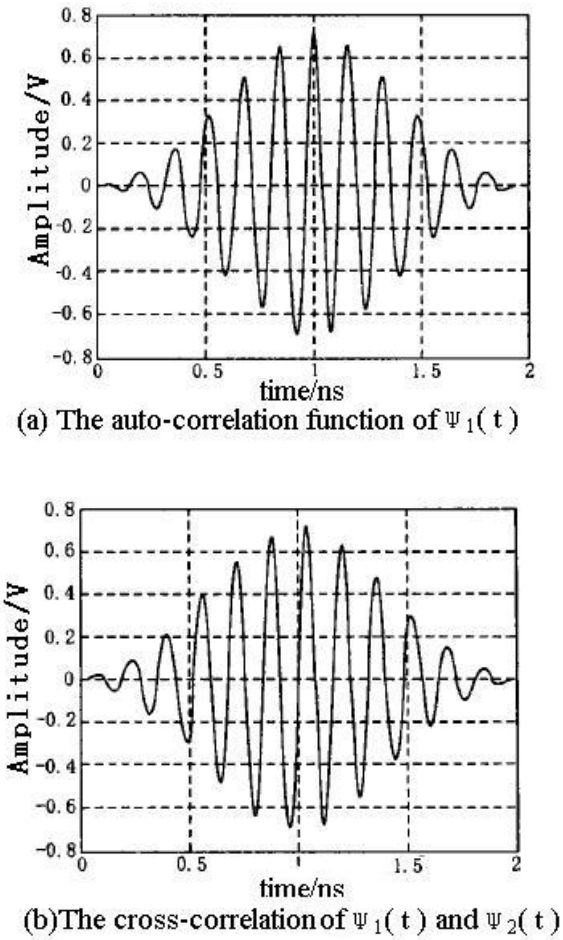


Figure 3 The correlation function of UWB shape impulse

References

[1] B. Parr, B. Cho, K. Wallace, Z Ding, "A novel ultra-wide band pulse design algorithm," IEEE Com-mu.Lett, 2003,7(5), pp.219~ 221

[2] C. A. Corral, S. Sibecas, S. Emami, G. "Stratis. Pulse spectrum optimization for ultra-wideband communication," IEEE Conf.on on Ultra Wideband Systems and Tech, 2002, 5, pp.31~ 35

[3] G. Roberto, G. Rogerson, Ultra-wideband wireless systems. IEEE microwave magazine, 2003, 7, pp.36~ 47

[4] J. Han, C. Nguyen. "A new ultra-wideband, ultra-short monocycle pulse generator with reduced ringing," IEEE Microwave Wireless Compon.Lett, 2002, 12(7), pp.206 208

[5] L. B. Michael, M. Ghavami, R. Kohno, "Multiple pulse generator for ultra-wideband communication using Hermite polynomial based orthogonal pulses," IEEE Conf.on Ultra Wide-band Systems and Tech, 2002, 5, pp. 73 ~77

[6] L. B. Michael, M. Ghavami, "Multiple pulse generator for ultra-wideband communicationusing Hermite polynomial based orthogonal pulses," IEEE Conf.on Ultra Wide-band Systems and Tech, 2002, 5, pp.47~ 51

[7] M. Z. Scholtz, Impulse radio: How it works, IEEE Commun.Lett, 1998 (2), pp.36 ~38

[8] M. Kowatsch, J. Lafferl, "A Spread-Spectrum Concept Combining Chirp Modulation and Pseud-onoise Coding," IEEE Trans on communications, 1983, 31 (10), pp.1133~ 1142

[9] R. C Qiu, "A study of the ultra-wideband wireless propagation channel and optimum UWB receive design," IEEE Journal on Selected Areas in Communications, 2002, 20 (9), pp.1628~1637

[10] R. A. Scholtz, Multiple access with time-hopping impulse modulation, Proceedings of MIL-COM93, 1993, 10, pp.447~450

[11] S. Hengstler, D. P. Kasilingam, A. H. Costa, "A Novel Chirp Modulation Spread Spectrum Techniqufor Multiple Access," IEEE 7thint.Symp.On Spread-Spectrum Tech & Appl, Prague, CzechRepublic, 2002, 9, pp. 73 ~77

[12] W. Chu, C. Colbourn, "Sequence designs for ultra-wideband impulse radio with optimal correlation properties," IEEE trans. on information theory, 2004, 50 (10), pp.2402 ~ 2407

Research on Peer-to-Peer Media Streaming Systems

Wei Shi

School of Information Technology, Jiangnan University, Wuxi, Jiangsu 214122, China
Email: awei88866@126.com

Abstract

A conventional solution for media streaming system on the Internet is the client-server service system. With the widespread penetration of broadband accesses and the development of peer-to-peer (P2P) technology, P2P technology has been employed to provide media streaming services. Recently, the research on P2P media streaming systems has drawn a lot of attentions from both academy and industry. Various P2P media streaming algorithms have been studied, and the systems have been developed. In this paper, some representative P2P media streaming systems are compared and analyzed by several key designs in systems, including system topologies, node connections and data scheduler. This paper is concluded in the end.

Keywords: P2P, media streaming, tree-based systems, mesh-based systems

1 Introduction

With the widespread penetration of broadband accesses and the development of P2P technology, researches on P2P media streaming systems [1, 2] have recently drawn people's attention.

A conventional solution for media streaming system over the Internet is the client-server service system. A client builds a connection with a media source server and content is downloaded directly from the source server. As the client population increases gradually, the bandwidth provision must grow proportionally. Obviously this solution suffers from the bottleneck at the source server.

In P2P media streaming systems, P2P technology is employed, which is a new apotheosis to build distributed network applications. In this system, users, namely as nodes, act as both clients and servers at the same time. That is a node not only downloads data from other nodes, but also uploads the downloaded data to other nodes in the system. The uploading bandwidth is efficiently utilized to reduce the bandwidth burden otherwise placed on the servers. P2P media streaming systems greatly reduce the cost of running the source server, and make large scale multimedia data delivery with low server cost feasible.

In the rest of the paper, the P2P media streaming systems are described in Section 2. Subsequently, some representative P2P media streaming systems are compared and analyzed in Section 3. Finally, the paper is concluded in Section 4.

2 P2P Media Sreaming Systems

Conventional media streaming systems are composed of six parts: media encoder and other tools, peripheral equipment, network, media database, media server, media player. (see Figure 1)

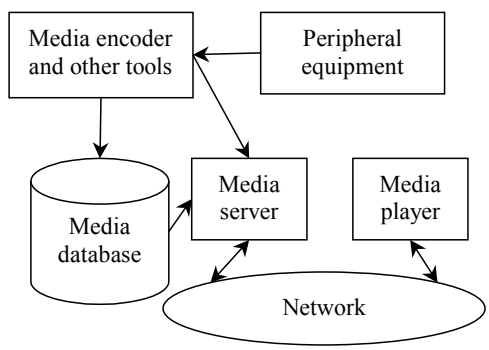


Figure 1 Media streaming system

P2P media streaming systems achieve the media streaming transport based on P2P networks. The

so-called P2P network refers to make use of the P2P computation pattern building a logical cover network which is based on application-layer multicast [3, 4], also known as P2P overlay network. This network determines the topology and management of connecting nodes. P2P media streaming systems can provide favorable media services for the upper layer by P2P overlay network. In this system, the source sever does distribute data not to all nodes by unicast, but to partial nodes. And then these nodes provide services to other nodes. From the above analysis we can see that the architecture of the conventional media streaming system isn't changed in P2P media streaming systems. Only the original service and transport way has been changed.

3 Comparison and Analysis

Nowadays, P2P streaming systems can be broadly classified into categories based on the overlay network topology. They are tree-based and mesh-based [5, 6]. In this section, some representative P2P media streaming systems will be compared and analyzed according to this classification.

3.1 Tree-based Systems

Tree-based systems have well-organized overlay structures and typically distribute media content by actively pushing data from a node to its children nodes. Tree-based systems include single-tree and multi-tree systems.

3.2 Single-tree Systems

Single-tree systems are a single multicast tree that's composed of all nodes in logic. (see Figure 2) In this system, a node has only one parent node and downloads all content of the media stream from the parent node. Therefore, a node departure will temporarily disrupt data delivery to all nodes in the sub-tree rooted at the departed node. An exemplification of single-tree systems is PeerCast [7].

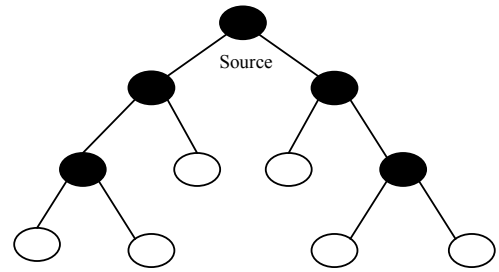


Figure 2 A simple single-tree systems

PeerCast

A media stream is a time ordered sequence of packets that is logically composed of two channels: data (served using unreliable RTP/UDP) and control (sent using reliable RTSP/TC) channels. In PeerCast, peer nodes are organized into application-layer multicast trees (see Figure 3), with each data packet being diffused at the same structure. When a peer node receives a data packet, it also forwards copies of the packet to each of its children. Because of the unpredictability of nodes' behavior, when an application-level multicast tree is built, the question that nodes joins and leaves should be solved.

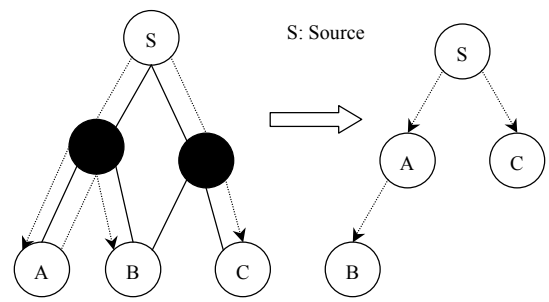


Figure 3 An application-level multicast tree

Node Join

At first, the node trying to join sends a setup request to the server. If the server is unsaturated, it accepts the request. If the server is saturated, it forwards the request to one of immediate children. The child decides whether to accept the request according to own resources estimate of the situation. This course will be completed until the node joins.

A node which is unsaturated always accepts a setup request. However, a saturated node N needs to forward the request R to another node in the network. Since a node only knows its local topology, N can only forward

R to one of N's immediate children, or its parent. Some of route selection strategies are the following:

a. Random: N chooses one of its children at random as the target T, and forwards R to T. Such a strategy requires little information of state at N. On an average, the tree is expected to be balanced.

b. Round-Robin: N sets up a list of its all children, and forwards R to the child C at the first of the list. And then the child C is then moved to the last of the list. Such a strategy requires some information of state maintenance at N, but is expected to keep the tree balanced.

c. Smart-Placement: Each node maintains the network locations of its children. R is sent to N. N forwards R to a child that has least access latency to the node bringing forward R. Such a strategy helps in creating trees taking network proximity into physical topology. Packet losses and delays are expected to be minimized.

d. Smart-Bandwidth: As the name suggests the difference of between the strategy c and d lies in merely the criterion distinguishing least access latency. The former is "placement", but the latter is "bandwidth".

Node Leave

When a node N leaves, all of its descendants will be lost and need to contract a new suitable valid target T to recover. Each node is definitely aware of two nodes in the network: its parent P and the source S. Therefore, there are two models for the section of T. Firstly, each descendant can try to recover by contacting T. Secondly, only the children C of N attempt to recover by contacting T. The rest of the descendants of C automatically recover when C contact T successfully. Hence, we have the following strategies theoretically:

a. Root-All: N chooses S as T. Starting from N, a redirect message is recursively forwarded to all the descendants of N specifying S as T.

b. Grandfather-All: N chooses P (which is the grandfather of C) as T. As in the a strategy, all the descendants of N are recursively redirected to P.

c. Root: N chooses S as T. Only nodes in C attempt to recover by contacting T. The rest of the descendants rely on C.

d. Grandfather: N chooses P as T. Only nodes in C attempt to recover by contacting t.

The strategy advantage of choosing S is that the tree depth could be decreased, which reduces the loss rate of packets and transport delay. The strategy advantage of choosing P is that the effects of failures are localized. Note that the policy to recover from failure is similar to leave, once a failure of the parent is detected. However, in this case, the identity of the parent of the failed node is not known to the descendants of the node. Thus, only the Root and Root-All strategies are relevant.

3.3 Multi-tree systems

Single-tree systems are inherently not well matched to a cooperative environment. The reason is that the small subsets of interior nodes carry all the burden of forwarding multicast messages. All the leaf nodes don't contribute their uploading bandwidth. This conflicts with the expectation that all nodes should share the burden.

To address this problem, a multi-tree system is beneficial, which constructs a forest of multicast trees. In the forest, each tree takes the source node as the root. (see Figure 4) In a multi-tree system, the forwarding load is distributed subject to the bandwidth constraints of the participating nodes in a decentralized, scalable, efficient and self-organizing manner. SplitStream [8] is arguably an example of the most typical multi-tree systems nowadays.

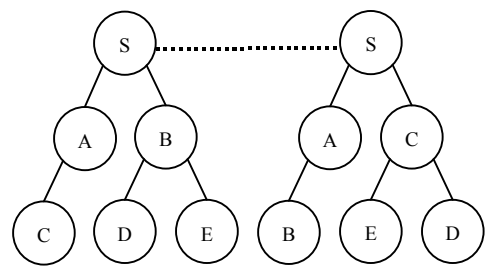


Figure 4 A simple multi-tree system

SplitStream

SplitStream depends on a structured peer-to-peer overlay network called Pastry, and on Scribe, an application-level multicast system set up this overlay.

The key idea of SplitStream is to split the multicast content into k stripes, and use separate multicast trees to distribute each stripe. These trees form a forest of multicast trees. In this way, the multicast content can be spread across all participating nodes.

Figure 5 illustrates a basic approach of SplitStream. In this simple example, the original content is split into two stripes. An independent multicast tree is constructed for each stripe. For simplicity, let us assume that the original content has a bandwidth requirement of BW, and that each stripe has half the bandwidth requirement of the original content. As shown in Figure 5, a node is an interior node in one multicast tree and a leaf in the other and forwards the stripe to two children, yielding a bandwidth requirement of no more than BW. Notice that node D is a leaf node in both sub-trees and doesn't contribute to video uploading. This is because the six contributes one unit of bandwidth and only five units of node uploading bandwidth are needed to stream to six nodes.

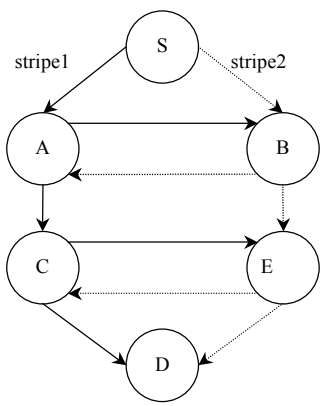


Figure 5 A simple example illustrating the basic approach of SplitStream.

The key challenge in the design of SplitStream is to efficiently construct a forest of multicast trees such that an interior node in one tree is a leaf node in all the remaining trees and different capacity limits of individual participating nodes are satisfied.

Node Join

In Pastry, a nodeId refers to nodes assigned random identifiers from a large sparse id space. In Scribe，a pseudo-random Pastry key, known as the groupId, is chosen for each multicast group. A multicast tree related with the group is built by the union of the Pastry routes from each group member to the groupId's root (which is also the root of the multicast tree). Messages are multicast from the root to the members using reverse path forwarding. Each node is able to establish a group. Each group have a only groupId, which includes a nodeID and a rendezvous (which is the nearest node to the nodeId).

When a node sends its join request to the group whose groupId is 0001, the request will be forwarded to the rendezvous of 0001 by the Pastry routes. Group tabulations of each node in the route path are examined to ensure whether it is the transponder of 0001. If one node is right, the node wanting to join will be accepted as its child , and added to its children list. Otherwise, this node will firstly be changed the transponder of 0001 by sending a joining message to the next node in the route path. Then, the rest is the same as the former situation.

Node Leave

Before a node leaves, its parent node will be found in the route list. And then the node sends its leave request to the parent node. When the parent node accepted the request, it changes itself into a child node in the route list. The node leaves successfully.

3.4 Mesh-based Systems

In tree-based systems, nodes are confined to a static topology. If a node's parent leaves, the node, as well as its descendants, cannot get streaming feed until it connects to another parent. To address this problem, many recent P2P streaming systems introduce mesh-based streaming approach. In a mesh-based system, the relationship between nodes is established and terminated dynamically based on the content availability and bandwidth availability on nodes. At any given time, the relationship is maintained with multiple neighboring nodes. A node may download/upload media from/to multiple neighbors simultaneously. If a node's neighbor leaves, the node can still download media file from remaining neighbors. At the same time, the node will find new neighbors to keep a desired level of

connectivity. The high peering degree in Mesh-based streaming systems makes them extremely robust against the damage brought about by nodes departure. Recently, DONet [9], a Data-driven Overlay Network, is designed by introducing a simple and straightforward data-driven design.

DONet

DONet employs a gossiping protocol [10], which has become popular solutions to multicast message dissemination in peer-to-peer systems. In DONet, there are three key modules: membership manager, partnership manager and data scheduler. (see Figure 6) A node always forward data to others that are expecting the data. Such a design is suitable for overlay with high dynamic nodes.

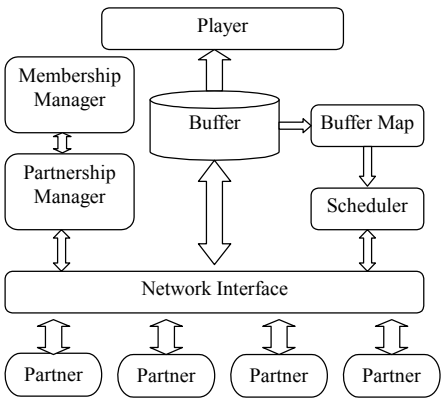


Figure 6 A system diagram for a DONet node

An example of the partnership in DONet is shown in Figure 7. Each DONet node must maintain the information of partnership nodes to select one of partnership nodes to fetch data. Hence, each node has a unique identifier, such as its IP address. And a list of identifiers for other partnership nodes is saved in cache. When a node sends a joining request to a source node, a node will be randomly selected as an agent of the new node from the cache of the source node. The new node obtains its identifier list from the agent, and contacts those nodes of the list to establish its partners in the overlay. To adapt the dynamical network, a seasonal exchange between nodes will take place. Nodes announce its exist and update its identifier list by the exchange.

In DONet, either the partnerships or the data transmission directions are unfixed. Their data scheduler is carried on according to the data situation in cache. Therefore, data contents in each other cache between nodes need to be clarified. A Buffer Map (BM) is defined to denote the availability of the segments in the buffer of a node. For example, a BM is recorded by using 120 bits, and each bit corresponds with a segment, with bit 1 indicating that a segment is available and 0 otherwise. However, because DONet is a semi-synchronized system, the data in the BM of each node is also different. Hence, the sequence number of the first segment is record by another two bytes. Each node continuously exchange its BM with the partners, and then schedules which segment is to be fetched from which partner accordingly.

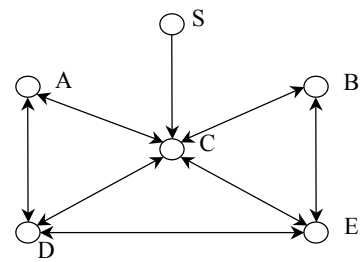


Figure 7 An example of the partnership in DONet

4 Conclusions

In this paper, some representative P2P media streaming systems are compared and analyzed. The core of tree-based systems is the construction and maintenance of multicast trees. In single-tree systems, the structure of multicast trees is simple. Each node has a unique parent node. Nodes are greatly affected by its parent node. And all the leaf nodes don't contribute their uploading bandwidth. This easily results in the burden unbalance of multicast trees. In multi-tree systems, the problem is solved in a certain extent. However, the multi-tree management is complicated. Mesh-based systems usually need high buffer. But the management cost of this system is lower, and the robust of this system is stronger. Mesh-based systems can well adapt to P2P dynamical network environment. Generally

speaking, mesh-based systems have superior performance than tree-based systems.

References

[1] D Xu. M. Hefeeda, S. Harnbrusch. and B. Bhargava.: On peer-to-peer media streaming. In: Proc. ICDCS'O2. Jul. 2002

[2] Hefeeda, M. M. and Bhargava, B. K.: On-Demand Media Streaming Over the Internet. In: Proceedings of the 9th IEEE Workshop on Future Trends of Distributed Computing Systems.San Juan, Puerto Rico ,2003 On-Demand Media Streaming Over the Internet

[3] Suman Banerjee, Bobby Bhattacharjee: A Comparative Study of Application Layer Multicast Protocols. Submitted for Publication, October 2002

[4] S. Ratnasamy, M. Handley, R. Karp, and S. Shenker, Application-level multicast using content-addressable networks. In: Networked Group Communication, 2001, pp. 14-29

[5] Magharei N, Rejaie R, Guo Y. :Mesh or multiple-tree: a comparative study of live p2p streaming approaches. In: Proc. IEEE INFOCOM, 2007

[6] N. Sebe, Y. Liu, and Y. Zhuang (Eds.): Challenges on Peer-to-Peer Live Media Streaming.In: MCAM 2007, LNCS 4577, pp. 37 - 41, 2007

[7] Deshpande H, Bawa M., Garcia-Molina H. :Streaming live media over peers. Stanford database group technical report (2001-20) ,August 2001

[8] Castro M, Druschel P, Nandi A, Rowstron A, Singh A.: Split-Stream: High-bandwidth content distribution in a cooperative environment, In: Proc. the International Workshop on Peer-to-Peer Systems, Berkeley, CA , February 2003

[9] X Zhang, J Liu, B Li, et al. CoolStreaming/DONet: A Data-driven Overlay Network for Live Media Streaming[C]. In: Proc. IEEE INFOCOM, 2005

[10] A. J. Ganesh, A.-M. Kermarrec, and L. Massoulie: Peer-to-peer membership management for gossip-based protocols. In: IEEE Transactions on Computers, 52(2), Feb. 2003

Research of Asynchronous Transfer of Control in JVM

Xinyu Wang

College of Information, Jiangnan University, Wuxi, 214122, China
Email: james.w1982@163.com

Abstract

Firstly, this paper discusses the difference between ATC mechanism and the normal Java interrupt mechanism, and presents a method(Public Space) of adding ATC mechanism in the standard java virtual machine-SableVM. Public Space which could be accessed by every threads run in JVM is a region in memory for storing AIE. In the end, I prove that in different circumstances, changed sablevm which is added Public Space and ATC which runs in Realtime System have same running results.

Keywords: ATC, RTSJ, Realtime, JVM

1 Introduction

At present, embedded systems have very rapid development, they has large but growing number ,and their applications are more widely. the systems themselves are becoming more diverse and complicated. However, comparing with desktop systems, development tools of embedded systems are the traditional C / C + + still ,even compiling language. But these development tools are complex to use, inefficient, lots of error to cause, poor portability, cross-compiler to need and other shortcomings. These shortcomings become obstacles of embedded systems' rapid development. Hence, there must need a new development language and the development mean to promote the development and application of embedded systems.

Java programming language is an object-oriented universal programming language, comparing with the traditional C / C ++language, it is safer, more practical and more complicated, and there are more advantages and progress. Because embedded systems are generally real-time system, but the Java language specification [1] and the Java Virtual Machine specification [2] [3] focusing on the issues of real-time is not enough, they discuss very broadly, and different virtual machine has inconsistent performance of Thread scheduling . At the same time, traditional Java technology used latest analysis (analysis when using this method or this section), dynamic class loading, garbage collection and other mechanisms have also affected the system's real-time ability.

For the lack of Java in real-time field, Sun Company joint IBM, Microware, QNX, Ada, Nortel Networks and other companies formulated a real-time specification of Java (RTSJ) [4].This file provide a reference to realization of real-time Java platform. Real-time Java standard improved at memory management, thread scheduling, asynchronous event handling, asynchronous transfer of control (ATC), and other real-time impacts, make Java be applied in real-time system.

This paper is focusing on asynchronous transfer of control mechanism of RTSJ, and propose one way to achieve asynchronous transfer of control in the open-source Java Virtual Machine---SableVM[5].

2 Interrupt Mechanism of Java

2.1 Thread Interrupt in Context

Thread interrupt is an old feature of the Java language. Calling a thread's interrupt method puts the thread into the interrupted state Blocking services that do not have effects other than blocking(like sleep、join、wait、waitForAll、getNextEvent,and others),unblock and

throws an InterruptedException if interrupt is called on the blocking thread or if the thread is in interrupt state when the call is made. I/O calls are also supposed to unblock and throw an exception if the thread is interrupted. For I/O operations, the exception is InterruprfIOException, which allows the I/O operation to include information about how far the I/O operation got before it was interrupted.

The standard interrupt mechanism is insufficient because of the following behaviors:

1. A conforming implementation of the JVM can completely ignore interrupts during I/O operations or other blocking methods. A real-time application needs a stronger assurance than this.

2. The conforming standard JVM can even ignore interrupts during non-I/O blocking operation.

3. The interrupt actually causes an exception (optimally)only when the thread is blocked. At other times the thread must poll for an interrupt.

4. Interrupts are not tagged. A standard Java platform makes no provisions for a thread that is interrupted while it is already in interrupted state.

2.2 Asynchronous Transfer of Control in RTSJ

The RTSJ mechanism for asynchronous transfer of control addresses these difficulties, except for termination of malicious threads.

1. Consistent conversion to an exception. AsynchronouslyInterruptedException cannot be ignored without serious effort. An application is unlikely to ignore the exception by mistake.

2. Hard to discard inadvertently. An AIE can be caught inadvertently, but the platform will raise the exception again if it is not handles according to a particular formula. That is ,it can be caught inadvertently, but it will not stay caught.

3. Safe with lock. Interrupting synchronized code is unwise, and the RTSJ does not do it. Synchronized blocks, even when they appear in methods that throw AsynchronousluInterruptedException, cannot be interrupted. If an algorithm really wants to find out about an AIE in the middle of a synchronized block, it can polling the isInterrupted method for it.

4. Provides for reinterrupt. A RealtimeThread can be reinterrupted while an AIE is in flight. Rules in the RTSJ determine whether the new exception will replace the previous one.

In addition, asynchronous transfer control mechanism of RTSJ has its unique advantages in consistent implementation, scheduling polling and security legacy code.

3 DESIGN of Asynchronous Transfer of Control

This paper improve details of this process in SableVM. SableVM is Java Virtual Machine which is robust, simple, easy to maintain and expand, very lightweight and in accordance with norms of JVM.

3.1 Interrupt Mechanism of SableVM

SableVM is a standard Java Virtual Machine, and its interrupt mechanism has described in 2.1. However, because of its building on standard Java specification, SableVM is not involved in ATC of RTSJ. Of course, SableVM can not identify AIE and there are no ATC system in it. My duty is to make SableVM can identify AIE in Java procedures (2.2 mentioned), and add the control of the asynchronous transfer mechanism in SableVM.

3.2 ATC Mechanism

Asynchronous transfer of control(ATC) is a mechanism(broadly similar to Thread.stop) that lets one thread throw an exception into another thread. ATC is important for some classes of real-time applications. The RTSJ defines a class, called AsynchronouslyInterruptedException (AIE), for asynchronous exceptions, with special rules that make the exception safe for use with the Java programming language:

1. The target thread will not service an asynchronously interrupted exception until it reaches a method that explicitly throws AsynchronouslyInterr-

uptedException.

2. The AIE will be asserted immediately if the target thread is in a method that throws it; otherwise, the AIE will remain pending until it enters such a method.

3. When control enters a method that throws AIE and an AIE is pending, it is thrown immediately.

4. The pending AIE remains pending until it is serviced.

5. An AIE can be thrown at a thread that already has an AIE pending. The new AIE will replace the pending one if it is aimed at less deeply neated method.

3.3 Improvement in Structure and Principle

From ATC mechanism we know that ATC is a communication measure among different threads , AIE must be accessed by all threads moved in JVM.

In SableVM, a structure named env represent a thread; multi-thread in SableVM mean there are lots of env operated at the same time, then virtual machine arrange these env into a chain-table, and schedule them.

3.3.1 Preparation

To identify AsynchronouslyInterruptedException (AIE) string in Java program, we must add some relational APIs in SableVM. But after this, we have to alloc memory for this exception, and create its object and implement in memory. Then, we need to amend these files :"error.list" "method_invoke . list "and" bootstract.m4.c ".Compiler will automatically generate "bootstrap.c"and other files. There are lots of methods using on identify AsynchronouslyInterruptedException string,initialize,alloc memory and impletment.We could use these methods.Then SableVM has been able to identify AsynchronouslyInterruptedException string in Java program, and also can initialize, alloc memory space and implement for identified object in SableVM.

3.3.2 Design of Public Space

Public space's structure seems like Figure 1. Data structures in it is explained as follows:

AIE_Flag is a sign , when turning to TRUE, it says the AIE in current process is activated.

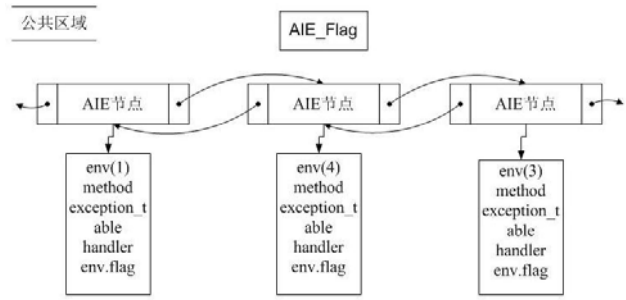


Figure 1 Public Space Structure

AIE node contains a data structure---struct AIE. In this structure, I store method address where could throw out a AIE(but not necessarily throw here) in memory, and thread address in memory which this method is belong to(env), the exception_table address of this method (exception_table). handler is a pointer, which point to the "catch" address corresponding to nearest "try" and in this try-catch section might throw a AIE. Env has been changed. I add a flag in it named "flag". This flag is used like this: When it turn to TRUE, it signs that this thread contains a AIE string, but JVM did not know whether this AIE is triggered. So this AIE can only be linked on AIE chain in Public Space. This may be extended, as long as AIE-nodes are in the public space, their "env.flag" is TRUE.

AIE link is a stack, which used the LIFO algorithm. As long as the value of a AIE_Flag turning to TRUE, SableVM throw out the AIE which is at top of stack. When meeting some special circumstances, we will adopt replacement rules will mentioned in 3.3.4.

3.3.3 ATC Rules

To achieve ATC mechanism in SableVM, these must be considered: different threads access the same section of procedures (such "buy ticket problem"), different threads access different section of AIE,and other circumstances. Following will prove that operation on the structure(public space) mentioned in 3.3.2 is fully consistent with the results of theoretical ATC mechanism at different circumstances.

[1]: Different threads access the same section of procedures

When thread run to a method which could throw

out AIE, it checks whether context has "try (…) catch (AIE e) (…)" structure, and this method will be pend on AIE chain in Public Space. Three threads as an example, the public space's structure at this is shown in Figure 2.

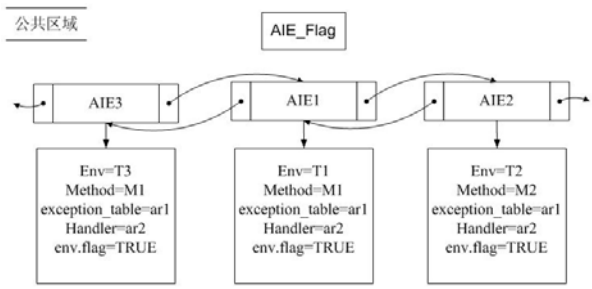


Figure 2 Public Space of [1]

In this figure, AIE stack's top is on the left. The reason why thread T2 and T3's AIE nodes is front of thread T1's AIE node is because different thread run in different speed, it is possible that thread T2 checking "try (…) catch (AIE e) (…)" section fastlier than T1, then T2's AIE T2 node will be pended into AIE stack in public space firstly. Therefore, these three nodes' order is uncertain, it depends on system's scheduling thread .

When AIE is triggered, SableVM will set AIE_Flag which is in public space to TRUE. Because of FIFO principle, thread T2's AIE will be thrown cause it is top of the stack, pc will be changed into handler's value of T2's AIE node. Similarly, when AIE_Flag is not to be TRUE, AIE of carious threads will not be thrown . This result correspond with the mechanism of the ATC (1) (2) (3) (4)which are mentioned in 3.2 .

[2]: Different threads access different section of AIE

[1] and [2] is similar at handling-method. The difference between them is not the same methods which throw different AIE ,different "exception_table" address and "handle" address must not be the same. It should be noted when you set their value. But they act a same principle. Public Space structure theory is also consistent with the behavior of ATC's (1) (2) (3) (4) mentioned in 3.2.

3.3.4 Replace Rules

When an AIE is fired at a thread, when it is thrown, or when it is propagated, it transfers control to the nearest suitable catch or finally block. The suitable target will be in an ATC-deferred region.

[3]:If the thread is blocked in wait、sleep、join、or in an I/O operation that throws InterruptedIOException, the AIE will unlock the thread.

[4]:If control leaves an ATC-deferred region because the code throws an exception other than an AIE, the pending AIE will replace the thrown exception.(figure 3).

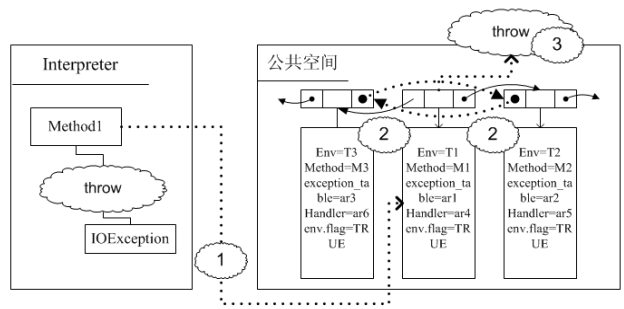


Figure 3 Interpreter and Public Space of [3]

When method2 throws a IOException in interpreter, this exception will be firstly sent to public space (Figure 3 ①), virtual machine will check the AIE stack to find out whether there are same method to throw AIE. If there are no AIE node belonging to method2 in the AIE stack, then throw this IOException. If this node is existence, AIE_Flag will be set to TRUE, this node will be out of the AIE stack (Figure 3 ②), and replaces IOException (Figure 3 ③), after these action, AIE_Flag will be reset.

If the thrown exception is an AIE, replacement rules apply. If the doInterruptible for aie1 is currently on the call stack, fire will return true, and the following action then occur:

1. If the AIE in flight is aie2,which is deeper on the stack than aie1,then the new exception will be silently ignored.[4]

2. If the AIE in flight targets aie3,which is less deeply nested on the call stack, then aie1 will replace aie3 and exception processing in the target thread will continue with the new exception.[5]

[4][5]is easy to understand. As RTSJ pays more attention to the real-time field, thread must accomplish

within tolerable time(user feels). When the method which could throw aie1 has been pended in invoking-queue, the system will not allow this method to wait for throwing aie1 until aie1 will poll to the time when aie1 is thrown. However, when not to affect overall operation of system, aie1 will be thrown to ensure the method throws aie1 complete at a tolerable time.

4 Conclusion

As described above, we could see that making the public space structure mentioned in 3.3.2 into SableVM, it make SableVM satisfy some requirements of ATC, and make ATC mechanism exist in SableVM.

References

[1] J. Gosling, B. Joy, G. Steele, and G. Bracha, "The Java Language Specification Second Edition," 2000

[2] T. Lindholm and F. Yellin, "The Java Virtual Machine Specification, 2nd edition," Addison Wesley, 1999

[3] B. Venners, Inside the Java Virtual Machine, Second Edition, 2003

[4] G. Bollela, J. Gosling, B. Brosgol, P. Dibble, S. Furr, D. Hardin, and M. Trunbull, "The Real-Time Specification for Java," Addison Wesley, vol. 1st edition, 2000

[5] SableVM. http://www.sablevm.org

[6] Garcia A F, Rubira C M F. A Comparative Study of Exception Handling Mechanisms for Building Dependable Object-oriented Software[J]. ACM Transactions on Programming Languages and Systems, 1998,20(2):294-301

[7] Cristian F. Exception Handling and Tolerance of Software Faults[J]. InSoftware Fault Tolerance (M. Lyu, ed.), John Wiley & Sons, 1995:81

[8] Dony C, Purchase J, Winder R. Exception Handling in Object-oriented System[C]. Report on ECOOP91 Workshop W4, 1991:17-30

[9] Sun Microsystems．Inc．JavaTM 2 SDK，Standard Edition Documentation[EB/OL]. http://java.sun.com/j2se/ 1.4.2/ docs/index.htm1

Exploring L-tryptophan Synthesis Metabolism Network Through Extreme Pathway Analysis

Dong Wang[1] Li Liu[1] Wenbo Xu[1] Zhijun Zhao[2] Jing Wu[2]

School of Information Technology, Jiang Nan University, Wuxi, 214122, China
Email: seaingwang@yahoo.com.cn

State Key Laboratory of Food Science and Technology, Jiang Nan University, Wuxi, 214122, China
Email: zhaozjun2004@163.com

Abstract

Metabolic pathway analysis is becoming more and more important for identification and quantification of all metabolites in (reconstructed) biochemical reaction networks. One of the most promising concepts for pathway analysis is extreme pathway analysis. Extreme pathways are a mathematically defined set of generating vectors that describe the conical steady-state solution space for flux distributions through an entire metabolic network. The objective of this study is to obtain a more detailed insight into L-tryptophan metabolism through network-based metabolic pathway analysis. Herein, the extreme pathways analysis is used for our study. Besides of "historical" biochemical pathways, the results also suggest that there are some novel pathways. And further experiments are needed to make a decision.

Keywords: Metabolic pathways, Extreme pathways, Elementary modes, L-tryptophan metabolism

1 Introduction

Metabolism involves the production of mass, energy, and redox requirements for all cellular functions, and thus provides the driving force for cellular activity. As one of the most thoroughly studied aspects of cellular function, it affords the best opportunity for the development of methodologies to characterize and analyze systems-level cellular properties of genome-scale models [1]. Metabolic engineering is a term used to describe the practice of improving cellular functions by directly manipulating enzymatic, regulatory, and transport processes in the cell [2]. So, understanding the structural design and capabilities of a cellular metabolic network clearly places the biochemical engineer in an advantageous position to manipulate the functional attributes of the system [3].

A metabolic network consists of the group of reactions and transport processes associated with the production and depletion of cellular metabolites. Biochemical pathways are thought of as functional units of metabolic networks. As such, the definitions and characterizations of metabolic pathways allow for a detailed analysis of robustness, physiological capabilities, and other systemic features of these complex reaction networks. These emergent properties necessitate the development of clear and mathematically precise definitions for a metabolic pathway. Mathematical and systemic definitions of metabolic pathways have been proposed [4]. A central concept in metabolic pathway analysis is that of extreme pathways [5, 6].

Extreme pathways are flux maps through a biochemical network that characterize the functioning of the network; in other words, extreme pathways account for the relative magnitude of the number of molecules undergoing each reaction in the cell. Extreme pathway analysis has the following characteristics: (1) it generates a unique and minimal set of systemic pathways; (2) it describes all possible steady-state flux distributions that the network can achieve by non-negative linear combinations of the extreme pathways; and (3) it enables the determination of time-invariant, topological properties of the network.

The calculation of extreme pathways is computationally challenging and for large networks, generates a tremendous amount of numerical data [3, 7].

The method of extreme pathways has been used in a number of metabolic networks, including the human red blood cell [8], the microorganisms E.coli [9], Haemophilus influenza [3, 10] and Helicobacter pylori [1, 11].

L-tryptophan, as one of eight essential amino acids in human nutrition, plays an important role in physical mechanisms functions and is widely applies in medicine, food ,feed but also many other aspects . How to enhance the produce rate of tryptophan through fermentation is an important topic. In recent years, people try to change internal metabolism fluxes distribution of the cell through metabolism project method, as to produce more products. Herein, we employ the extreme pathways to analyze the Tryptophan biosynthesis metabolism network, with the purpose of exploring metabolism of L-tryptophan (L-TRP). Even though L-TRP metabolism has been well studied experimentally, powerful theoretical methods such as pathway analysis are needed to discover novel mechanisms, because of metabolism is often much more plastic than the set of pathways defined in traditional biochemistry textbooks, as well as validated experimentally.

2 Meterials and Methods

Metabolites and metabolic reactions

We choose L-tryptophan synthesis metabolism network as an in silico model to perform this study, for L-tryptophan synthesis metabolism network is complete sequenced and extensive annotated. So, it can afford sufficient information for this study.

The starting point is provided by information collected from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database23 (http://www.genome.ad.jp/kegg/). We checked every enzymatic reaction carefully. All metabolites and metabolic reactions are included in L-tryptophan synthesis metabolism network are shown in Table 2. Those metabolites that can be exchanged with the system boundary are identified for calculating the extreme pathways (for detail see Table 1). Metabolites that can be thought of as being transported across the cell membrane and exchanged with the environment are called primary exchanges, while currency exchanges are exchanges of cofactors such as ADT, ATP, NADH etc [6]. Note that we don't include the enzyme which catalyze enzymatic reaction does not significantly change the extreme pathway analysis of the system, such as aldehyde dehydrogenase (EC 1.2.1.3), 4-aminobutyrate aminotransferase (EC 2.6.1.19) etc. Then, the full L-tryptophan metabolism stoichiometric matrix is derived from these reactions and the corresponding metabolic model could be constructed (Figure 1).]

GLC: D-glucose; G6P: D-glucose 6-phosphate; F6P: D-fructose 6-phosphate; FDP: D-fructose 1,6-bisphosphate; DHAP: glycerone phosphate; GA3P: D-glyceraldehyde 3-phosphate; 13DPG: 3-phospho-D-glyceroyl phosphate; 3PG: 3-phospho-D-glycerate; 2PG: 2-phospho-D-glycerate; PEP: phosphoenolpyruvate; PYR: Pyruvate; 6PGL: 6-phospho-D-glucono-1,5-lactone; 6PGC: 6-phospho-D-gluconate; RL5P: D-ribulose 5-phosphate; X5P: D-xylulose 5-phosphate; R5P: D-ribose 5-phosphate; S7P: sedoheptulose 7-phosphate; E4P: D-erythrose 4-phosphate; F6P: D-fructose 6-phosphate; DAHP: 3-deoxy-D-arabino-hept-2-ulosonate 7-phosphate; DHQ: 3-dehydroquinate; Q: L-quinate; DHS: 3-dehydroshikimate; SHIK: shikimate; S3P: 3-phosphoshikimate; EPSP: 5-O-(1-carboxyvinyl)-3-phosphoshikimate; CHA: chorismate; L-Gln: L-glutamine; ANTA: anthranilate; L-Glu: L-glutamate; PRPP: 5-phospho-alpha-D-ribose 1-diphosphate; PRAA: N-(5-phospho-D-ribosyl)-anthranilate; PPi: diphosphate; CDRP：1-(2-carboxyphenylamino)-1-deoxy-D-ribulose; 13GP: 1-C-(indol-3-yl) glycerol 3-phosphate; L-Ser: L-serine; L-Trp: L-tryptophan; ADP: Adenosine diphosphate; ATP: Adenosine triphosphate; NAD: Nicotinamide adenine dinucleotide; NADH: Nicotinamide adenine dinucleotide (R); NADP: Nicotinamide adenine dinucleotide phosphate; NADPH: Nicotinamide adenine dinucleotide phosphate (R); PI: Phosphate; H: Hydrogen Ion; H2O: Water.

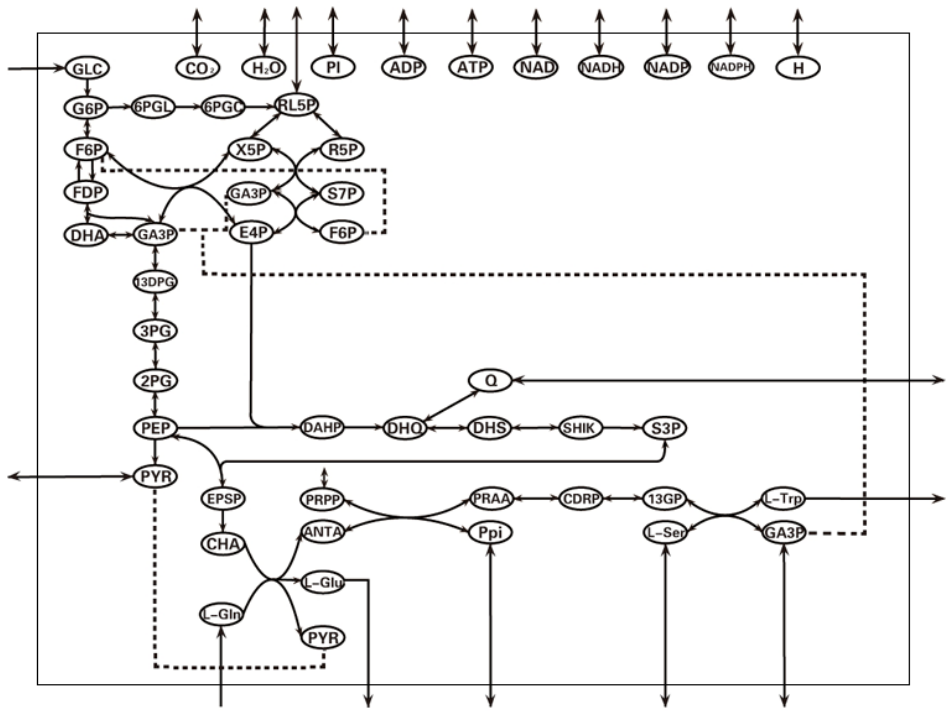


Figure 1 Metabolic network map for L-tryptophan synthesis

Pathway analysis and classification

Extreme pathways define the edges of high dimensional cone that circumscribes all possible flux distributions in a metabolic network [6]. They are calculated from a stoichiometric matrix, S, which has dimensions of (m by n) where m is the number of metabolites in the reaction network and n is the number of reactions. Therefore, the columns contain the stoichiometric coefficients of the metabolic reactions and rows correspond to the metabolites. Reversible internal reactions are decoupled into two separate reactions for the forward and reverse directions. The application of mass balance and reaction directionality Eq. (1) constraints defines a set of solutions to the metabolic networks. The flux distribution through a reaction network can be represented as:

$$S \cdot v = 0 \quad (v_i \geq 0, \forall i) \tag{1}$$

Where v is the flux vector containing the individual reaction fluxes, v_i. The set of extreme pathways is the convex basis of S, defined by Eq. (1). An algorithm for calculating extreme pathways has been previously published [6]. Herein, we employ the program EXPA (for the current version see http://gcrg.ucsd.edu/downloads/expa.html) to calculate the extreme pathways.

Extreme pathways can be divided into three basic categories based upon their exchange fluxes [6]. Each of these three categories can be understood relative to energy usage.

Type I extreme pathways are those that have exchange fluxes across the system boundaries that correspond to no currency metabolites (Figure 2, left). These extreme pathways can be energetically interpreted analogously to charging a battery. These extreme pathways drive the cycling of metabolic currencies, such as ATP, which then drive other cellular processes.

Type II extreme pathways are those that have "exchange fluxes" corresponding only to currency metabolites (e.g., ATP, NADH), with the rest of the pathway being an internal cycle (Figure 2, center). These pathways represent futile cycles and are analogous to draining a battery. They are unidirectional in the absence of a driving force on the cofactor pool.

Type III extreme pathways are those that have no exchange fluxes (Figure 2, right). Thus, these represent internal cycles. The fluxes through interior cycles must

necessarily be zero to satisfy the loop constraints.

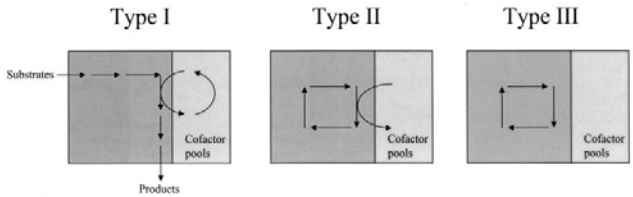


Figure 2 Types of Extreme Pathways.

3 Results

The computation of the extreme pathways for the L-tryptophan synthesis metabolism network result in 19 type I, 1 type II, and 20 type III extreme pathways. Herein, we will focus on type I and type II extreme pathways which are the most interested. Table 1 contains the net reactions (exchanges only) for both type I and type II extreme pathways.

D-glucose to D-ribulose 5-phosphate/ D-glyceraldehyde 3-phosphate (GRGA)

These extreme pathway GRGA1 uses glucose to glyceraldehyde 3-phosphate through the route of glycolysis, and GRGA2 uses glucose to ribulose 5-phosphate through pentose phosphate pathway (PPP). GRGA3 is used the combination between glucose and glyceraldehyde 3-phosphate to produce ribulose 5-phosphate, In agreement with common biochemical knowledge.

D-ribulose 5-phosphate/D-glyceraldehyde 3-phosphate conversion (RGA)

Pathways RGA1 to RGA4 represent the different conversation between ribulose 5-phosphate and glyceraldehyde 3-phosphate that can occur in the cell. These pathways are the basic pathways to analysis the whole metabolism network to produce the maximum of tryptophan with the lowest reactants.

D-glyceraldehyde 3-phosphate to Pyruvate (GAP)

This pathway use glyceraldehyde 3-phosphate to pyruvate through glycolysis, which is clear from common biochemical knowledge.

Table 1 The net reactions for all type I and type II (DIS1-3) extreme pathways including both primary and currency exchange fluxes. Net reactions include system inputs (negative integers) and outputs (positive integers), and do not include internal reactions

P_#	GLC	L-Gln	RL5P	GA3P	Q	PRPP	PPi	L-Ser	L-Glu	L-Trp	PYR	ADP	ATP	NAD	NADH	NADP	NADPH	PI	CO2	H	H2O
GRGA1	1	0	0	2	0	0	0	0	0	0	0	2	-2	0	0	0	0	0	0	0	0
GRGA2	1	0	1	0	0	0	0	0	0	0	0	1	-1	0	0	-2	2	0	1	2	-1
GRGA3	2	0	3	-1	0	0	0	0	0	0	0	2	-2	0	0	0	0	0	0	0	0
RGA1	0	0	1	-2	0	0	0	0	0	0	0	0	0	0	0	-2	2	1	1	2	-2
RGA2	0	0	3	-5	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	-2
RGA3	0	0	-1	1	0	0	0	0	0	0	0	0	0	0	0	-4	4	0	2	4	-2
RGA4	0	0	-3	5	0	0	0	0	0	0	0	2	-2	0	0	0	0	0	0	0	0
GAP	0	0	0	-1	0	0	0	0	0	0	1	-2	2	-1	1	0	0	1	0	0	1
TRP1	0	-1	-1	-1	0	-1	1	-1	1	1	1	-1	1	-2	2	-1	1	2	2	4	3
TRP2	1	-1	1	-2	0	-1	1	-1	1	1	1	0	0	-2	2	1	-1	2	1	2	4
TRP3	0	-2	-1	-3	0	-2	2	-2	2	2	2	-2	2	-4	4	2	-2	4	2	4	8
TRP4	0	-1	1	-4	0	-1	1	-1	1	1	1	-1	1	-2	2	1	-1	3	1	2	3
TRP5	0	-1	-2	1	0	-1	1	-1	1	1	1	0	0	-2	2	1	-1	2	1	2	4
QT1	0	0	-1	-1	1	0	0	0	0	0	0	-1	1	-1	1	-1	1	1	1	2	-1
QT2	1	0	1	-2	1	0	0	0	0	0	0	0	0	-1	1	1	-1	1	0	0	0
QT3	0	0	-1	-3	2	0	0	0	0	0	0	-2	2	-2	2	2	-2	2	0	0	0
QT4	0	0	1	-4	1	0	0	0	0	0	0	-1	1	-1	1	1	-1	2	0	0	-1
QT5	0	0	-2	1	1	0	0	0	0	0	0	0	0	-1	1	1	-1	1	0	0	0
QTRP	0	-1	0	0	1	-1	1	-1	1	1	1	0	0	-1	1	0	0	1	1	2	4
DIS	0	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	1	0	0	-1

Table 2 Reactions included in the model of L-tryptophan synthesis metabolism network

Abbreviation	Chemical Reaction	EC Number	Reversibility
V0	ATP + GLC —> G6P + ADP	2.7.1.2	IRREV
V1	G6P <—> F6P	5.3.1.9	REV
V3	ATP + F6P —> ADP + FDP	2.7.1.11	IRREV
V4	FDP+H2O —> F6P + PI	3.1.3.11	IRREV
V5	FDP <—> DHAP + GA3P	4.1.2.13	REV
V7	DHAP <—> GA3P	5.3.1.1	REV
V9	GA3P + PI + NAD <—> 13DPG + NADH + H	1.2.1.12	REV
V11	ADP + 13DPG<—>ATP + 3PG	2.7.2.3	REV
V13	3PG <—> 2PG	5.4.2.1	REV
V15	2PG <—>PEP + H2O	4.2.1.11	REV
V17	ADP + PEP + H —>ATP + PYR	2.7.1.40	IRREV
V18	G6P + NADP+ —>6PGL +NADPH + H	1.1.1.49	IRREV
V19	6PGL + H2O —> 6PGC	3.1.1.31	IRREV
V20	6PGC + NADP+ —> RL5P + CO2 + NADPH	1.1.1.44	IRREV
V21	RL5P <—> X5P	5.1.3.1	REV
V23	RL5P <—> R5P	5.3.1.6	REV
V25	R5P + X5P <—> S7P + GA3P	2.2.1.1	REV
V27	E4P + X5P <—> GA3P + F6P	2.2.1.1	REV
V29	S7P + GA3P <—> E4P + F6P	2.2.1.2	REV
V31	PEP + E4P + H2O —>DAHP + PI	2.5.1.54	IRREV
V32	DAHP —> DHQ + PI	4.2.3.4	IRREV
V33	DHQ <—> DHS + H2O	4.2.1.10	REV
V35	Q + NADP <—> DHQ + NADPH + H	1.1.1.282	REV
V37	DHS + NADPH + H <—> SHIK + NADP	1.1.1.25	REV
V39	ATP + SHIK —> ADP + S3P	2.7.1.71	IRREV
V40	PEP + S3P <—> PI + EPSP	2.5.1.19	REV
V42	EPSP —> CHA + PI	4.2.3.5	IRREV
V43	CHA + L-Gln—> ANTA + PYR + L-Glu	4.1.3.27	IRREV
V44	ANTA + PRPP<—> PRAA + PPi	2.4.2.18	REV
V46	PRAA <—> CDRP	5.3.1.24	REV
V48	CDRP <—> 13GP + CO2 + H2O	4.1.1.48	REV
V50	L-Ser + 13GP <—> L-Trp + GA3P + H2O	4.2.1.20	REV

L-tryptophan production(TRP)

These pathways shows the basic way to the L-tryptophan. TRP1 produces the standard one ATP, two NADH ,and one L-tryptophan from one glyceraldehyde 3-phosphate molecule and one ribulose 5-phosphate molecule. TRP2 uses D-glucose to D-fructose 6-phosphate through the route of glycolysis, and then with one molecule glyceraldehyde 3-phosphate to produce one L-tryptophan through PPP and Tryptophan Biosynthesis (TB). TRP3 takes TRP1 a step further and actually cycles through PPP repeatedly. The net result is an increase in the net production of L-tryptophan. TRP 4 and PRP5 are similar with the production of one L-tryptophan molecule through glycolysis, PPP and TB. The only difference between them is the Reactant. The reactant of TRP4 is just glyceraldehyde 3-phosphate; the other one is just ribulose 5-phosphate molecule.

L-quinate production (QT)

The pathways (QT1-5) are identical to those from group4 (TRP1-5), the only difference is the production of L-quinate instead of L-tryptophan.

L-quinate to L-tryptophan (QTRP)

This pathway produces one L-tryptophan molecule through TB, as the following work of QT.

Dissipation of ATP (type II pathways, DIS)

This type II pathway represent the futile cycle

between D-fructose 6-phosphate and D-fructose 1,6-bisphosphate as to dissipate excess ATP.

4 Discussion

Extreme pathway analysis has been applied to L-tryptophan synthesis metabolism network. The resulting extreme pathways were analyzed and classified based on their structure and functional capabilities. Even though there are some "historical" pathways being computed as extreme pathways. However, a majority of the extreme pathways are nontraditional such as pathways: TRP2, TRP3, TRP4, TRP5 and QTRP. They use the combination of different reactants to the L-tryptophan production. Are these pathways existent as salvage pathways? Such nontraditional pathways are a good example of how extreme pathway analysis can elucidate systemic properties resulting from complex network which highly connected.

The results from the extreme pathway analysis show that network structure and capacity constrains provide a strong basis for analysis and interpretation of physiology of the L-tryptophan synthesis metabolism network. Genome-scale metabolic networks can now be reconstructed from genomic and other data sources [12].The use of extreme pathways for analysis of such reconstructed network and their relation to whole-cell functions have become critical in the advancement of systems biology.

As the extreme pathways of L-tryptophan synthesis metabolism network have been established, the reactions that always occur together in each of the extreme pathways in which they are active can be calculated. In combination with experimental determination information and bioinformatics analysis, it might give a good examining their regulation at the genetic level. The future of our work will be expanding the scope of the L-tryptophan synthesis metabolism network to have the more detailed information of the L-tryptophan synthesis metabolism network.

References

[1] Price ND. Determination of redundancy and systems properties of Helicobacter pylori's metabolic network using genome-scale extreme pathway analysis. Genome Res 12:760-769, 2002

[2] BAILEY, J. E. (1991). Toward a science of metabolic engineering. Science 252, 1668-1675

[3] Schilling CH, Palsson BO. Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis. J Theor Biol 203:249-283, 2000

[4] Jason A. Papin, Nathan D. Price, Bernhard Ø. Palsson. Extreme Pathway Lengths and Reaction Participation in Genome-Scale Metabolic Networks. Genome Res. 2002 12: 1889-1900

[5] Schilling CH et al. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era, Biotechnol Prog 15:296-303, 1999

[6] Schilling CH, Letscher D, Palsson BO. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective, J.Theor.Biol 203:229-248, 2000

[7] Samatova, N.F., Geist, A., Ostrouchov, G., and Melechko, A.V. 2002. Parallel out-of-core algorithm for genome-scale enumeration of metabolic systematic pathways. Proc. First IEEE Workshop on High Performance Computat. Biol. (HiCOMB2002), Ft. Lauderdale, FL

[8] Wiback SJ, Palsson BO, Extreme pathway analysis of human red blood cell metabolism, Biophys J 83:808-818, 2002

[9] Schilling CH et al, Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems, Biotechnol Bioeng 71:286-306, 2000

[10] Papin JA et al, The genome-scale metabolic extreme pathway structure in Haemophilus influenzae shows significant network redundancy, J Theor Biol 215:67-82, 2002

[11] Schilling CH et al, Genome-scale metabolic model of Helicobacter pylori 26695. J Bacteriol 184:4582-4593, 2002

[12] Covert, M. W., C. H. Schilling, I. Famili, J. S. Edwards, I. I. Goryanin, E. Selkov, and B. O. Palsson. 2001a. Metabolic modeling of microbial strains in silico. Trends Biochem. Sci. 26:179-186

FPGA Implementation of Digital Filter

Fan Liu

School of Information Technology,Jiangnan University, Wuxi, Jiangsu, 214122, China
Email: liufanchinchen@126.com

Abstract

This paper introduces the basic knowledge of digital filter, and summing up the general design of digital filters, as well as some of the more efficient algorithms. Analysis of the advantages and disadvantages of various methods of design, the paper contends many kinds of structural plans. Then we focus on the CSD code which is used to achieve multiplier application. In this way, the hardware resources can be saved. And also, it's quit a good way to increase the multiplier's speed.

Keywords: digital filter, FPGA, pipeline, CSD code

1 Introduction

A digital filter is essentially a system or network, it has the option to change the signal waveforms, amplitude, frequency and phase characteristics. Generally, filter is designed to improve the quality of a signal, or extract information from the signal, or separate the two or more signals combined in the past. Digital Filter has some characteristics that traditional analog filter can't reach. It hardly affected by the external environment (for example, the change in temperature). And also it achieves higher accuracy than analog filter, but it has a smaller size, lower power consumption.

2 Theoretical Foundation

At present, there are three ways to achieve FIR filter: use of generic digital monolithic integrated circuit filter, DSP devices and programmable logic devices. Monolithic common digital filter is convenient, but because of the specification of word length and order is not enough, so we can not fully meet the actual needs. Although the use of DSP devices is a simple, but because of procedures for the implementation is ordinal, so the speed of execution will be inevitably slow. FPGA has a regular internal logic array of resources and a mass of connections, so it's particularly suitable to digital signal processing tasks, in relation to serial computing-led General DSP chip, it'll be easier to achieve parallel. But for a long time FPGA has been used for logic or timing control system, very few signals processing application use FPGA, the main reason is because the FPGA do not have effective structure for multiplication. Now this issue has been resolved so that the FPGA in digital signal processing has developed by leaps and bounds.[8]

In accordance with the impulse response function's time-domain features, Digital filter can be divided into two types: infinite impulse response (IIR) filter and finite impulse response (FIR) filters. They both have their own advantages and disadvantages.

Designing a digital filter includes the following five steps:

(1) Standard of the filter requirements.

(2) Suitable filter coefficient calculations.

(3) Using an appropriate structure to express the filter.

(4) Analyze the effects of finite word-length effect

(5) Using hardware to achieve filter.

3 FPGA Implementation of Digital Filter

With the increase of capacity and speed of the programmable logic devices, to achieve single-chip integrated system (System On Chip) has become

possible. Using programmable logic device for digital filter is a good. It is the realization of the hardware parallel algorithm, so it is particularly applicable to real-time requirements of some high occasions. Here are some more methods that are often used in the FPGA implementation of Digital Filter.

3.1 FPGA Implementation of FIR filter

FIR filter's (finite impulse response filter) output depends on the input in the past and has nothing to do with the output of the past. It can be express with a difference equation:

$$y(n)=\sum_{k=0}^{N-1}h(k)x(x-k) \tag{1}$$

h(k) is the impulse response coefficients of the digital filter. y(n) is the output in the time "n". Its system function form is:

$$H[z]=\sum_{k=0}^{N-1}h(k)z^{-k} \tag{2}$$

In the design, the first step is to determine the filter indicator in accordance with the specific application. We can use MATLAB the FDA Tool (Filter Design & Analysis Tool) to determine the filter coefficients. [1] The next step is the structural design of filter. The simplest structure of FIR filter is the direct type. As shown in Figure 1[2]:

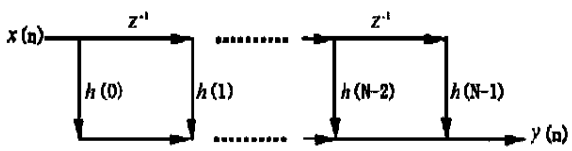


Figure 1 N-direct FIR filter

This is one of the simplest forms, but it's not the most effective form both in speed and in saving energy. Here is an optimization method:

Using linear-phase structure:

One of FIR filter's advantage is that its result have Linear Phase. Because of the symmetry factor, its impulse response unit h (n) has symmetric features:

$h(n)=h(N\text{-}1\text{-}n)$ or $h(n)=-h(N\text{-}1\text{-}n)$

Based on this characteristic, filters can be written in mathematical expression:

$$y(n)=\sum_{n=0}^{N-\frac{1}{2}}(x(n)+x(N-1-n))h(n) \tag{3}$$

This can save 50% of multipliers in each filter cycle. This will greatly increase the computing speed and reduce the wear and tear of the hardware resources. [3] Its structure is shown in Figure 2.

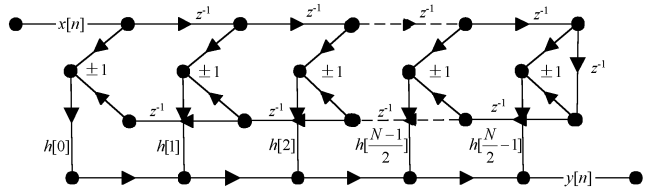


Figure 2 direct linear phase structure

3.2 FPGA Implementation of IIR filter

IIR filter's (infinite impulse response filter) output depends on the input and the output of the past. [4] It can be expressed with difference equation:

$$y(n)=\sum_{i=0}^{M}b_i x(n-i)-\sum_{i=1}^{N}a_i y(n-i) \tag{4}$$

$a_i b_i$ is the filter Coefficient. y(n) is the output in the time "*n*".

Its system function form is:

$$H(Z)=\frac{\sum_{r=0}^{M}b_r Z^{-r}}{1+\sum_{k=1}^{N}a_k Z^{-k}} \tag{5}$$

In the design, Preliminary work is almost the same as FIR filter. First we should decide the specific application of the filter indicator. We can also use MATLAB the FDA Tool (Filter Design & Analysis Tool) to determine the filter coefficient ai, bi.

When designing an IIR filter, following structure is often used:

Second-order cascade structure is often used to design IIR digital filter. The advantage of this method is that each subsystem is related to one pair of pole and zero, when adjusting coefficient, only one pair of pole and zero is affected without affecting the other pole and zero. With this structure, it's easy to achieve accurate filter zeros and Poles. Also, it facilitates the adjustment of filter frequency response performance. Another

advantage is that it uses less storage unit. The form of second-order transfer functions is:

$$H_k(z)=\frac{b_{k0}+b_{k1}z^{-1}+b_{k2}z^{-2}}{1+a_{k1}z^{-1}+a_{k2}z^{-2}} \quad (6)$$

For this kind of system, there are many programmes can be used, including direct Type I, direct Type II. Compare to direct type I, direct Type II can save more memory, so it's more efficient and has a wide range of applications. Its structure is shown in Figure 3.

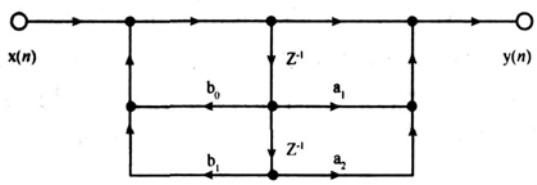


Figure 3 IIR filter second-order structure (direct Type II)

After the design of second-order structure, we can make 4-order or even more orders filter with cascade structure. Cascade structure is shown in Figure 4.

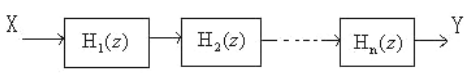


Figure 4 cascade structure

3.3 Optimization for the design

CSD code:

Multiplier design is the key to the filter design, the existing parallel multipliers' structure are almost base on Booth algorithm or improved Booth algorithm, using Wallace adder tree to reduce the number of partial product, the final use of CLA(carry look-ahead adder) to generate product. The disadvantage of this structure is that the coding circuit of partial product is too complicated. When the digit of the multiplier increases, it's hard for the multiplier circuit to expand. Using CSD code can save adder unit, and then save on the FPGA resources consumption.

The CSD coding algorithm has been put forward for a long time; binary complement code plays an important role for the CSD code to achieve high-speed multiplier. CSD code is a numerical system that can be expressed by (-1, 0, 1), there is only one form for each binary code that has the least 1 and -1. Therefore, the CSD code is the main application code in the Multiplier for reducing the partial product, and then reducing the adder and subtracter. The times of add or subtraction for a n-bit multiplication will be less than n / 2, with the increase in length, the times will reduce to n / 3.

This is the realization of CSD coding algorithm:

(1) Replace all the '1' serials which are bigger than 2 with 10…0-1 from the least significant digit. And replace 1011 with 110-1.

(2) Replace all the 10-1 with 011.

This form let the coefficient have the least nonzero bit.[5][6]

We use an example to illustrate:

$$93x=2^6x+2^4x+2^3x+2^2x+1x \quad (7)$$

After CSD coding:

$$93x=2^6x+2^5x-2^2x+1x \quad (8)$$

When a number multiplied or divided by 2, the actual operation is just one bit shift to the left or right. If a number multiplied by the sum of several powers of 2, we only need to add the sums which are based on the digit shift. And this makes the multiplication more easily. As shown in Figure 5:

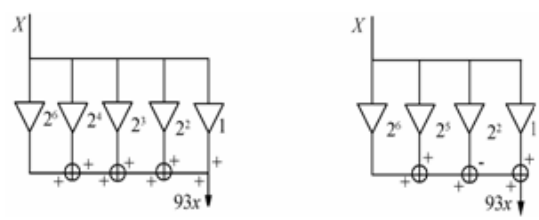


Figure 5 Binary Multiplier and CSD coding multiplier

In order to increase the throughput, we put pipeline into the system. Divide one operation into several steps, and then do the job in a higher speed. [7] As shown in Figure 6:

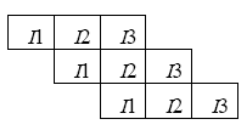


Figure 6 pipeline

After CSD coding and the pipeline design, the structure of the system is like this:

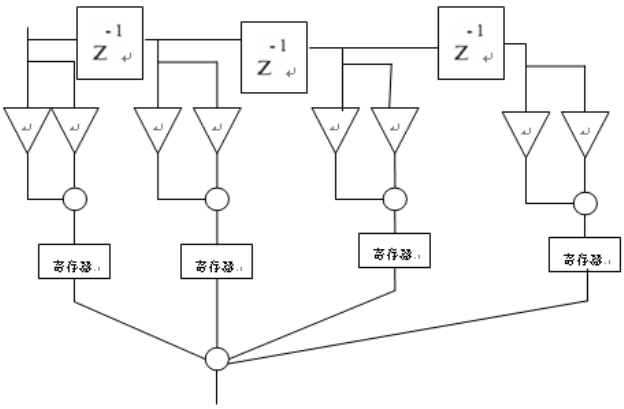


Figure 7 structure of the syste

4 Implementation Result

A multiplier with CSD code is design with Verilog HDL. Using a 16×6 ROM as the look-up table LUT, and then put it into EAB. In the design, we use 3 steps to complete the multiplication of the implementation process. The first step, download the operations and lookup table. The second step, use shift-add to complete the implementation of multiplication. The third step put the product into the output register, then compile and synthesis the program with MAX+PLUS II. At last, download it to EPF10K10. In Table 1, we can see the advantage after we use CSD code in the design.

Table 1 resource expend in the design

	Logic element used	Time used
Don't use CSD code	90	2.3μs
Use CSD code	50	0.9μs

From the analysis, we can see that, using CSD code in the design of multiplier can save hardware resource and increase computing speed. According to experimental statistics, the use of CSD code in the multiplier design can reduce 33% of the occupied resources.

5 Conclusions

In this paper, methods of digital filter design with FPGA are introduced. And some Special skills are included. CSD code is used in the multiplier's coefficient. With this method, we can use less adders and subtracters. In this way, the hardware resources can be saved. And also, it's quit a good way to increase the multiplier's speed. With pipeline, the system's throughput can be increased to a large extent.

References

[1] Zhang Dengqi, Zhou Ting, "Design of IIR digital filter based on Matlab", Journal of Hunan Institute of Science and Technology, 20(3), 2007, pp.26-28

[2] Zhu Youlian, Tao Weige, "Implementation of IIR Digital Filter Based on FPGA", JOURNAL OF EEE, 29(2), 2007, pp.52-58

[3] Wang Xinhuan, "High Speed FIR Digital Filter Design Based on FPGA", Modern Electronic Technology" , 254(15), 2007, pp.184-187

[4] Joyce Van de Vegte, Fundamentals of Digital Signal Processing, Beijing: Electronics Industry Publishing House, 2004

[5] Emmanuel C. ifeachor, Barrie W.Jervis, Digital Signal Processing A Practical Approach, Beijing:Electronics Industry Publishing House, 2004

[6] Liang Ruiyu, Jiang Bing, "The Structure Design of An Efficient IIR Filter and Its FPGA Implementation", JOURNAL OF HOHAI UNIVERSITY CHANGZHOU, 19 (3), 2005, pp.39-41

[7] Zhang Chi, Guo Lili, Sun Yan, "A high speed FIR filter design based on MATLAB and FPGA device", Information Technology, 7, 2006, pp.31-33

[8] Shi Hongyu, "IIR Digital Filter Realized by FPGA Based on Distributed Algorithm", mechanical management and development, 85(4), 2005, pp.67-69

[9] Wang Weibing, "Description of the FPGA of High step IIR Digital Wave Filter", modern Electronics Technology, 207(16), 2005, pp.3-5

[10] Zhang Degnqi, Zhou Ting, "Design of IIR digital filter based on Matlab", Journal of Hunan Institute of Science and Technology, 20(3), 2007, pp.26-29

Portability Analysis and Experiment of Open64 Compiler

Qiuhong Li[1] Zhongsheng Li[2]

1 Department of Information Wuxi Science Technology College, Wuxi, China 214083
Email: liqiuhong_ lili@163.com

2 Jiangnan Computing Technology Institution, Wuxi, China 214083
Email: lizhsh@163.com

Abstract

Open64 is an open source compiler system which comes from commercial compiler , It implements a set of integrated and robust compiler technology . Open64 is a good platform for use of reference both in research and in compiler development , because Opern64 compiler only has the version for IA-64 ,this paper do some research to see weather Open64 can be ported to other environment with good performance , and do some analysis about the transplanting. we do some experiments to verify that the Open64 compiler has good portability .

Keywords: Register Allocation; Instruction Scheduling, Intermediate Representation Inter-procedure Analysis, feedback

1 Introduction

Open64 compiler comes from an open source project named as SGI Pro64. Now the Open64 compiler is maintained by Delaware University and Alberta University , ORC project is developed by Intel and ICT which is extended from Open64. In 1999 , SGI launched the Por64 project to provide an advanced compiler for EPIC structure . Because Open64 comes from commercial compiler, it implemented an integrated and robust compiler techniques , including loop optimization and software pipelining .

2 Open64 Compile Introduction

Open64 compiler translates the source code of different languages including C/C++, Fortran90 into uniform middle format via the grammar and semantic analyzing programs for different languages at front end, On the basis of middle format, the compiler goes through inter-procedure analysis/optimization module, loop optimization/parallel module and global optimization module to reach the last code-generated module, then allocates register and schedules instructions, finally generates the binary coding for aim machine[4][5][6].

2.1 Front End

The front end of Open64 includes C/C++ front end and Fortran90/95 front end. C/C++ front end comes from GCC compiler, Fortran90/95 front end comes from MIPSpro Fortran compiler. Among these two front ends, only Fortran90/95 front end supports OpenMP.

2.2 Intermediate Representation

Open64 defines the intermediate representation WHIRL. There are five levels for WHIRL , they are super high level , high level, middle level, low level and super low level. WHIRL is a tree structure, The nodes include expression, statements and control unit such as loop and if statement. The aim for WHIRL is an abstract machine based on C language. The transformation of WHIRL contains as a subset of C language, which is irrelevant with special machine. This characteristics of WHIRL makes it easy that transplanting the compiler to another architecture. The work we need to do is only to change the CGIR module, It does not touch the WHIRL optimization.

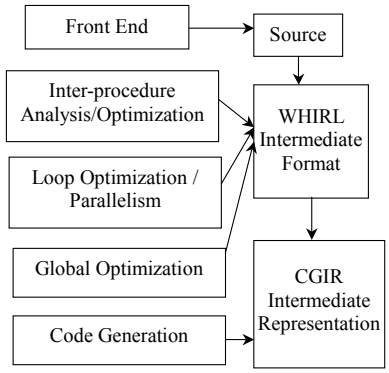


Figure 1 Open64 Compiler Framework

WHIRL is the main representation for Open64 compiler, It uses symbols to store the type information. But WHIRL can not include all information.. Open64 has other proper structure to store special information. The control flow info is embedded in WHIRL for WHIRL includes level control flow unit, Collecting the control info to form a sub language is the strategy of Open64 to avoid traversing all abstract grammar tree. Another important aspect of describing language is the definition and reference for variables. The according sub language is static single assignment, SSA is the main method in global optimization. In Open64, the relative graphics describing the same variable has three levels, LNO DDG represents the dependency among array visit. The last transformation of WHIRL tree is to transform the tree into the instruction sequences for special machine. Convert_WHIRL_To_Ops initializes the code generator. The code generator representation is the main one for CG. For instructions are part of basic module, CFG is implicitly in CGIR level.

A good compiler must provide tools to verify the correctness of optimization, First, IR can be transformed at different phrases, but it is not enough for terminal users, for terminal users can not understand the transformation store. The better way is that IR is transformed into original program language, keeps or recreates the WHIRL tree , So we can modify the ASCII WHIRL manually to debug the compiler.

2.3 Inter-procedure Analysis (IPA)

Open64 compiler supports inter-procedure analysis and optimization. IPA includes the info of some control unit, so that it can get something useful from global information. Inter-procedure analysis includes alias name analysis, array segment analysis and code distribution analysis. Inter-procedure optimization includes inline/cloning, useless function deleting, inter-procedure constant propagation, clarification of memory reference for accurate alias analysis. The implementation of IPA is based on the restructure of global abstract grammar tree. Therefore, Open64 compiler dumps the intermediate representation when generates the super high level WHIRL tree, The dumped info can be encapsulated into .o file even .a file. When the compiler links these files, it will do inter-procedure analysis and optimization if the compiler finds the WHIRL tree in .o or .a files.

The inter-procedure analysis of Open64 is transparent to users, You need not change the makefile file, It can provide alias analysis and process attribute for loop optimization, global optimization and code-generator.

2.4 Loop Nested Optimization/Parallelism (LNO)

LNO works in super high level of WHIRL, It can benefit from inter-procedure optimization. Loop nested optimization is suitable for all the languages it can support including OpenMP. Loops in programs are the source of many optimizations leading to performance improvements, particularly on modern high-performance architectures as well as vector and multithreaded systems. Loop optimization includes loop peeling, loop fusion ,loop spreading and loop swap, loop separation etc. Among the optimization techniques, loop peeling is an important technique that can be used to parallelize computations. The technique relies on moving computations in early iterations out of the loop body such that the remaining iterations can be executed in parallel. A key issue in applying loop peeling is the number of iterations that must be peeled off from the loop body. Current techniques use heuristics or ad hoc techniques to peel a fixed number of iterations or a speculated number of iterations. To our knowledge, no

formal or systematic technique that can be used by compilers to determine the number of iterations that must be peeled off based on the program characteristics. Loop fusion is a common optimization technique that takes several loops and combines them into a single large loop. The vector dependency info in LNO is passed to CG.

2.5 WHIRL Optimization (WOPT)

WOPT is the default optimizer of Open64 compiler. It shares an alias manager with code generator. WOPT adopts SSA as the unique representation of a program, Each optimization keeps the SSA format, If necessary, each optimization may repeat this process for several times, Open64 extends SSA :

Add alias representation and indirect memory reference representation; Integrate partial redundant code elimination.

Support speculative code move.

Raise register by adjusting load and store position.

2.6 Feedback

Open64 provides feedback mechanism, The compiler inserts some tracing code while compiling , then executes it in advance, It collects the run-time info through tracing code, and optimizes the code next time using the info. In Open64 compiler, you can insert instructions at any phases at back end, The inserted instructions mixes up with the original code, Compiler maintains the inserted code and check the consistency at the program transformation phases.

3 Transplanting Open64

Because Open64 adopts intermediate representation independent of machine, the optimization module of WHIRL can be transferred to different machine directly. The most work of transplanting Open64 compiler concentrates on machine description and code generation module.

3.1 Machine Description

The machine description files of Open64 are V11-ia64-extra.knb and V26-ia64-41-external.knb. These two files defines a mass of machine description information, The first step is to modify these two files. There are plentiful information of machine description in these two files. Below this paper introduces some main kinds of machine description information and shows how to transplant Open64 compiler to Alpha machine.

3.1.1 Instruction Description

An instruction can de defined as format string, for example, In order to add a lda instruction We first define the lda instruction:

opcode+=”0,lda,fullALL,alpha, 136,NULL “;

It shows that the instruction type is fullALL, the instruction format is 136,Below we define the 136 instruction format.

Opndsgrp+=”136,-litl6/OU_offset,-int64/OU_base, +int64”;

It shows the lda instruction format is :

Lda d_reg offset(s_reg).

You can define the executable unit for lda instruction,for example:

We define the executable unit is I_UNIT:

Unitprop[“lda”]:=bitmask(I_Unit);

3.1.2 Register Description

In alpha structure there are two classes known as integer and float, we need modify the parameters of original register type, For example we can define the integer register as:

Regclassprop+=”integer,all_isa_mask,32,64,1,0,int eger,$%d,REG_integer_special”;

REG_integer_special[“29”]:=”$gp”;

REG_integer_special[“30”]:=”$sp”;

It indicates that integer registers include 32 registers with 64 bit, and gives the name of two special registers .

According to special stipulation, define special classes as bellows :

Regprop+-"integer,allocatable,0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,\16,17,18,19,20,21,22,23,24,25,27,-1";

Regprop+-"integer,callee,9,10,11,12,13,14,26,30,-1";

Regprop+-"integer,caller,0,1,2,3,4,5,6,7,8,15 ,

16,17,18,19,20,21,22,23,24,25,27,28,29,-1";

Regprop+-"integer,func_arg,16,17,18,19,20,21,-1";

Regprop+-"integer,func_val,0,-1";

Regprop+-"integer,global_ptr,29,-1";

Regprop+-"integer,frame_ptr,15,-1";

Regprop+-"integer,stack_ptr,30,-1";

Regprop+-"integer,zero,31,-1";

Regprop+-"integer,ret_addr,26,-1";

3.1.3 Cache Structure Description

Accurate Cache description can improve the optimization effect of compiler, Open64 provides the method for definition of cache parameters at every level .The cache structure for alpha can be described as bellows:

Type cache_names_t = enum(L1I,L1D,L2);

CACHE_RelativeLevel[L1I]:=0;

CACHE_Content[L1I]=0;

CACHE_PolicyWrite[L1I]:=policy_write_other;

CACHE_PolicyRepl[L1I]:=policy_repl_lru;

CACHE_PolicyAlloc[L1I]:=policy_alloc_other;

CACHE_Lines[L1I]:=512;

CACHE_BytesPerLine[L1I]:=64;

CACHE_Ways[L1I]:=2;

CACHE_Ports[L1I]:=1;

CACHE_Port0AccessTypes[L1I]:=bitmask(access_read);

CACHE_ReadLatency[L1I]:=2;

Detailed definition for the meanings of classes can be found in the [3] document.

3.2 Kapi and Targ_info

Kapi is the API for handling machine description ,You can read the machine description file using kapi. You need not modify the code of kapi, but you can reference the kapi format when you modify the description files.

3.3 Code Generator (CG)

Entering the code generator module, the WHIRL must be degraded into machine instruction format. The compiler makes some optimizations relative to aim machine mainly including instruction scheduling and register allocating. All these optimizations are dependent on a series of machine description information. These machine description information comes from the two machine description files processed by kapi.

During transplanting, the most modification work concentrates on WHIRL_To_TOP. In WHIRL_To_TOP, the compiler transforms the WHIRL nodes into the operation code of aim machine, the transformed code is relative to the machine directly, bellows is an example. The example implements the transformation of a WHIRL node which is a integer constant.

With the –O0 option, above work can reach the requirement of transplanting, Of course ,These methods are not enough to acquire a high performance compiler. There are many optimization modules in CG which are related to architecture closely. The most typical module is software parallelism module (SWP), Open64 compiler adopts a scheduling algorithm names as Modulo which can not be found in alpha, so if you want to transplant SWP, you must consider it at the algorithm level.

4 Transplanting Experiment Results

We do some concrete transplanting experiments on a ALPHA21264(LINUX).Two benchmark programs which are matrix multiple and cg kernel loop can run in the transplanting environment ,The result is as bellows:

Table1 performance comparison between gcc and opencc

	scale	runtime	runtime
Matmul	1000*1000	gcc(-O3)	opencc(-O3)
cg	A	37.3	31.8

5 Summary

Open64 compiler is closely to commercial

compiler,It provices another good choice for compiler researches besides GCC, Our experiment is just a start , there are much work to do yet ,such as the transplanting of SWP . Otherwise, we need add optimization method for aim machine , It is part of the transplanting work.

References

[1] Richard A.Huff. Lifetime-Sensitive Modulo Scheduling SIGPLAN's PLDI Conference, June 1993

[2] Compaq Computer Corporation.Alpha Architecture Handbook October 1998

[3] Compaq Computer Corporation. Compiler Writer's Guide for the Alpha 21264 Jun 1999

[4] Appraisement conference files for IA-64 open source compiler system 2000.12

[5] Sebasitan Pop. Interface and Extension of the Open Research Compiler http://www-rocq.inria.fr/~pop/

[6] Guang R.Gao et. The SGI Pro64 Compiler Infrastructure http://www.capsl.udel.edu/~pro64

Safety Testing and Assessment of Software Based on Importance Sampling and AHP

Guozhu Liu[1] Junwei Du[2]

1 School of Information Science and Technology , Qingdao University of Science and Technology
Qingdao 266061,China
Email: lgz_0228@163.com

2 School of Information Science and Technology , Qingdao University of Science and Technology
Qingdao 266061,China
Email: d-jw@163.com

Abstract

According to different dangerous degree of failures of safety critical software, the safety criterion should be established scientifically firstly. Then, both the basic theory of importance Sampling (IS) and using Analytic Hierarchy Process (AHP) to get sampling probability are presented. This paper mainly discusses and studies in the combination of Importance Sampling and AHP. It is showed throughout the result of experiment that this methodology solves the problem of low testing efficiency of traditional safety critical software and unascertainable failure rate for other accelerated test techniques and doesn't affect the accurate result of safety assessment.

Keywords: Safety-Critical, Importance Sampling, safety testing, safety assessment

1 Introduction

Safety-critical software is needed in such fields as missile control, nuclear power plant, railway signaling system or space shuttle system where the failures of the system will cause many deaths or tremendous financial losses. Most of Chinese railway signaling systems are currently in the progress of carrying out the technology changing from relay-based to computer control interlocking, while some have already introduced Computer Interlocking System, but the users and the administrators always doubt whether the system can be safe, especially in some special situation.

Software reliability and safety are generally accepted as the key factors in software dependability since they quantify software failures that can make a powerful system inoperative or deadly [1]. According to software engineering, validation is an important phase in the life cycle of software development. Especially for safety critical software, no validation means less dependability.

The control and safety protection of Safety Critical System is an important field of computer application. The software of Safety Critical System is always called Safety Critical Software. Because once failure occurs, it may cause catastrophic result or fateful economic loss. Thus, how to assess safety of safety software has great significance.According to the theory of mutual complementarily between software failure rate vector and classified safety, safety of the safety software is evaluated on the basis of software failure rate vector. To assess and predict safety, statistic test result is currently needed. However, there exists very small probability events associated with causing catastrophic result event closely in input event set. If these events occur in the natural environment, vast time and test cases will waste for conventional statistic testing measure. Therefore, on these conditions, it is impossible to test high level safety criterion. In order to increase the testing efficiency of high level safety criterion, fault injection and error data injection etc. are put forward to accelerate the occurrence of these infrequent events. However, testing

cases are not able to show the practical operation state completely, and failure rate of testing is usually higher than the fact. So then the safety of the safety software can not be assessed exactly and effectively. Importance Sampling is not only used to expose more deficiencies in finite time, but can transform the failure rate of testing to the original rate by calculating accurately the amplification coefficient. Taking railway signal control software for instance, how to assess and test based on Importance Sampling and AHP are mainly discussed and researched in this paper.

2 Safety Testing and Assessment

2.1 Classification of Test Cases

As mentioned in the introduction, our method focuses on test results based on importance sampling of the safety critical software. However there is no clear frontier between 'robust' and 'non-robust' safety critical software, because safety can only be increased by reducing the risks inherent in software system instead of be guaranteed.

Document[7] conveys stress testing which is used to test safety critical software. In stress testing, it is intended to exercise rare conditions as much as possible.

Rough hierarchical testing adopts both routine testing and stress testing. In rough hierarchical testing, test cases include[8,9]:

- Level A. Regular test cases, without any abnormal inputs, are used to test the rudimentary functions which the software should meet according to specification;
- Level B. Allowable stress test cases, with one abnormal input or combined abnormal inputs which frequently happen in reality, are used to test the protective functions that are detailed in software specification.
- Level C. Light stress test cases, with one abnormal input or combined abnormal inputs which is a probable or occasional case, may cause critical or catastrophic consequences.
- Level D. Strong stress test cases, with one abnormal input or combined abnormal inputs which have a remote or improbable possibility, may cause critical or catastrophic consequence.

The collection of stress test cases is significant and always the key of the testing efficiency. It can be abstracted from expert experiences and classified by its frequency of occurrence and consequence of failure in reality. Since these concepts are fuzzy and can not be definitely supplied, the classification of test cases can only be rough. Here we classify the test cases into four levels. If there are more detailed requirements in software performance, the test cases can be classified into more levels.

2.2 Safety assessment Critical

Safety testing of software is established on the basis of occurrence and dangerous degree analysis of software failure throughout the testing process. Software failures which appear in unique environment will lead to unexpected state of a running system. Without effective fault-tolerance measure in time, one failure will appear. It may cut down the software safety, and different software failure has different effect on safety.

According to classification rule of safety software dangerous degree, safety failures of safety software are sorted into four degrees [3]:

- Catastrophic (C):

Personnel death, fateful economic loss, system reject, or out of control

- Serious(S)

Serious personnel injured, occupational disease, property loss, or system damage

- Light(L)

Slight wounds, light occupational disease, property loss or system damage

- Slight(SL)

Less loss than Light degree

In terms of definition of software safety, safety can be measured by safety rate, failure rate, mean time between accidents (MTBA) software accident rate (SAR), etc. However, these safety criterions may not assess safety well, because effect of different safety software on safety is not necessarily identical. In light of

four dangerous degrees, their corresponding software failure rate parameter constitute 4-dimension vector. Thereby, software failure rate vector can be used to represent classified safety of software. Assumed that ξ_1, ξ_2, ξ_3, ξ_4 are catastrophic, serious, light, slight accident probability respectively, then software failure probability vector $(\xi_1,\xi_2,\xi_3,\xi_4)$ can be generated.

3 Overview of Importance Sampling

3.1 Basic Conception

Assume that a random variable X has probability density function (*p.d.f*) $f(x)$ and that $Y=h(X)$ is a function of X. Our goal is to estimate the expected value of *Y*,

$$E(Y)=E(h(X))=\int_{-\infty}^{+\infty}h(x)f(x)dx \tag{3-1}$$

Through sampling, that is, we generate a sample $\{x_1,x_2,x_3,\ldots,x_n\}$ according to $f(x)$, therefore generating $\{y_1,y_2,y_3,\ldots,y_n\}$, and then calculate

$$E(Y)=\tilde{Y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}\sum_{i=1}^{n}h(x_i) \tag{3-2}$$

The sampling testing of *p.d.f* $f(x)$ is traditional statistic methodology essentially. Therefore, a large number of testing samples are needed when generating high reliability expected value of Y. However, if we change the original *p.d.f* by enlarging small probability and occurrence probabilities of catastrophic incidents, sample size will be effectively reduced.

In importance sampling, we change *p.d.f* of *X* from $f(x)$ to $g(x)$ such that those X's which are of importance in our parameter estimation have higher occurrence probabilities in $g(x)$. We use X to represent the variable which has *p.d.f* $g(x)$. By Equation (3-1), we have

$$E(Y)=\int_{-\infty}^{+\infty}h(x)f(x)dx=\int_{-\infty}^{+\infty}h(x)\frac{f(x)}{g(x)}g(x)dx$$

where $k(x)=\dfrac{f(x)}{g(x)}$, is called *likelihood ratio.*

$$E(Y)=\int_{-\infty}^{+\infty}h(x)k(x)g(x)dx \tag{3-3}$$

Let $Y'=h(X)k(X)$ then Equation (3-3) becomes

$$E(Y)=\int_{-\infty}^{+\infty}y'g(x)dx=E[Y']$$

Thus, instead of sampling from $f(x)$ to estimate the expected value of *Y*, the experiment is changed to sampling from $g(x)$ to estimate the expected value of Y'. That is, we generate a sample $\{x_1,x_2,x_3,\ldots,x_n\}$ according to $g(x)$, therefore generating $\{y_1,y_2,y_3,\ldots,y_n\}$, and then calculate

$$\tilde{Y}'=\frac{1}{n}\sum_{i=1}^{n}y_i'=\frac{1}{n}h(x_i')k(x_i') \tag{3-4}$$

- Definition 1：

Software operation route and software input have one-by-one corresponding relations. So, probability distribution of various inputs of software, which appear in whole set D, is called *software operation profile*.

$$OP=\{<prob_j,I_j>;I_j\in D\},\quad \sum_{I\in D}prob_j=1$$

- Definition 2：

Probability distribution of various testing inputs of software, which appear in whole set D, is called *software testing profile*.

$$TP=\{<d_j,I_j>;I_j\in D\},\quad \sum_{I\in D}d_j=1$$

Equation (3-4) shows that failure rate of software testing profile after changing probability distribution can transform into one of normal operation profile.

3.2 Application in Testing & Assessment

According to failure serious degree of four different dangerous degree of safety software, we divide input set D of testing data into four different set (D_1,D_2,D_3,D_4). They respectively stand for input set of catastrophic, serious, light and slight consequences, and their corresponding input occurrence probability are θ_1, θ_2, θ_3, θ_4. We have

$$\sum_{j\in D_1}prob_j+\sum_{j\in D_2}prob_j+\sum_{j\in D_3}prob_j+\sum_{j\in D_4}prob_j$$
$$=\theta_1+\theta_2+\theta_3+\theta_4=1$$

Generally, $\theta_1<<\theta_2<<\theta_3<<\theta_4$.

On the principle of importance sampling, occurrence probability of each task in input set should be changed. For serious hazard input, it should be enlarged in testing process. Contrarily, reduction of

occurrence probability should be carried out for light or no hazard input. Assume that $\theta_i^{'}(i=1,2,3,4)$ is occurrence probability of each degree in input set and then calculate,

$$k_i = \frac{\theta_i}{\theta_i^{'}} \quad (i=1,2,3,4) \tag{3-5}$$

And then assume that $\xi_i^{'}(i=1,2,3,4)$ is failure probability of each set under Importance Sampling testing profile. In this way, each operational failure probability under normal operational profile can be obtained by equitation (3-4).

$$\xi_i = k_i \cdot \xi_i^{'} \tag{3-6}$$

Sampling testing starts with probability distribution of $\theta_i^{'}(i=1,2,3,4)$. According to Bayesian statistic, we can deduce that *p.d.f* of $\xi_i^{'}(i=1,2,3,4)$ has *transcendental distribution* function $f(\xi_i^{'})$.

$$f(\xi_i^{'}) = \frac{(\xi_i^{'})^{a_i-1}(1-(\xi_i^{'})^{b_i-1})}{B(a_{i-1},b_{i-1})}$$

where $i=1,2,3,4, a_i>0, b_i>0; B(a_i,b_i)$ is *Beta* function. Then transcendental expected value of $\xi_i^{'}(i=1,2,3,4)$ is

$$\int_0^1 \xi_i^{'} f(\xi_i^{'}) d\xi_i^{'} = \frac{a_i}{a_i+b_i}$$

After every testing set has executed for n_i times and has occurred x_i failures, a *posterior distribution* of *p.d.f* of failure probability is

$$f(\xi_i^{'} \mid x_i, n_i, a_i, b_i) = \frac{(\xi_i^{'})^{a_i+n_i-1}(1-\xi_i^{'})^{b_i+n_i-x_i-1}}{B(a_i+x_i, b_i+n_i\_x_i)}$$

Then, posterior expected value of $\xi_i^{'}(i=1,2,3,4)$ is,

$$\xi_i^{'} = \frac{x_i+a_i}{n_i+a_i+b_i} \tag{3-7}$$

3.3 Identify $\theta_i^{'}$ using AHP

Through analysis of section 3.1 and 3.2, changed *p.d.f* $\theta_i^{'}$ has to be identified. Due to distinctness of dangerous degree after software failure, more serious consequences should have higher sampling frequency in testing set. In terms of this principle, different classified importance is quantified by AHP. Finally, changed *p.d.f* is calculated.

Above all, a hierarchy model should be constituted. In reference to respective importance distinction of integrative risk assessment and expertise, pairwise factors' ratio is determined by *1~9 ratio calibration arithmetic*.

Table 1 List of Judgment matrix calibrations and meanings

calibration	Meaning
1	Equality in importance
3	Slight (important than latter)
5	Distinct
7	Intense
9	Extreme
reciprocal	Contrary
2，4，6，8	Between odd numbers above

Table 2 Judgment matrix using AHP

O	C	S	L	SL
C	1	3	5	7
S	1/3	1	3	5
L	1/5	1/3	1	3
SL	1/7	1/5	1/3	1

By square root arithmetic [5], we obtain value of *q*: $\theta_1^{'}=0.56$, $\theta_2^{'}=0.26$, $\theta_3^{'}=0.12$, $\theta_4^{'}=0.06$.

4 Practical Application

Computer Interlocking software for railway is Safety critical software. Computer software may lead to input occurrence probability θ_i of dangerous degree under software operation profile, shown in Table 3.

Table 3 Input appearance probability of dangerous degree

Dangerous degree	Input appearance probability
C	10^{-5}
S	0.001
L	0.1
SL	0.89899

By Equation (3-5), we have $k_1=1.79\times10^{-5}$, $k_2=3.85\times10^{-3}$, $k_3=0.83$, $k_4=15$

Assume that one station has 10 tasks per hour. Referring to safety quantitative criterion of railway

safety control and protecting system in EN50129[6] (European Railway Standard), Table 4 shows the new safety criterion of the station.

Table 4 Classified quantitative criterion of dangerous output in interlocking system

Dangerous output degree	Acceptable failure rate
C	$<10^{-12}$
S	$<10^{-1}$
L	$<10^{-8}$
SL	$<10^{-6}$

4.1 Experimental Conclusion

Assume that $a_i = 1, b_i = 1 (i = 1,2,3,4)$ (without a priori information).

Table 5 shows the smallest sample size of traditional statistic and Importance Sampling testing method separately. It summarized that Importance Sampling can greatly reduce the sample size and increase testing efficiency effectively.

Table 5 Contrast of sample size of traditional statistic and Importance Sampling testing methods

Conspicuous level α	Traditional method	Importance sampling
0.3	1.2×10^{12}	1.76×10^{8}
0.25	1.4×10^{12}	2.2×10^{8}
0.2	1.6×10^{12}	2.48×10^{8}
0.15	1.9×10^{12}	2.93×10^{8}
0.1	2.3×10^{12}	3.55×10^{8}

4.2 Result of Assessment

Safety of interlocking software is assessed based on Importance Sampling. Assume that total number of input testing tasks is 10; the numbers of catastrophic, serious, light, slight failures are fixed as 0, 0, 1, and 5, respectively. By Equation (3-7), we can obtain failure rate under each testing profile of input set. And then we can also calculate the failure rate vector as (3.1×10^{-14} 1.4×10^{-11} 1.3×10^{-8} 1.2×10^{-6}) under operation profile by Equation (3-5). That vector shows catastrophic and serious accident rate can meet the requirement of safety quantitative criterion of interlocking software. However, it cannot meet the requirement of the light and slight accident rate. Therefore, by equation (2-3), where $c = 1$, and Table 4, we have $Rs = 2.671\times10^{-12} > 1\times10^{-12}$. According to integrative risk value, this interlocking software also cannot achieve the safety requirement.

5 Conclusion

By taking use of safety classified characteristic of safety software, this paper establishes scenario and criterion of safety assessment of interlocking software and mainly discuss the safety assessment based on Importance Sampling. Importance Sampling has two main contributions to high safety software. Firstly, it effectively reduces testing cost by accelerating the occurrences of rare probability incident. Secondly, it can accurately obtain the amplification coefficient of small probability incident.

References

[1] WU F, Huang L, "Analysis and Safety Assessment of Automatic Testing for Safety-Critical Software", *Proceedings of the 12th Asian Test Symposium (ATS 2003)*, IEEE, Xi'an, China, 2003

[2] WU Fang-Mei, *Testing and assessment of safety software for railway*, China Railway Press, Beijing, China, 2001, 128-130

[3] XU Zhong-Wei, WU Fang-Mei, "Quantitative Safety Assessment of Safety-Critical Software Based on Testing", *Computer Engineering & Science*, China,2001. 5

[4] Myron Hecht, Herbert Hecht, "Use of importance sampling and related techniques to measure very high reliability software", *Aerospace Conference Proceeding*, IEEE, 2000, (4), 533-546

[5] Saaty.T.L., *The analytic hierarchy process*, China Colliery Press, Beijing, China, 1988

[6] STANDARD EN 50129, *Railway Application: Safety Related Electronic Systems*, Ver. 08, 2001

[7] Kumar K.Goswami, Ravishankar K.Lyer. DEPEND:A Simulation-Based Environment for System Level Dependability Analysis. IEEE Trans. on Computer, 1997, Vol.46, No.1: P60-74

[8] William Perry. Effective Methods for Software Testing. Wiley,1995

[9] Tu Haiying, Li Weiwei, Xu Zhongwei and Wu Fangmei. T&AP on Microcomputer Interlocking Safety Software. CFTC-7, 1997,12. Guangzhou, China. P324~327

Simulated Moving Bed Modeling and Application

Wang Min

School of Information Technology, Jiangnan University, Wuxi 214036, China
Email: wangmin_cq@yahoo.com.cn

Abstract

Simulated moving bed (SMB) technology as an increasingly used chromatographic separation and preparation technique is reported in this paper. The principle, model, new development, optimization of SMB and major applications of SMB chromatography (SMBC) technology for separation are also reviewed. In addition, a new model solving method, the space-time conservation element and solution element (so-called CE/SE),is also discussed in a brief introduction. Applying the CE/SE method for SMB chromatographic problems, accurate solutions can be obtained and fast calculation can be achieved, respectively.

Keywords: SMB technology, chromatographic separation, SMB model, new progress, CE /SE method

1 Preface

Since high performance preparative chromato graphy (HPPC) is an effective facility being up against challenging of product cost、quality criterion and production efficiency in the yielding process of pharmacy、biology and chemical industry etc [1~2].

SMB chromatography (SMBC) ,one of the HPPC technics, is an primary continuous chromatography. In contrast to conventional batch chromatography's disadvantage of low utilize of absorbent、vast expending of solvent、low product's concentration、non-continuous operation and low efficiency, SMB chromatograph possesses advantages of low consume of absorbent、low cost investment、great impulse of mass transport、convenience in auto-control、mass preparation amounts and high preparation efficiency.

The SMB separation process model which is based on speed can better represent works in chromatographic column and includes partial differential equations, and solving about it is relative easy. Therefore, here, its equilibrium-diffusion model is introduced.

2 Principle of SMBC separation

The process of SMBC separation consists of serials of stationary bed chromatographic columns、inlet-outlet pipelines and valves[6]. The system of SMBC controls changing of inlet ports to "simulate" moving-bed by electromagnetic valves, therefore excellence of chromatography and advantages of moving bed, respectively, are obtained in this system. The SMBC system simulates relative countercurrent of stationary phase and mobile phase efficaciously by inlet-outletports periodically switching along direction of fluid flow [3~4].Fig.1 is the SMBC sketch map.

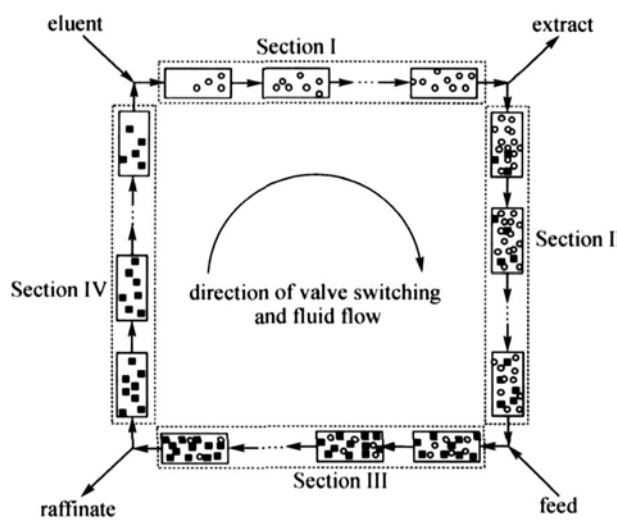


Figure 1 Schematic diagram of SMB chromatography○strong retained component weak retained component

The SMB system divides into four different

function sections. Section one for desorbing strong retained component; section two for desorbing weak retained component; section three for adsorbing strong retained component and section four for adsorbing weak retained component. Consequently, if we select reasonable setting parameters、sorbent、eluent solution and operational variables, weak retained component remain in the raffinate solutions and strong retained component remain in the extract solutions, respectively. Accordingly, the system achieves continuous separation. Simultaneity, High purity and desired concentration of strong and weak component also can obtained at the two outlet under the proper cooperating of four section's velocity [3~5].However, study of how to select reasonable parameters and various variables is theory of optimization.

3 Operation Optimization of SMB Process

In recent yeas, many scholars have deeply studied the optimization techniques of SMB process due to strong coupling of its separation process and complexity of its technical mechanism. Nowadays, study of operation optimization has become hot spot. Most of these studies generally based on early triangular theory [7] and genetic algorithm [9].Applying of optimization algorithm provide a different effective way for operation optimization of SMB process. Using optimization methods for SMB chromatographic problems, satisfactory performance targets and accurate solutions are obtained and fast calculation is achieved or other benefits acquired etc. These methods generally examine effects of some parameters upon the results with certain test examples. There are a mass of papers discuss optimization question from different aspects. Such as documents [1]、[5] and [8]~[15] introduced in detail. In addition there are lots of articles published in magazine of Journal of Chromatography、Chemical Engineering science、Computer & Chemistry and Chromatography etc. Below, I will introduce an optimization algorithm to discretize and solve the SMB model.

4 Discrimination and Solving Principle of SMB Model

Equilibrium-diffusion model, as mentioned earlier in this paper, consists of following equations.

Material equilibrium equation:

$$\frac{\partial C_i}{\partial t}+\frac{1-\varepsilon}{\varepsilon}\frac{\partial q_i}{\partial t}+V\frac{\partial C_i}{\partial z}=D_L\frac{\partial^2 C_i}{\partial z^2}$$

Here, substitute definition $F=\frac{1-\varepsilon}{\varepsilon}$, $x=\frac{z}{L}$, Pe = $\frac{V\cdot L}{D_L}$ in above equation and obtain equation:

$$\frac{L}{V}\frac{\partial C_i}{\partial t}+F\frac{L}{V}\frac{\partial q_i}{\partial t}+\frac{\partial C_i}{\partial x}=\frac{1}{Pe}\frac{\partial^2 C_i}{\partial x^2} \tag{1}$$

In condition of solute impetus of mass transport is linear,

Mass transport equation:

$$\frac{\partial q_i}{\partial t}=k\left(q_i^*-q_i\right) \tag{2}$$

Every component of chromatographic separation system consists of the above two partial different equations.

In the above Eq. (1) and Eq. (2) ,initial and boundary conditions are required.

initial condition:

$$C_i\left(x,t=0\right)=0 \quad 0\le x\le 1 \tag{3}$$

boundary conditions:

$$x=0 \quad \begin{cases} C_i-\frac{1}{Pe}\frac{\partial C_i}{\partial x}=C_i^0 & t\le\Delta t \\ C_i=0 & t>\Delta t \end{cases} \tag{4}$$

$$x=1 \qquad \frac{\partial C}{\partial x}=0 \tag{5}$$

Notation explanation as follows:

C — concentration in fluid phase, D_L — axial dispersion coefficient, V—velocity of flow in fluid phase, q—concentration in stationary phase, q*—q corresponding equilibrium to C(determined by adsorption isotherm), L — column length, z — axial direction of column, x — dimensionless column length(=x*L),k—lumped mass transport coefficient, t—time, i—number subscript of component type.

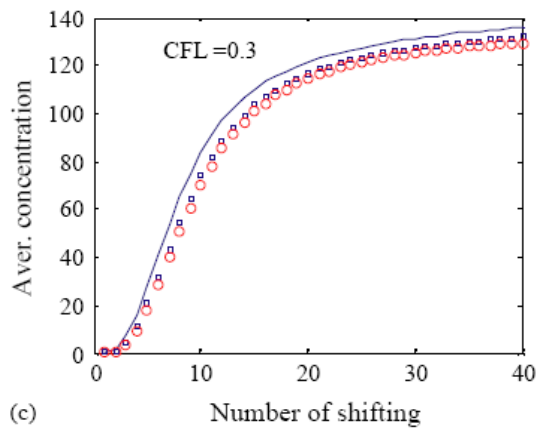
The partial differential equation system, which is often solved, after finite element partition of solving

domain, by orthogonal collocation method in every finite element. Here, finite element, finite difference methods are used for Discretization. However, solution procedure of this combinative method may be inadequate for multi-column, multi-component and multi-dimensional systems or be numerically dissipative. Thus, in this paper, a new numerical method, the so-called space-time conservation element and solution element (CE/SE), is proposed to enhance accuracy and computational efficiency. The CE/SE method is based on the Gauss' divergence theorem and enforces both local and global flux conservation in space and time to achieve a numerically non-dissipative feature. First, Equations are transformed using the Gauss' divergence theorem, then each integral variable can be approximated by a first-order Taylor expansion, then the time and space derivatives are reformulated through the chain rule, Finally, the unknowns are obtained from the known values[15].

In this way, The SMB model is solved with accuracy and a reduced computational time, using the CE/SE method. Ultimately, we can design emulate system to conduct example emulate on compute.

5 Performance Analyses of CE/SE Method

As mentioned in the above section, The space-time conservation element/solution element (CE/SE) method uses a simple stencil structure, which is two points at previous time level and one point at present time level, on staggered space-time grids (shown in Fig.2), and enforces flux conservation in space and time. Stable, accurate, fast solutions and simple treatments can be gained under some conditions being satisfied. Only two main parameters (CFL number and number of mesh points) affect the numerical solution results. I will examine their effects on the results through actual separation process examples in my following study, there I will acquire conclusions: CFL number affects little the numerical solution under somewhat high Peclet number and somewhat low Stanton number but sufficient number of mesh points are required; correspondingly, if Peclet number decreases and Stanton number increases, low CFL number is well and small number of mesh points is suitable. The above conclusions can easily be summarized by comparison between experimental and simulation results. Such as in the experiment of Fructose-Glucose SMB process, I change the CFL number and number of mesh points to observe the average concentration variation of fructose and glucose in the extract and raffinate solution, respectively. Similar concentration variation diagram as follows:



(c)

6 New progress of SMB technology

SMB technology, developed in age of sixty of twentieth century[6], acquired prodigious advancements through development in more than forty years(especially under recent ten years' fast development);Its theory gradually be perfected and new technology constantly turn up; It has been popularly applied in many fields.

SMB technology is prevalent day by day owing to its superiority by technique, as follows [16]:

1. SMB chromatography is easy to realize auto operation and steady control of products' quality due to its continuous separation process.

2. SMB technology owes higher separation efficiency than other preparative chromatography. One hand, SMB chromatography only need less feeds under the same demand of product's purity; On the other hand, research indicates that production capacity of SMB merely reduce 10% when efficiency of chromatographic

column reduce 20%,but the one of ordinary chromatography reduce 50%.

3. Separation process of isomer can enlarge quickly and reliably under analytical chromatographic conditions by SMB technology.

SMB chromatography as well as has modified operation model, subsequently, further study and process spring up. Temperature gradient SMB and solvent gradient SMB shift the solute's adsorption intensity in each zone through the change of temperature and solvent form, respectively, thereby the performance of separation capacity get perfected. Multi-components SMB increase counts of component by increasing counts of zones, for instance, adopting SMB system with four-zones and five zones to separate-purify three components structure. Power-feed Operation SMB change liquid velocity when inlet-outlet switching. VariCol Operation SMB is a multi-columns chromatography operation; It's principle is length of zone changes by time with non-coordination shifting of inlet-outlet pipeline. Supercritical fluid chromatography(SFC) is famous for using supercritical fluid as mobile phase, it can make single mobile phase used in multi-purpose separation through changing operation temperature or operation pressure [7]、[16].

7 Application of SMB

The first system of SMB was exploited by Universal Oil Products Co.(UOP) in age of sixty of twentieth century. UOP first used it to separate products of Oil [4]. SMB technology application areas expand ceaselessly along with its continually developing. SMB technology spread over the area of Oil 、Chemical-Industry 、 Biochemical 、 bio-ferment 、 pharmaceuticals、food industry etc. Particularly, SMB technology displays its unique performance in separation of mixture of hand isomer-pharmaceuticals、glucide、organic acid、amino acid etc. Meanwhile, it appears huge superiority on separating two-component system [2].

Nowadays, the industry of offering SMB technology has appeared in several advanced nations. UOP and AST of USA、NOVASEP united by France and Germany [4].

8 Conclusions

In recent years, SMB technology has gained great progress and its application fields still are expanding. Besides, scholars pay great attention to the study of separating high additional value products of biomedicine、fine chemical industry by device of SMBC separation day by day. Large SMB equipment's preparation amounts can achieve megaton-level per year. Simultaneity, it just consumes little solvent, can saves solvent above 90%.SMBC will certainly be quite useful in a good many fields such as oil、biochemical、food industry and pharmaceuticals (especially in medication by hand),as well as owe wide application foregrounds. So constantly deep study of SMB has significant importance to industry and society.

The advantages and values of SMB technology are obvious to all. However, problems about it still exist. Forecast 、optimize and control to it is complicated because it is a dynamic and nonlinear system. Besides, Application range of SMB is restricted because it is a two-component separation device; the equipment is so complex that it is hard to use and maintain. In our country, study and application of preparative chromatographic technology is weak and the preparative scale is small because of our economy and technology. Therefore, the SMB technology still need to energetically develop; It is necessary to strengthen study、develop、extend the work about SMB technology (especially our country).

References

[1] WU Xian-dong etc. Application of Particle Swarm Optimization in process of Non-linear Simulated Moving Bed Chromatographic Fractionation. Control and Instruments in Chemical Industry. 2006.33(4)

[2] CAI Yujie etc. Simulated Moving Bed Technology and Its Applications. Chinese Journal of Chromatography. 2004

[3] ZHANG Hong-li etc. The Advance of Simulated Moving

Bed Technology.Analysis and Testing Technology and Instruments. 2005

[4] LIN Bing-chang. Application of SMBC Technology for Separation of Effective Components in Chinese Herb. Fine Chemicals. 2005

[5] WU Xiandong etc. Multi-objective optimization of simulated moving bed chromatography separation based on NSGA-Ⅱ algorithm.Journal of Chemical Industry and Engineering (China). 2007

[6] Ruthven D M etc. Counter-current and simulated counter current adsorption separation process. Chemical Engineering Science, 1989

[7] WANG Xuejun etc. Process of Preparative Chromatography. Journal of Qin-Dao University. 2001

[8] MAXXOTTI etc. Optimal Operation of Simulated Moving Bed Units for Nonlinear Chromatography Separation[J]. Journal of Chromatography A. 1997

[9] Michalewicz Z. Genetic Algorithms Data Structures Evolution Programs, AI Series. New York: Springer Verlag,1995

[10] CAI Yujie etc. Optimization of Conditions of Simulated Moving Bed Chromatography for Separating Mother Liquid of Xylitol with Real Coded Genetic Algorithms. Chinese Journal of Chromatography. 2003

[11] CHEN Xueguo etc. Optimization of Separation Conditions of Chromatographic linear gradient with Anneal Genetic Algorithm . Chinese Journal of Chromatography. 2004

[12] XU Ling etc. Study of Optimal Design of Simulated Moving Bed Chromatographic Separation Process. Control and Instruments in Chinese Industry. 2004

[13] LIU Yanchao etc. Optimal Design of Simulated Moving Bed Chromatographic Separation Process. Journal of Southern Yangtze University. 2004

[14] HE Fan etc. Modeling and Optimization for the Xylose and Xylitol SMB Chromatographic Separation Process. Journal of Wuxi University of Light Industry.2002

[15] Young-I1 Lim etc. A fast and accurate numerical method for solving simulated moving bed (SMB) chromatographic separation problems. Chemical Engineering Science . 2004

[16] High Performance Preparative Chromatography (HPPC)

Moving Target Removal in Video Sequence Using Boundary Tracking

Fei Chen[1] Xunxun Zeng[2] Meiqing Wang[2]

1 School of Sciences, Jimei University, Xiamen, Fujian, 361021, China
Email: chenfei0390@sina.com

2 College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian 350002, China
Email: xun.xun@163.com, mq_wang@163.com

Abstract

Moving target removal as one of the video special effect techniques refers to removing a foreground moving object from background and filling in the missing regions in a visually consistent manner. Usually for fill-in, we have to manually create masks for moving targets frame by frame by human intervention. This paper introduces a convenient algorithm to implement moving target removal from constrained complex background in a video sequence. We start from a given mask of the moving target in the first frame, and propose a new method which combines frame difference with boundary tracking to automatically generate moving target masks for other frames, then use image synthesis and PDE-based inpainting to repair the region of the moving target in the first frame. Finally, we fill the mask regions of moving target in other frames to finish the removal of moving target in video sequence.

Keywords：Target removal; boundary tracking; frame difference; image synthesis; PDE

1 Introduction

Moving target removal refers to the process of removing a foreground moving target of a video sequence and filling in the missing regions in a visually consistent manner. Recently, as one of the key techniques in post-processing [1] it has many important applications in movie and film making industry. For example, some common digital special effects are composition of graphic objects, such as removal of unnecessary objects from a movie, or insertion of virtual objects to a video sequence. Thus, moving target removal attracts the attention of many researchers [1,2,3,4]. In[1], Soon-Yong et al. introduced a background completion technique to remove a moving foreground object in a video sequence which obtained by using a moving camera. Kedar A. et al. [2] presented a framework of video inpainting, and proposed a solution under constrained camera moving.

In the early research, the masks of moving objects were created manually frame by frame using graphic software by human intervention to erase the moving target in a video sequence, and the empty target regions were filled according to the background information from other frames. Based on this reason, earlier methods are inconvenient and very time-consuming.

In this paper we address a constrained but important case of moving target removal. We assume that the camera is stationary, the scene essentially consists of stationary or complex background with a moving foreground, and the intersection of moving foreground targets in two consecutive frames is not empty which means the speed of the moving target cannot be too fast. As we will see below, these assumptions are explicitly present in Section 4.

The method introduced in this paper is able to conveniently remove the object that moves in any fashion and can change scale as it moves. Starting from

*This work was partially supported by NSFC under Grant No.10771036

a given mask of a moving object in the first frame, the proposed method employs frame difference and boundary tracking to automatically generate different masks for other frames, and fills in the mask region of every frame with appropriate other frame information to implement the removal of moving target.

This paper is organized as follows. In Section 2 we briefly describe the frame difference and boundary tracking. Section 3 introduces the PDE-based inpainting. In Section 4 we present the convenient algorithm of moving target removal. Section 5 shows the experiment results for two video sequences. Section 6 contains conclusions.

2 Frame difference and boundary tracking

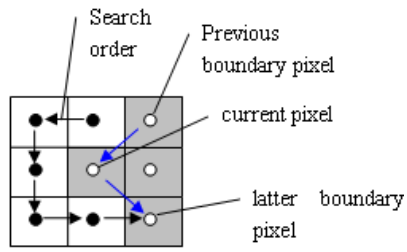
Frame difference technique is used to detect moving objects, which is simply finding the absolute difference between two consecutive frames. Suppose the i th frame image in original video sequence is $F_i\ (i=0,1,\cdots,n)$, the frame difference is DF_i, defined as follow:

$$DF_i = F_i - F_{i-1} \qquad (1)$$

To get the definite moving area, the images in the video sequence are converted into binary images by a threshold value T, which is a constant that may be adjusted for a specific application [5]. If the absolute difference between the pixel values is greater than T, the movement has been significant and the pixel is called a boundary pixel whose value is set to 1 . Otherwise, if the change is less than this threshold, the pixel value is set to 0. The obtained binary image FDM_i is called Frame Difference Mask (FDM), defined as the following formula:

$$FDM_i = \mathrm{BW}(DF_i) = \begin{cases} 0 & |DF_i| < T \\ 1 & |DF_i| \geq T \end{cases} \qquad (2)$$

Boundary tracking is an effective method to locate the edges of the target in an image. It has many important applications in the area of digital image segmentation and visual analysis [6]. Boundary tracking can be easily applied to binary image. Once a single boundary point is found, the operation seeks to find all other pixels on the boundary of objects.



(0-pixel: black point, 1-pixel: white point)

Figure1 Find next boundary pixel

One 8-neighbors boundary tracking approach is described as follows:

1. Find first boundary pixel.

2. Check the 8-neighbors of the current pixel and go to the first boundary pixel found in anti-clockwise order. Set this boundary pixel as the current pixel

3. Stop if this is the first boundary pixel.

4. Check the 8-neighbors of the current pixel until meeting the first boundary pixel found starting from the previous boundary pixel in anti-clockwise order (see Figure 1).

5. Go to step 3.

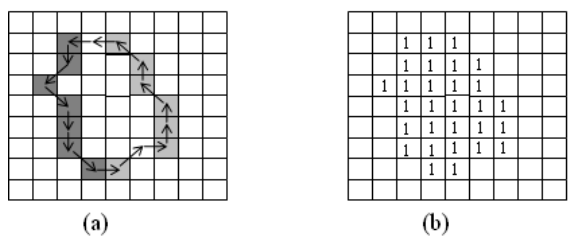


Figure2 Restore the region from boundary coordinates

After the determination of boundary, we can restore the region in the binary image by boundary coordinates. It consists of two steps:

1. Determine the left boundary and right boundary. In Fig. 2(a), darkish pixels denote the left boundary pixels, and grayish pixels are the right boundary pixels.

2. Convert those pixels between left boundary and

right boundary from 0-pixel to 1-pixel (see Figure2(b)).

3 PDE-based inpainting

Inpainting is a technique of modifying an image by filling in the missing information on prescribed domains in an undetectable form. Image inpainting was first introduced by Bertalmio et al. [7, 8] involves the use of a partial differential equation (PDE) model based on the transport theory. The idea used in this model is to smoothly propagate information from the surrounding areas in the isophote directions. Chan and Shen [9, 10] extended the classical TV denoising model of Rudin-Osher-Fatemi [11] and developed the total variation (TV) inpainting model from the point of view of variational principles and image prior models. Later, they considered the connectivity principle, and proposed the curvature driven diffusions (CDD) inpainting model [12] by adding the curvature of the isophotes to the TV conductivity coefficient. In [13], Chan and Shen develop Euler elastica (EE) inpainting algorithm based on connecting appropriate level lines by Euler elastica curves. This method turns out to be a generalization of the transportation mechanism of BSCB model and the CDD model. The PDE-based inpainting methods can be used to repair non-texture image, and have many interesting applications [7, 10].

Because the TV inpainting is easy to be numerically implemented and works well with small narrow inpainting domains, we will apply TV inpainting to the removal of moving target. Let Ω be the entire domain of the image. D denotes the inpainting domain where the image information is missing or damaged. ∂D is the inpainting boundary $\partial D \in \Omega \backslash D$, and $\Omega \backslash D$ denotes the available part of the original image z on Ω and u is the restored image (see Figure3).

Figure3 Inpainting the domain D by filling the missing information

TV inpainting methods are used to find a solution u on the inpainting domain D by using information from the domain $\Omega \backslash D$ by minimizing TV function,

$$\min E[u] = \int_D |\nabla u| dxdy \quad \text{such that} \quad u = z|_{\partial D} \quad (3)$$

4 Moving target removal

Our basic approach to remove a moving target is outlined in Fig. 3. First, we obtain the frame difference masks from the original video sequence by frame difference. Next, we employ boundary tracking to automatically generate moving target masks for other frames by a given mask of the first frame. In order to make the image boundary smooth and continuous and easy to track, some filtering operations can be used. Then, we fill the given mask region of moving target in the first frame with appropriate information from other frames, which is called 'image synthesis'. If the mask region can not be completely filled in, we use PDE-based inpainting technology for the remaining parts. Finally, the regions of moving target of other frames are filled by synthesizing useful information from the previous frames, and a new video sequence is generated. Fig. 3 shows a flow diagram of our system.

In this section, we firstly present the moving target removal from static background, then we consider the constrained complex background with moving objects, and apply frame difference combining with boundary tracking to the moving target removal, finally we give the implement algorithm of our method.

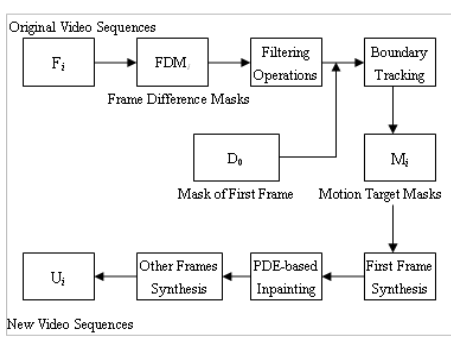


Figure3 Diagram of moving target removal system

4.1 Moving target removal from stationary background

We firstly assume our problem is in a stationary background with a moving target. Therefore, it is easy to track the target using frame difference (see section2). In most cases, image or video resources are often received in poor condition, mostly with noise or defects making the resources hard to read. So there still exist many useless noise spots in the background and foreground, after frame difference images are binarized.

In this paper, the mathematical morphology is used to process the binary images. The two most basic algorithms of mathematical morphology are dilation and erosion. The basic effect of dilation operator on a binary image is to change some background pixels adjacent to the target region from 0-pixel to 1-pixel. After that the region is expanded and smoothed, meanwhile inner gaps are filled. Oppositely, erosion can change some pixels from 1-pixel to 0-pixel. It causes the region to shrink, at the same time, it can remove noises. A combination of dilation with erosion can form an effective noise filter. After denoising, we get the new binary image sequence denoted by moving target masks (MTMs).

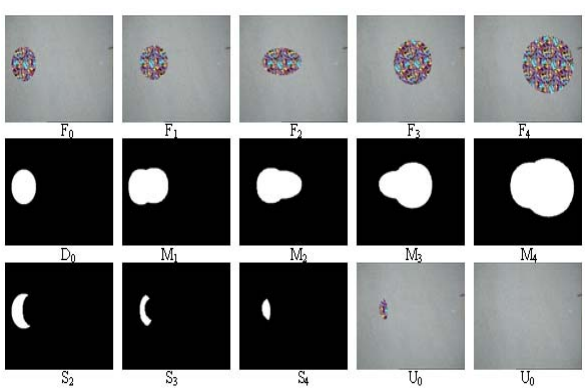


Figure4 A simple example of moving ball tracking

For a binary image(0: background, 1: object), we can consider the 1-pixels to all comprise a set of values from the "universe" of pixels in the image. Throughout the paper, when we refer to an binary image A, we mean the set of 1-pixels in that image. Therefore, we can use standard set notation to describe binary image opration: intersection ($\cap$) and minus ($-$). Let M_i be the i th MTM, and D_0 the original given binary mask of the moving target in the first frame (see Fig. 4). We define D_i as the intersection of the binary images M_i and D_i as Eq.(4).

$$D_i = M_i \cap D_{i-1} \quad i = 1,2,\cdots,n \qquad (4)$$

If D_i is empty, it means the region D_0 of the first frame has been filled completely by synthesizing the first i frames. Then we define S_i as the difference between D_i and M_i (the 1-pixels in D_i that aren't in M_i) as following (see Figure4):

$$S_i = D_{i-1} - M_i \qquad i = 1,2,\cdots,n \qquad (5)$$

When S_i is empty, it means the i th MTM makes no contribution to the first frame.

Let U_i be the i th frame image in the new video sequence. We take two steps to fill the missing part in video sequence caused by the removal of unwanted objects.

1. Inpaint the first frame by synthesizing other frames.

$$U_0 = F(0,\Omega\backslash D_0) \oplus F(1,S_2) \oplus F(2,S_3) \oplus \cdots \oplus F(k-1,S_k) \quad (6)$$

Where $F(i,S_{i+1})$ represents the pixels of the region S_{i+1} of the i th frame in the original video sequence, and the operator ($\oplus$) is used to union these regions together.

If D_k is not an empty set, we employ total variation inpainting to the remaining unfilled region D_k of the first frame (see the last two images in Fig. 4), described as follows:

$$U_0 = U_0(0,\Omega\backslash D_k) \oplus \mathrm{TV}(0,D_k) \qquad (7)$$

2. Fill in the MTM regions of other frames by synthesizing useful information from the previous frames. This can be written as

$$U_i = U(i-1,M_i) \oplus F(i,\Omega\backslash M_i) \quad i = 1,2,\cdots,n \qquad (8)$$

The algorithm of the moving target removal is as follows:

Alg.1 MovingTargetRemovalFromStaticBackground ( )

D[1]←D0
For** i=2 **to** n **do
MinusBinaryImage(D[i-1], M[i], S[i]);//Si=Di-1-Mi
MinusBinaryImage (D[i-1], S[i], D[i]);// Di=Di-1-Si
If** IsEmptySet(D[i]) **then
***Break**;*

```
End do
For j=2 to i do
    UnionImagesByMask (F[j-1], S[j], F[0]);
End do
If  IsNotEmptySet(D[i])  then
TotalvariationInpainting(F[0], D[i], iterations);
For t=1 to n do
      UnionImagesByMask(F[t-1], M[t], F[t]);
End do
```

4.2 Moving target removal from constrained complex background

Here, we consider the constrained complex background with moving objects. Before we go forward, we give two assumptions in our solution.

1. Assume that the intersection of foreground moving targets in two consecutive frames is not empty.

2. If background contains many moving objects, we suppose the foreground moving target do not interact with background moving objects.

Then we propose a solution to remove moving target from the constrained complex background. Notice that our problem involves several difficulties, e.g., how to track a moving target by a given mask, frame difference masks we got sometimes are not integrated, there may be holes in FDM.

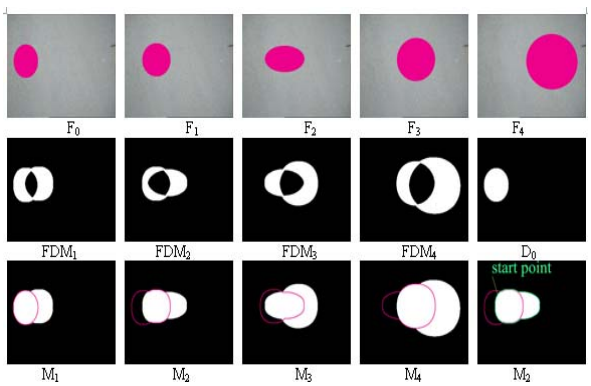


Figure 5 A simple example of tracking a simple color moving ball

In the example above, we consider another case. If the moving target is simple colore, we will get a big hole in the FDM (see Figure. 5), because that frame difference does not always get enough information. There are many methods can be used to solve this problem. However, because a foreground target has a distinct color distribution compared to the surrounding background, we use boundary tracking to track the positions of the moving objects and create the corresponding MTMs. Moreover, boundary tracking is also an effective method to fill in the inner holes in MTMs. Suppose we obtain appropriate FDMs by using frame difference and mathematical morphology filter, our algorithm consists of the following steps for boundary tracking (see Fig. 5): finding the start point (the first boundary pixel) in the current frame which depends on the mask region of the previous frame (this step is due to our assumption 1), tracking all the boundary pixels and restoring the so-called MTM regions from the boundary coordinates in the current frame, repeating the above steps to obtain all the MTMs frame by frame.

A similar technique is employed in the moving target tracking from complex background which contains many moving objects. Fig. 6(a) is one frame from traffic sequence. We can see many cars are moving. If we choose the left black moving car as the foreground target, other background moving cars will affect the moving target tracking by using frame difference, see Fig. 6(b). After boundary tarcking, they may be not integrated in MTM, see Fig. 6(c). Aiming at this case, we apply morphology filter to smooth boundaries, remove short branches and remedy the narrow gaps of boundaries in order to make integrated MTM, see Fig. 6(d). Fig. 6(e) is the new frame generated by using our algorithm (see Alg. 2) to implement the moving target removal.

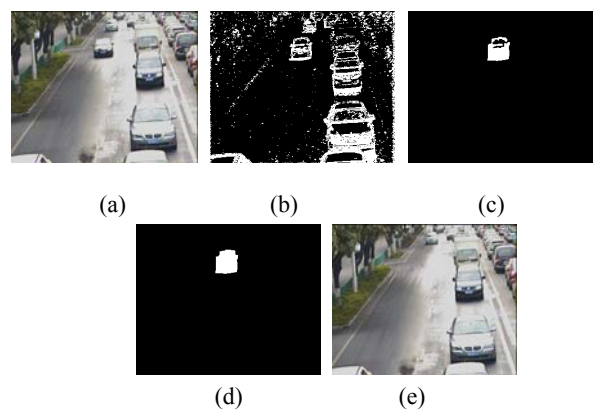


Figure6 One frame of the traffic sequence

Alg.2 MovingObjectRemovalFromComplexBackground()

```
For t=1 to n do
      FrameDifference(F[t],F[t-1],DF);
GaussFlatImage(DF);
      ConvertToFDM(DF, FDM[t]);
```

```
            Dilation (FDM[t]);
            Erosion (FDM[t]);
        CopyBinaryImage (FDM[t], M[t]);
End do
    For t=1 to n do
      a=0,b=0;
            FindStartPoint(M[t], M[t-1], &a, &b);
            BoundaryTracking(M[t], a, b);
        RestorationByBoundary(M[t]);
            Dilation (M[t]);
        Erosion (M[t]);
End do
MovingTargetRemovalFromStaticBackground ( )
```

5 Experimental Results

We adopt CCD camera and image collection card to shoot the moving objects and gather a series of images named by video sequence and store these video frames as BMP format. These frames are used to evaluate processing methods and algorithms subjectively and objectively.

We carried out two experiments with our algorithm and successfully erased the moving object from the video sequence. Figure. 7 shows the results of the first experiment by using a video traffic sequence. We want to remove a left black car from the sequence. The 1st two rows are ten frames from the original video sequence (50 frames). We notice that the size of the car changes from far to near. The 2nd two rows are FDMs, which reflect all the tracks of the moving objects. The 3rd two rows are MTMs. The last two rows are new video sequence obtained by our methods. In the second experiment we want to remove a person from a scene. In Figure. 8, the 1st two rows are ten frames from the original video sequence. In the last two rows the moving person is successfully filled in the object areas in all frames.

Figure7 Traffic sequence: the 1st two rows are ten frames from the original video sequence (50 frames). The 2nd two rows are FDMs. The 3rd two rows are MTMs. The last two rows are new video sequence.

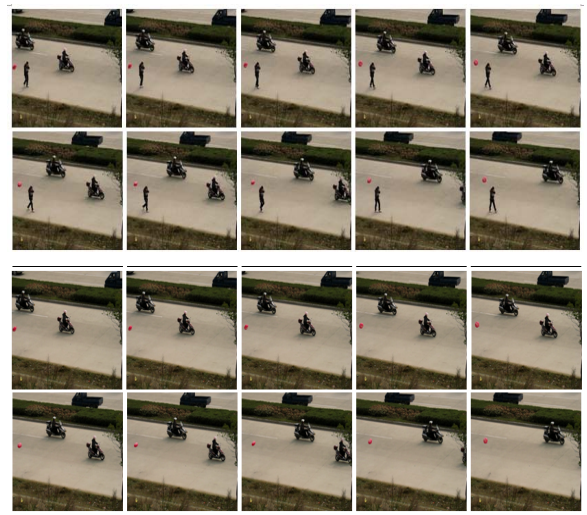


Figure 8 Removal of a moving person: the 1st two rows are ten frames from the original video sequence. The last two rows are new video sequence.

6 Conclusions

A convenient algorithm based on frame difference and boundary tracking for moving target removal from constrained complex background has been presented. It consists of the following steps: computing appropriate FDMs by using frame difference and mathematical morphology filter, automatically generating MTMs for other frames by boundary tracking according to the given mask of the first frame, using image synthesis and PDE-based inpainting to repair the region of the moving target in the first frame, filling in other frames to finish the removal of moving target in video sequence. This algorithm can be further extended for removing multi-targets with constrained complex background.

References

[1] Soon-Yong Park , Chang-Joon Park , and Inho Lee, "Moving Object Removal and Background Completion in a Video Sequence", Image and Vision Computing New Zealand, 2005

[2] K. A. Patwardhan, G. Sapiro, and M. Bertalmio, "Video Inpainting Under Constrained Camera Moving", IEEE Transactions on Image Processing, 16(2), 2007, pp:545-553

[3] Sen-Ching S. Cheung, Jian Zhao, M. Vijay Venkatesh, "Efficient Object-Based Video Inpainting", ICIP 2006, pp:705-708

[4] Y.-T. Jia, S.-M. Hu, and R. R. Martin, "Video completion using tracking and fragment merging" The Visual Computer, 21(8-10),2005,pp:601-610

[5] Yasushi Mae, Yoshiaki Shirai, Jun Miura, "Object Tracking in Cluttered Background Based on Optical Flow and Edges", In Proc. of 13th ICPR, 1996, pp:196-200

[6] Pratt W K, Digital Image Processing, 2nd Ed. NY:John Wiley & Sons, 1991

[7] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image Inpainting", In Proc. ACM Conf. Computer Graphics (SIGGRAPH), 2000, pp:417-424

[8] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, "Navier-Stokes, Fluid Dynamics, and Image and Video Inpainting", In Proc. ICCV 2001, IEEE CS Press 1., 2001,pp:1335–1362

[9] T. F. Chan and J. Shen, "Variational restoration of non-flat image features: models and algorithms", SIAM J. Appl. Math., 2001,61(4),pp:1338–1361

[10] T. F. Chan and J. Shen, "Mathematical models for local non-texture inpaintings", SIAM J. Appl. Math., 2001, 62(3),pp: 1019–1043

[11] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms", Physica D, 1992, 60, pp: 259–268

[12] T. F. Chan and J. Shen, "Non-texture inpainting by curvature driven diffusions(CDD)", J. Visual Comm. Image Rep., 2001, 12(4),pp: 436–449

[13] T. F. Chan, S.-H. Kang, and J. Shen, "Euler's elastica and curvature based inpaintings", SIAM J. Appl. Math., in press. Available at UCLA CAM Report 2001-12 at: www.math.ucla.edu/˜imagers, 2001

Parallelism of Image Inpainting Technology Based on BSCB Model*

Shumin Guo[1] Meiqing Wang[2]

College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian,350002, China

Email:1 yuluziyue@163.com; 2 mq_wang@163.com

Abstract

Image inpainting based on Partial Differential Equation (PDE) requires a lot of computation which makes it hard to be applied to industry. In this paper, to realize high image processing speed and flexibility, the development of parallel algorithms for image inpainting based on BSCB model under Message Passing Interface (MPI) environment is presented. Three methods to partition an image inpainting task are proposed which successfully avoid much communication among subtasks. Two images are tested and the experimental results on a cluster of processors show good speedup.

Keywords：image inpainting; parallelism; PDE; BSCB

1 Introduction

Image inpainting based on PDE has been the state-of-art technology in recent years due to the simplicity and efficiency of the PDE models. Some common used models include of the BSCB model [1], TV model [2], and CDD model [3]. The main idea of these models is to diffuse the information of the regions surrounding the areas to be inpainted into the areas to be inpainted. This diffusion of information is completed by using numerical integration along the temporal axis which is time-consuming.The bigger the region to be inpainted, the longer the running time. The high computing complexity makes the image inpainting methods based on PDE difficult to be applied to industry.

In this paper the parallelism of image inpainting technology based on BSCB model is investigated by using the Message Passing Interface (MPI) [4] which is a commonly used message passing library for parallel environments. Numerical experiments are performed in distributed parallel environments and good speed-up ratios are observed.

The paper is organized as follows. First, the BSCB model and its numerical format are introduced; second the possible parallel methods of the BSCB model are investigated; then the distributed environments used in this paper are introduced and the numerical experiments are presented; finally conclusions are given.

2 Image inpainting based on BSCB model

2.1 BSCB image inpainting model

BSCB model uses Laplacian operator to measure the information of the neighborhood of the region to be inpainted and smoothly propagate the information to the region to be inpainted along the isophote direction. At the same time, to avoid the prolongation lines from crossing each other, anisotropic diffusion function is used. That is, the model contains two steps: inpainting and diffusion.

Let Ω be the region to be inpainted , $\partial\Omega$ be the boundary of Ω. The relationship of Ω and $\partial\Omega$ is depicted as Fig.1. And $I(i,j):[0,M]\times[0,N]\to\mathbf{R}$ be a discrete 2D gray level image where $\mathbf{R}$ stands for the real space.

* This work was partially supported by NSFC under Grant No.10771036 and partially supported by Program for New Century Excellent Talents in Fujian Province, China, under Grant No. 003393.

The BSCB model can be described as follows:

$$\frac{\partial I}{\partial t} = \nabla L \cdot \vec{T} \quad (1)$$

$$\frac{\partial I}{\partial t} = g_\varepsilon k |\nabla I| \quad (2)$$

Eq. (1) is used for inpainting where L is some kind of information and $\vec{T}$ is the isophote direction[7]; Eq. (2) is used for diffusion where k is the Euclidean curvature of the isophote of I, Ω^ε is a dilation of Ω with a ball of radius ε and g_ε is a smooth function in Ω^ε.

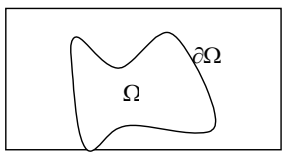


Figure1 An illustration of damaged area and its boundary

In a simple case, the information is substituted by the Laplacian on I. The Eq.(1) and Eq.(2) can be discretized as follows[8,9]:

$$I^{n+1}(i,j) = I^n(i,j) + \Delta t I_t^n(i,j), \forall (i,j) \in \Omega \quad (3)$$

$$I^{n+1}(i,j) = I^n(i,j) + \Delta t \bullet g_\varepsilon(x,y) k(x,y,n) | \nabla I(x,y,n) |, \forall (i,j) \in \Omega^\varepsilon \quad (4)$$

where $I^n(i,j)$ or $I(i,j,n)$ is the intensity of the pixel located at (i,j) in the n-th iteration image and $I_t^n(i,j) = \nabla L \cdot \vec{T}$ in the n-th image. It is noted that $I(i,j,0) = I_0(i,j)$ and $\lim_{n\to\infty} I(i,j,n) = I_r(i,j)$ where $I_0(i,j)$ is the input image and $I_r(i,j)$ is the output of the algorithm. Δt is the rate of improvement.

2.2 Sequential algorithm of BSCB image inpainting model

Sequential implement of image inpainting is as follows:

(1) Read the image file and store the image data in an array.

(2) Search the array, and store the position of the pixels to be inpainted in another array named mask.

(3) Begin the iteration loop, restore the image using the inpainting function and diffusion function.

(4) Write the output into the result file.

3 Parallel implement of image inpainting algorithm using MPI

Digital image processing can be exploited for parallel processing because of their several characteristics. A remarkable characteristic is that a same operation is processed from pixels to pixels or regions to regions[5-6]. Base on this characteristic, the main problem for parallelism is the way to partition the task.

3.1 Partition of the task

At present, partitioning an image into small pieces according to the number of nodes averagely is a common used method in the parallization of an image processing task[10]. That is, if there are N computing nodes that can be used in a distributed parallel environment, the initial image is divided into N smaller non-overlap subimages, each of which processed on one node.

For image inpainting problem, the position of the region to be inpainted is random. To divide the image directly into smaller pieces averagely may lead to the following problems:

Firstly, if a small piece of the image does not contain missing area, the corresponding node will do nothing, which leads to a waste of resource.

Secondly, a certain region to be inpainted may be partitioned into different subtasks. During the processing, these subtasks have to exchange information of the neighboring data, which will lead to the increase of time for communication and decline of efficiency.

Due to these reasons, some other methods are proposed in this section to partition the task.

(1) Partition based on three colour channels

A colour image with RGB model can be easily partitioned into three colour channels. Each channel can be considered as a grey scale image. Accordingly, an image processing task can be decomposed into three subtasks each processing one grey scale image.

(2) Partition based on region

For a grey scale image which has $k > 1$ continuous missing areas to be inpainted, it can be partitioned into k regions, each region contains one continuous missing area.

This method may decrease communication among processes.

(3) Region－channel partition

For a colour image with more than one continuous missing areas to be inpainted, the two partitioning methods described above may be combined together. Let variable *nregion* be the number of continuous missing areas to be inpainted, then the whole inpainting task can be divided into 3**nregion* subtasks: the colour image is first partitioned into *nregion* regions and each region is partitioned into three colour channels.

4 Implement of parallel image inpainting algorithm

4.1 Construction of MPI parallel environ ment

MPI is a common used message passing library for parallel environment which specify a collection of routions which facilitate communication among processors. MPICH2-1.0.5 is chosen to construct the parallel environment in this paper.

There are two models for MPI parallel program design: master-slave model and peer-to–peer model. In this paper, the master-slave model is used.

When using the master-slave model, MPI parallel program contains the following parts：

(1) initialization：at the beginning of the program, the following functions are used to initialize each process.

MPI_Init(&argc, &argv);

MPI_Comm_rank(MPI_COMM_WORLD, &rank);//rank:the index of process

MPI_Comm_size(MPI_COMM_WORLD, &size);//size:the number f processes.

(2) Body of the program：Computation and communication

(3) End the program：the following function is used.

MPI_Finalize();

In MPI parallel program design of master-slave model, two parts are contained in the program body: body of the master process and body of the salve process.

The master process mainly accomplish the followi ng work:

(1) Read the image data from the BMP file;

(2) Partition the task;

(3)Dynamically assign the subtasks to subprocesses and wait for the output of each subprocess;

(4) Check whether all the subtasks have been done, if not, repeat Step3

(5) Combine all the output received from the subtasks reconstruct the new image.

(6) Processes finalize

The sub processes mainly accomplish the following work:

(1) Wait for the subtasks;

(2) Work on its own task to do the image inpainting process and send the result back to the master process;

(3) Repeat step (1) and (2) until no more subtask is assigned;

(4) Processes finalize.

4.2 Experimental results

In this paper three partition methods are tested. There are two 24-bit color images with size of 256×256 for the experiments, depicted in the Figure.2(a) and Figure.3(a). Figure.2(b) and Figure.3(b) depict their damaged versions with the red regions to be the regions to be inpainted containing 3542 pixels and the white regions 3817 pixels respectively.

These tests are performed on a a distributed memory system with 6 processors each of 3.0 GHz with 1GByte RAM. The operating system is Windows XP. The programs are written with C language and MPI Library.

Test 1: Partition based on three colour channels for Fig. 2(b);

Test 2: Partition based on region for Fig.3(b);

Test 3: Region－channel partition for Fig.2(b).

The result image from Test 1 and Test 3 are the same, depicted as in Fig. 2(c). Fig.3(c) is the result image from Test 2.

In each test, since the master process doesn't do the computation task, two processes are started on it, one is the master process and the other is sub process .On other nodes ,only one process is started.

(a) Original Image (b) Damaged Image (c) Inpainted Image

Figure2 Image tested in experiment 1 and experiment 3

(a) Original Image (b) Damaged Image (c) Inpainted Image

Figure3 Image tested in experiment 2

Table 1 performance of Test 1

nodes	Running time（ms）	Speedup	Parallel efficiency
1	59694		
2	32152	1.86	0.93
3	20386	2.93	0.976

Table 2 performance of Test 2

nodes	Running time（ms）	Speedup	Parallel efficiency
1	23750		
2	13312	1.78	0.89
4	6406	3.71	0.927

Table 3 performance of Test 3

nodes	Running time（ms）	Speedup	Parallel efficiency
1	58716		
2	29880	1.96	0.980
3	20200	2.91	0.970
4	16007	3.67	0.918
5	15558	3.77	0.745
6	9578	5.24	0.873

The speedup $S(n)$ and the parallel efficiency $E(n)$ defined as follows are compared in Table 1 to Table 3.

$$\text{Speedup: } S(n)=\frac{T(1)}{T(n)} \qquad (5)$$

$$\text{Parallel Efficiency: } E(n)=\frac{S(n)}{n} \qquad (6)$$

where n stands for the number of computing nodes, $T(1)$ stands for the running time of the image inpainting task on 1 processors and $T(n)$ stands for the running time of the image inpainting task on n processors.

5 Conclusion

In this paper parallelism of image inpainting technology based on BSCB model is studied. Partition based on three colour channels, partition based on region and the combination of two partition methods are proposed and the experimental results based on MPI library show the high efficiency of parallelism. It should be noted that the partition methods used in this paper do not need communication between computing nodes. In future work, some complex partition may be surveyed which may lead a high quality of inpainting images.

References

[1] M Bertalmio, G.Sapiro, V.Caselles and C.Ballester. Image inpainting. Computer Graphics Processings, Annual Conference Series, ACM SIGGRAPH, New Orleans,LA,2000

[2] T F Chan,J Shen,Mathematical models for local nontexture inpaintings[J].SIAM Journal on Applied Math,2001,62 (3):1019-1043

[3] T F Chan,J Shen.Non-texture inpainting by curvature-driven diffusions(CDD)[J].Journal of Visual Communication and Image Representation, 2001, 4(12):436-449

[4] Zhihui Du .Parallel programming of High Performance computation——MPI Parallel programming. Tsinghua publishing house. (in Chinese),2001

[5] K. Hwang, Ed., Special Issue on Computer Architectures for Image. Processing, *IEEE Cornput.,* vol. 16, Jan. 1983

[6] Lee, S.-Y.; Aggarwal, J.K, A system design/scheduling strategy for parallel image processing. Pattern Analysis and Machine Intelligence. IEEE Transactions on Volume 12,Issue 2.Feb. 1990

[7] C. Kenney and J. Langan. *A new image processing primitive*:reconstructing images from modified flow fields. University of California Santa Barbara Preprint, 1999

[8] P. Perona and J. Malik Scale-space and edge detection using anisotropic diffusion. IEEE-PAMI 12, pp. 629-639, 1990

[9] L. Alvarez, P.L. Lions, J.M.Morel. Image selective smoothing and edge detection by nonlinear diffusion. SIAMJ.Numer. Anal. 29, pp. 845-866, 1992

[10] Jie Lv,Tianxu zhang,Biyin zhang.Applications of MPI parallel-computing on image processing Infrared and Laser Engineering2004

The Application of Numerical Interpolation in PDE Image Inpainting*

Chao Zeng[1] Chensi Huang[2] Meiqing Wang[3]

College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian, 350108, China

Email:1 zengchao999_000@163.com; 2 hcs7997@126.com; 3 mqwang@fzu.edu.cn

Abstract

Classical PDE methods cannot inpaint "big" missing region. In order to overcome this problem, a new technology combining numerical interpolation with PDE method is proposed in this paper. First, an interpolation method is used to fix the curve(s) to connect the edge which is interdicted by "big" inpainted region block(s), thus the block(s) can be divided into smaller parts; then, a PDE equation is used to fill information into the inpainted region(s). The experimental results show the new method can achieve a good inpainting quality in restoring "big" missing regions.

Keywords：Image Inpainting'; PDE model; Big Missing Region, Cubic Splines; Local Multivariate Interpolation Method

1 Introduction

In recent years, image inpainting technology has been rapidly developed focusing on two fields: nontexture-based image inpainting and texture-based image inpainting. In the field of nontexture-based image inpainting technology, researchers mainly adopt high order PDE-based inpainting algorithms. Following the methodology of restoring old artworks used by artists, M. Bertalmio et al proposed a PDE model which inpaints along the directions of the isophotes[1]. Guided by the connectivity principle of human visual perception, Tony F. Chan and Jianghong Shen introduced a nonlinear third-order PDE inpainting model based upon curvature-driven diffusions for nontexture images[2]. Through modifying the conductivity coefficient in TV model[3], the diffusion gets stronger where the isophotes are having larger curvatures, this modification overcomes the TV's failure in completing a whole object for a large aspect ratio. Barcelos and Batista proposed an image restoration method with the functions of digital inpainting and noise removal, this method uses different equations in the inpainting domain and the noisy domain to process the degraded images[4]. M. Bertalmio[5] gave a third-order optimal PDE model combining the pixel value interpolation formula selected in [6] and a Taylor expansion, this model ensure continuation of level lines. A variational formula was introduced by Masnou and Morel[7,8] to fill-in the missing information in 2D grey images.

Along with the development of image processing technology, numerical interpolation method is gradually applied to image processing. A B-Spline interpolation-based adaptive denoising method[9] was developed to improve the effectiveness of denoising with high noise levels. Numerical interpolation is more adopted on the field of image interpolation, so as to obtain smooth, clear and high quality images after being zoomed. By extending the normal based subdivision scheme for curve and surface design, a new algorithm for image interpolation[10] was proposed, this algorithm makes the interpolated images have clear edge, natural transition and relatively ideal visual effect. Liu Gang et al put forward a new $2*$ image interpolation based on gradient[11] to apply to image interpolation amplification.

* This work was partially supported by NSFC under Grant No.10771036 and partially supported by Program for New Century Excellent Talents in Fujian Province, China, under Grant No. 003393.

This method has the advantages of fast operation and keeping large information of the image edge. Based on the idea of partition digital images into homogeneous and edge areas based on the analysis of the local structure on the images, a fast edge-oriented algorithm for image interpolation[12] was introduced for real-time enlargement of video images which greatly improve the quality of the enlarged images. At present, adaptive interpolation technology for images is mostly used to keep the definition of the enlarged images, whereas, less applied to image inpainting.

Aiming at nontexture images, in this paper, cubic spline interpolation and the local multivariate interpolation method is used to fix the curve(s) to connect the edge which is interdicted by large damaged region, then the PDE inpainting equation proposed in [4] is used to fill information into the divided missing regions. The results of experiments show the effectiveness of this method in restoring large damaged regions.

2 The bscb model

Let $I(i,j)$ be a digital image, any PDE inpainting algorithm can be written in the following numerical form:

$$I^{n+1}(i,j) = I^n(i,j) + \Delta t I_t^n(i,j), \forall (i,j) \in \Omega \quad (1)$$

here the superindex n denotes the iteration number, Δt is the step of iteration, $I_t^n(i,j)$ stands for the update of the image $I^n(i,j)$ after each iteration, $I^{n+1}(i,j)$ is the inpainted version of $I^n(i,j)$. Ω represents the region to be inpainted, so this evolution equation is only active inside Ω.

Bertalmio et al proposed the first PDE digital image inpainting model - the BSCB model, its main idea is to propagate the surrounding information $L^n(i,j)$ of Ω into Ω along the direction of isophotes, simultaneously, the area of Ω will be shrunk gradually after certain times of iterations, ultimately, the region Ω can be entirely repaired by the surrounding information of 、 Ω.

The diffusion formula of BSCB model is

$$I_t^n(i,j) = \delta L^n(i,j) \cdot \vec{N}^n(i,j) \quad (2)$$

where $\delta L^n(i,j)$ is a measure of the change in the information $L^n(i,j)$, L defines as the Laplacian operator ∇I. $\vec{N}^n(i,j)$ is gained by the 90-degree rotation of the gradient, $\left|\nabla I^n(i,j)\right|$ is defined by the slope-limited version of the norm of the gradient of the image[1]. In order to make the image more smooth while keeping the features of the edge of the image, after several times of propagation, a few iterations of anisotropic diffusion are used to avoid the cross of the isophotes. In [1], the following diffusion equation is used:

$$\frac{\partial I}{\partial t} = g_\varepsilon k \,|\nabla I| \quad (3)$$

where k is the curvature of isophotes, g_ε is a smooth function in Ω^ε, Ω^ε is a dilated area of Ω with a ball of radius ε. The definition of function g_ε is

$$g_\varepsilon(i,j) = \begin{cases} 0 & (i,j) \in \partial\Omega^\varepsilon \\ 1 & (i,j) \in \Omega \end{cases}$$

where $\partial\Omega^\varepsilon$ is the boundary of Ω^ε.

3 The image inpainting methodcom bines with interpolation technology

Generally, PDE models can inpaint thin damaged regions (for example, the scratchings in the left image of Fig. 1) in images ideally. But the results of inpainting solely by PDE models is not satisfying when the damaged regions are large (the white area interdicting the range in the right image of Figure 1).

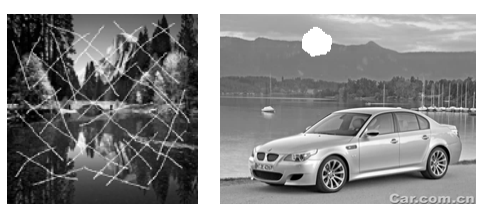


Figure1 Images with thin(left) and large(right) damaged region(s)

Digital image inpainting technology is to fill-in the damaged regions by their surrounding information of pixels, so as to determine the values of the pixels in the damaged regions. The value of each pixel $u(x,y)$ in

digital image u is the intensity function defined at the coordinate (x,y). Therefore, the problem of image inpainting can be dealt with by combining numerical interpolation.

The proposed method uses an interpolation method to fix the curve(s) to connect the edge which is interdicted by large damaged region block(s), thus the block(s) can be divided into smaller parts by the segmentation line(s) which are exactly the curve(s), then, a PDE inpainting equation is used to fill information into the inpainted regions. The steps of producing the segmentation line(s) will be illustrated in detail in section 3.3.

3.1 Cubic spline interpolation

Cubic spline function consists of polynomial pieces on subintervals joined together with certain continuity conditions. Suppose the $n+1$ points $t_0,t_1,\mathrm{L}\ ,t_n$ have been specified and satisfy $t_0<t_1<\mathrm{L}\ <t_n$. These points are called nodes. A cubic spline function of degree 3 having nodes $t_0,t_1,\mathrm{L}\ ,t_n$ is a function S such that:

1. On each interval $[t_{i-1},t_i)$, S is a polynomial of degree ≤ 3.

2. S has a continuous 2nd derivative on $[t_0,t_n]$.

Hence, S is a piecewise polynomial of degree at most 3 having continuous derivatives of all orders up to 2.

3.2 Local multivariate interpolation method

A local multivariate interpolation method of Franke and Little is designed so that the datum at one node will have a very small influence on the interpolating function at points far from that node. Given nodes $(x_i,y_i),1\le i\le n$, we introduce functions

$$g_i(x,y)=(1-r_i^{-1}\sqrt{(x-x_i)^2+(y-y_i)^2})_+^{\mu} \qquad (4)$$

The subscript + indicates that when the quantity inside the parentheses is negative, it is replaced by 0. This will occur if (x,y) is far from the node (x_i,y_i). The parameter μ influences the smoothness of the function.

3.3 Implementation of dividing the damaged area

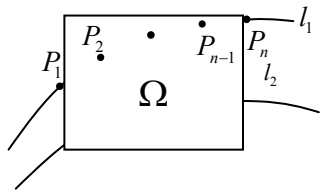


Figure2 The sketch map of selecting interpolation nodes

As shown in Figure 2, Ω is the damaged region within image u, l_1 and l_2 are the edges interdicted by the missing area. The following steps demonstrate the numerical procedure of producing the segmentation lines:

Step 1: Select n nodes $P_1,P_2,\mathrm{L}\ ,P_n$ orderly on the virtual curve which smoothly connects l_1 visually, where P_1, P_n should be the pixels with available values on l_1 as close as possible to Ω. Mark P_i with ordinal numbers, namely $T_i=i,i=1,2,\mathrm{L}\ ,n$.

Step 2: Get the coordinates (X_i,Y_i) of $P_i,1\le i\le n$.

Step 3: Figure out the cubic spline functions S_X and S_Y having the nodes X_i and Y_i respectively. Make the following table with T_i and X_i as number matches:

Table1 Data for Cubic Spline Interpolation

T_i	1	2	L	n
X_i	X_1	X_2	L	X_i

According to the data in Table 1, using the cubic spline function method in [13], we can attain the polynomial pieces $S_{X1},S_{X2},\mathrm{L}\ S_{Xn-1}$ on $[T_1,T_2],[T_2,T_3],\ \mathrm{L}\ ,[T_{n-1},T_n]$, so the cubic spline function can be written as

$$S_X=\begin{cases}S_{X1}(\theta) & \theta\in[T_1,T_2]\\ S_{X2}(\theta) & \theta\in[T_2,T_3]\\ \mathrm{M} & \mathrm{M}\\ S_{Xn-1}(\theta) & \theta\in[T_{n-1},T_n]\end{cases}$$

In the same way, S_Y can also be worked out.

Step 4: Select a series of fitting points in $[T_1,T_2],[T_2,T_3],\mathrm{L}\ ,[T_{n-1},T_n]$ and marked by t, where $t\in(T_{j-1},T_j)(2\le j\le n)$, put t into $S_X(t)$ and $S_Y(t)$ to figure out the abscissas $x=S_X(t)$ and ordinates

$y = S_Y(t)$ of these fitting points, so the coordinates (x, y) of all fitting points are determined.

Step 5: Use the pixels $P_1(X_1, Y_1), P_n(X_n, Y_n)$ with the local multivariate interpolation method to determine the values of the fitting points and the nodes $P_k (2 \le k \le n-1)$. Denote the set consisting of the fitting points and the nodes P_k as L and the coordinates of the points in L as (x_l, y_l). Let $\mu = 1$ in Eq. (4), we can obtain

$$g_1(x_l, y_l) = (1 - r_1^{-1}\sqrt{(x_l - X_1)^2 + (y_l - Y_1)^2})_+ \quad (5)$$
$$g_n(x_l, y_l) = (1 - r_n^{-1}\sqrt{(x_l - X_n)^2 + (y_l - Y_n)^2})_+$$

Then, the value of the pixel (x_l, y_l) in image u can be determined by the following equation:

$$u(x_l, y_l) = u(X_1, Y_1)g_1(x_l, y_l) + u(X_n, Y_n)g_n(x_l, y_l) \quad (6)$$

The five steps shown above can only determine one segmentation line; therefore, repeat this procedure one more time can determine one more segmentation line.

4 The experimental results

In this paper, the BSCB model and the proposed method are used to inpaint the damaged images to compare the restoration results of these two methods (Figure3: An image with the size of 256×192 and Figur4: an image with the size of 256×256); In addition, the restoration results of using the proposed method with the segmentation line(s) at different location(s)(Figure5: an image with the size of 192×192) as well as having different number (Figure6) are also compared. The image processed in Figure6 is identical to the one in Figure3.

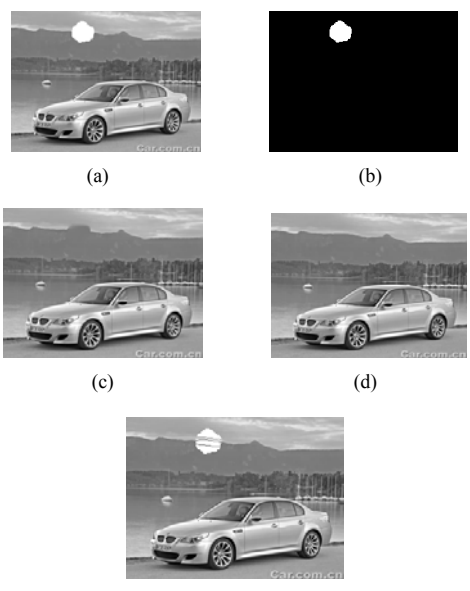


Figure 3 The contrast of the two schemes for the damaged BMW picture: (a) The damaged BMW image; (b) The mask for the damaged region; (c) The result of the BSCB model; (d) The result of the proposed method; (e) The segmentation lines(4 segmentation lines).

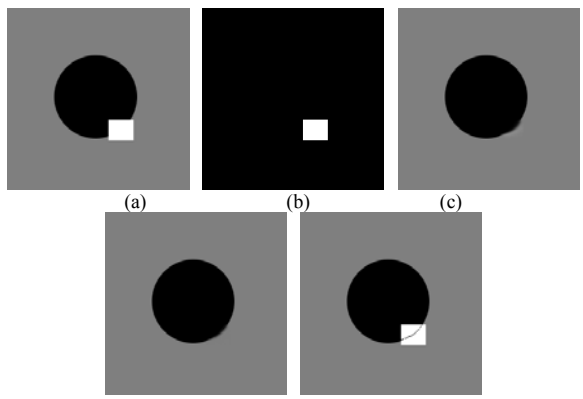


Figure5 (e)

Figure 4 The contrast of the two schemes for the damaged round: (a) The damaged round; (b) The mask for the damaged region; (c) The result of the BSCB model; (d) The result of the proposed method (e); The segmentation line(1 segmentation line).

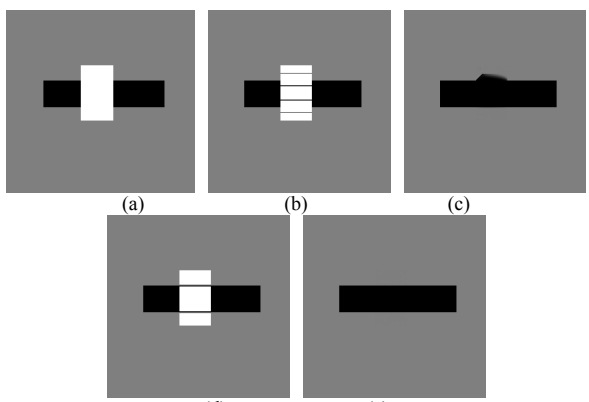


Figure5 The results influenced by the positions of segmentation lines: (a) An interdicted rectangle; (b) 4 segmentation lines at bad positions; (c) Bad result; (d) 4. segmentation lines at good positions; (e) Good result.

As shown in Figure3 and Figure4, the BSCB model can not restore large damaged region, particularly the edge of the image, ideally. In contrast, the proposed method connects the edge of the image with segmentation line(s), at the same time, the damaged region is divided into several parts, then these parts are inpainted by the PDE inpainting

equation proposed in [4], the inpainting result is harmonious.

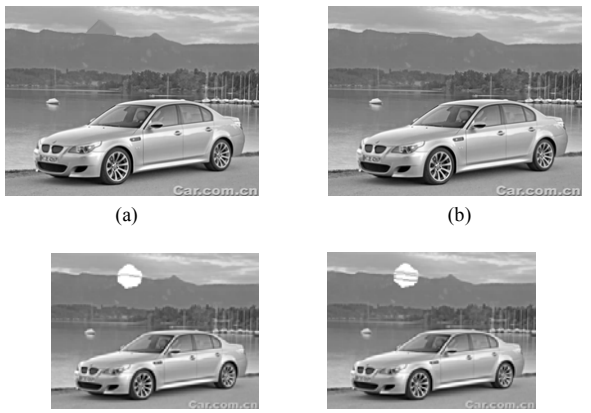


Figure6 The results influenced by the number of segmentation line(s): (a) Processed with only one segmentation line; (b) Processed with four segmentation lines; (c) The position of the sole segmentation line; (d) The positions of the four segmentation lines.

In Figure5, the experimental results show that fix the segmentation lines at the positions where they can connect the edge of the image then use the PDE inpainting equation proposed in [4] to inpaint the divided parts has much better performance than fix them at other positions.

Figure6 illustrates that, to gain good inpainting result, the number of the segmentation line(s) should be determined by the surrounding information of the damaged region. In this experiment, the damaged region interdicts the cloud, the sky and the range. Therefore, several segmentation lines are needed to connect the edges of these objects.

5 Conclusion and future work

In this paper, a new inpainting method is proposed aiming at repairing large damaged areas combining PDE model with interpolation technology. The experimental results show a good inpainting result. But it should be noted that the inpainting effect is impacted by the positions of the nodes selected. Additionally, this method has better performances when restoring relatively distinct edge whereas the performances are not ideal when repairing the edge with many details, therefore, further improvement is necessary for more effective performances.

References

[1] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image Inpainting", *Proc. SIGGRAPH ACM*, pp. 417-424, 2000

[2] Tony F.Chan and Jianghong Shen, "Nontexture Inpainting by Curvature-Driven Diffusions", Journal of Visual Communication and Image Representation 12, pp. 436-449, 2001

[3] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms", *Physica D* **60**, pp. 259–268, 1992

[4] Celia A.Zorzo Barcelos, Marcos Aurelio Batista, "Image restoration using digital inpainting and noise removal", Image and Vision Computing 25, pp.61-69, 2007

[5] Marcelo Bertalmio, "Strong-Continuation, Contrast-Invariant Inpainting With a Third-Order *Optimal* PDE", IEEE *Trans. Image Process*, vol.15, no.7, pp. 1934-1938, Jul. 2006

[6] V. Caselles, J.Morel, and C.Sbert, "An axiomatic approach to image interpolation", IEEE *Trans. Image Process*, vol. 7, no. 3, pp. 376-386, Mar. 1998

[7] Simon Masnou, "Disocclusion: A Variational Approach Using Level Lines", IEEE *Trans. Image Process*, vol. 11, no. 2, pp. 68-76, Feb. 2002

[8] S.Masnou and J. Morel, "Level lines based disocclusion", *Proc. 5th IEEE Int. Conf. Image Processing*, Chicago, IL, pp.259-263, 1998

[9] Liu Qi, Cui Huijuan, Tang Kun, "B-Spline interpolation-based adaptive denoising method", Journal of Tsinghua University(Natural Science Edition), 46(1), pp.42-45, 2006

[10] Luo Liyan, Yang Xunnian, "A Subdivision Approach to Image Interpolation", Journal of Computer-aid Design & Computer Graphic, 18(9), pp.1311-1316, 2006

[11] Liu Gang, Han Jiandong, "A New 2*Image Interpolation Based on Gradient", Infrared Technology, 28(6), pp.324-326, 2006

[12] Mei-Juan Chen, Chin-Hui Huang, Wen-Li Lee, "A fast edge-oriented algorithm for image interpolation", Image and Vision Computing, 23, pp.791-798, 2005

[13] David Ward, Kincaid Cheney, Numerical Analysis Mathematics of Scientific Computing(Third Edition), China Machine Press(Stack of Original Classics), pp. 308-464, 2003.4

Texture Classification Based Digital Watermarking Algorithmin Finite Ridgelet Transform Domain*

Zhibiao Shu

College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian 350002, China

Email: szb@fzu.edu.cn

Abstract

This paper proposes a watermarking technique based on texture classification for image in finite ridgelet transform domain. The texture classification for image is accomplished by using the ridgelet transform. Significant sub-blocks of the images are determined according to the texture classification automatically. Watermarks are inserted in locations which are identified by ridgelet coefficients corresponding to the strong image texture. Experiments show that the proposed method can achieve a better tradeoff between the robustness and invisibility.

Keywords： Wavelet; Ridgelet; Digital Watermark; Texture; Human Visual System

1 Introduction

Watermarking can be classified into two classes depending on the domain of watermark embedding, i.e. the spatial- and the frequency-domain watermarks. It is widely accepted that the frequency-domain watermarking algorithms can easily exploit the perceptual models based on characteristics of the Human Visual System (HVS) to achieve the best tradeoff between imperceptibility and robustness to image processing, and also easy to be implemented in compressed domain. Hence many algorithms have been developed in DCT or wavelet domain [1][2]. Wavelets are in some sense adapted to zero-dimensional singularities, whereas ridgelets [3] are adapted to higher-dimensional singularities; or more precisely, singularities on curves in dimension two, singularities on surfaces in dimension 3, and singularities on n-1 dimensional hypersurfaces in dimension n.

Xiao et al. [4] proposed a perceptually based watermarking algorithm in ridgelet domain is proposed. Based on the principle of multi-channel decomposition of ridgelet transform, HVS properties are exploiting to estimate the JND profile. Then watermark casting scheme which searches the perceptually significant ridgelet coefficients. The watermark sequence is cast into the selected coefficients to provide a higher tolerance to various attacks. Moreover, the fidelity of the protected image can be adjusted by using the weighting factor α of the casting watermark energy.

Guo et al. [5] proposed a digital watermarking algorithm based on image content in contourlet transform domain. The contourlet transform possesses not only spatial and frequency locality and multiresolution but also directionality and anisotropy. Authors make use of the trait of contourlet coefficient, and analyze the embedding location and depth. The watermarks are inserted in locations which are identified by contourlet coefficients corresponding to the strong image texture. And the embedding depth can be adjusted intelligently based on the noise visibility function (NVF). Besides, the detecting algorithm is further optimized.

In this paper, we proposed a digital watermarking algorithm based on image content in ridgelet transform domain. Combining the human visual system with the image local properties, the texture classification for image is accomplished by using the ridgelet transform which is flexibility and adaptability for feature clustering. The watermark embedding locations are identified adaptively according to the result of

classification. Watermark inserting is realized by changing coefficient of the ridgelet transformation.

2 Ridgelet transform

Wavelets in two dimensions are obtained by a tensor-product of one dimensional (1-D) wavelets and they are thus good at isolating the discontinuity across an edge, but will not see the smoothness along the edge. This fact has a direct impact on the performance of wavelets in many applications [6]. While simple, these methods work very effectively, mainly due to the property of the wavelet transform that most image information is contained in a small number of significant coefficients around the locations of singularities or image edges. However, since wavelets fail to represent efficiently singularities along lines or curves, wavelet-based techniques fail to explore the geometrical structure that is typical in smooth edges of images. Therefore, new image processing schemes which are based on true two-dimensional (2-D) transforms are expected to improve the performance over the current wavelet-based methods. To overcome the weakness of wavelets in higher dimensions, Candes and Donoho [7] pioneered a new system of representations named ridgelet which deal effectively with line singularities in 2-D. The idea is to map a line singularity into a point singularity using the Radon transform. Then, the wavelet transform can be used to effectively handle the point singularity in the Radon domain. Their initial proposal was intended for functions defined in the continuous R^2 space.

The continuous ridgelet transform of an integrable bivariate function f(x) is given by

$$CRT_f(a,b,\theta)=\left\langle f,\psi_\gamma\right\rangle=\int_{R^2}\overline{\psi}_{a,b,\theta}(x)\cdot f(x)dx$$

Where ridgelets $x_1\cos\theta+x_2\sin\theta$=const in 2-D are defined from a wavelet type function in 1-D $\psi(x)$ as

$$\psi_{a,b,\theta}(x_1,x_2)=a^{-1/2}\cdot\psi((x_1\cos\theta+x_2\sin\theta-b)/a)$$

In 2-D, points and lines are related through the Radon transform, thus the wavelet and ridgelet transforms are linked through the Radon transform. More precise denote the Radon transform as

$$R_f(\theta,t)=\int_{R^2}f(x)\delta(x_1\cos\theta+x_2\sin\theta-t)dx$$

Then the ridgelet transform is the application of a 1-D wavelet transform to the slices (also referred to as projections) of the Radon transform, and is denoted as

$$CRT_f(a,b,\theta)=\int_{R^2}\psi_{a,b}(t)R_f(\theta,t)dt$$

Instead of taking a 1-D wavelet transform on the Radon transform, the application of a 1-D Fourier transform would result in the 2-D Fourier transform. Let $\hat{f}$ be the 2-D Fourier transform of *f(x)*, and then we have

$$\hat{f}(\lambda\cos\theta,\lambda\sin\theta)=\int R_f(\theta,t)e^{-i\lambda t}dt$$

An invertible discrete ridgelet transform can be obtained by taking the discrete wavelet transform (DWT) on each FRAT projection sequence $(r_k[0],r_k[1],...,r_k[p-1])$ where the direction *k* is fixed. The overall result is called finite ridgelet transform.

3 Texture classification

A texture classification algorithm based on texture and non-texture is proposed in this section. Texture classification involves two phases, i.e., learning and classification. In the learning phase, the original image is decomposed using Finite Ridgelet Transform (FRT) as explained in Section 2.The original image is divided into Z clocks by $n\times n$ (*2n-1* must be a prime number) .Let $C_k(i,j)$ be FRT coefficients in kth clock，where $k\in[1,Z]$，$i\in[1,n]$，$j\in[1,n+1]$ are orientation numbers of sub-blocks. Then there are k×j orientations in single scale FRT domain on image I. Define orientation feature vectors about transform domain as following：

1) Mean m_{kj}

$$m_{kj}=\frac{1}{n}\sum_{i=1}^{n}C_k(i,j)$$

2) Standard deviation sd_{kj}

$$sd_{kj}=\sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}(C_k(i,j)-m_{kj})^2}$$

3) Contrast co_{kj}

$$co_{kj}=\sum_{i=1}^{n}(i-j)^2C_k(i,j)$$

4) Cluster shade cl_{kj}

$$cl_{kj} = \sum_{i=1}^{n}(i - M_{kx} + j - M_{ky})^3 C_k(i,j)$$

$$M_{kx} = \sum_{i=1}^{n}\sum_{j=1}^{n+1} iC_k(i,j)$$

$$M_{ky} = \sum_{j=1}^{n+1}\sum_{i=1}^{n} jC_k(i,j)$$

Above feature elements are quantizated during 0 and 1 , one may get orientation feature vectors $x_{kj}=\{m_{kj}, sd_{kj}, co_{kj}, cl_{kj}\}$ in single scale FRT domain.

Let X be data sample set of all orientation feature vectors. Let the clustering number be 2. For spatial adapt weighted index, iteration cut-off error and maximum number of iteration, the optimal cluster segmentation is realization by FCM algorithm on Matlab6.5.

Class v_1 in FCM algorithm of Matlab6.5 is first class. The orientations on which the sub-block texture is denser than others are concentrated in the class.

Better segmentation effects are achieved after the classified processing. Firstly, perform Finite Ridgelet Transform on images. Secondly, perform k-means algorithm in the sub-blocks of the images according to texture features. Thirdly, judge that sub-blocks of the images are texture or non-texture by using the region segmentation. The result of classification is a satisfaction.

Compared with a typical traditional method, the present approach shows visible improvements both in diminishing segmentation error, and in increasing the precision of boundary and region's harmony.

We call the class v_1 the texture domain and class v_2 the smooth domain. The texture partition algorithm is as following:

Step 1：Set membership grade threshold μ_T, $\mu_T \geq 0.5$;

Step 2：Calculate the number of orientation on which $\mu_{kj} > \mu_T$ for every sub-block;

Step 3：Select the sub-block N_k which satisfies $N_k > 0$ in Z sub-blocks.

The watermark is embedded in the selected sub-block. So one may embed adapt the watermark for different threshold μ_T and the membership grade may be intelligence adjust actor for watermark embedding depth.

From left to right, the gray images of Fig.1 are Lenna, the class v_1 image of size 511×511×256 for ridgelet based texture classification algorithm and the class v_1 image of size 512×512×256 for wavelet based texture classification algorithm. Threshold μ_T is 0.5. Ridgelet transforms are calculated in 7×7 sub-blocks, and there are 2106 sub-blocks which are texture sub-blocks. Wavelet transforms are calculated in 8×8 sub-blocks. And there are 2106 sub-blocks which are texture sub-blocks. The cluster effect of ridgelet transforms is better than that of wavelet transforms.

(a) (b) (c)

Figure1 (a) The image Lenna. (b) Class v_1 for ridgelet based texture classification algorithm. (c) Class v_1 for wavelet based texture classification algorithm.

4 Watermark Embedding Algorithm

The flow chart of watermark embedding algorithm is shown in Fig.2, where V_1 is sub-blocks which are adapt to embed watermark , V_2 is sub-blocks which are not adapt to embed watermark, α is the global embedding strong depth ,β is average energy sum for sub-blocks , γ is membership grade .

Let $ZZ=\{zz_k,k=1,2,...,p,p<Z\}$ be selected set of sub-blocks and $C_k(i, j)$ be ridgelet transform coefficients, where p depends on the threshold value μ_T of membership grade.

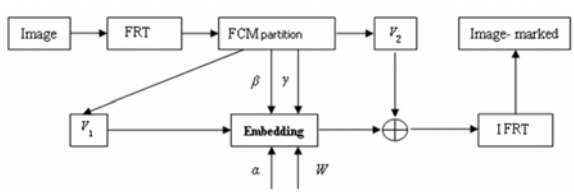


Figure2 Flow chart of watermark embedding algorithm

Denote by N_k $(k=1,2,...,kk,\ kk \le n+1)$ set of orientations which belong to class v_1 in kth sub-block .Compute energy sum E_j of orientations which belong to class v_1 in sub-block :

$$E_j = \sum_{i=1}^{n} C_k(i,j)^2 \quad j \in N_k$$

Let k_h be maximum energy orientation, i.e. $k_h = \max_{N_k}(E_j)$

Denote ridgelet transform coefficients matrix that belongs to p orientations by $CC=[R_1,R_2,...,R_p] = \{r_{ij}\}$,where R_h is ridgelet transform coefficient, $h=1,2,...,p;i=1,2,...,n;j=1,2,...,p$.

Let embedding watermark W be random real number sequence, $W=\{w_1,w_2,...,w_p\}$,which is a Gaussian distribution $N(0,\sigma^2)$.

One bit is embedded on every orientation and embedding algorithm is as following :

$$r_{ij}^W = r_{ij} + \alpha\beta_j\gamma_j \mid r_{ij} \mid w_j$$

Where α is the global embedding strong depth which is given by experiment, β_j is average energy sum for sub-blocks , γ_j is membership grade .Substitute for r_{ij} by corresponding r_{ij}^W and do inverse ridgelet transformation .

5 Watermark Detecting Algorithm

First compute finite ridgelet transformation for detected image with the same single scale as embedding. Then do FCM cluster and texture partition.

Construct ridgelet transform coefficients matrix

$$CC^*=[R_1^*,R_2^*,...,R_p^*] =\{r_{ij}^*\},\ i=1,2,...,n;$$
$$j=1,2,...,p.$$

Compute correlation ρ between detected image and watermark W :

$$\rho = \frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}(r_{ij} - r_{ij}^*)w_j$$

Define the correlation threshold value T_ρ by

$$T_\rho = 3.97\sqrt{\frac{TT}{np}\sum_{j=1}^{p}\sum_{i=1}^{n} r_{ij}^{*2}}$$

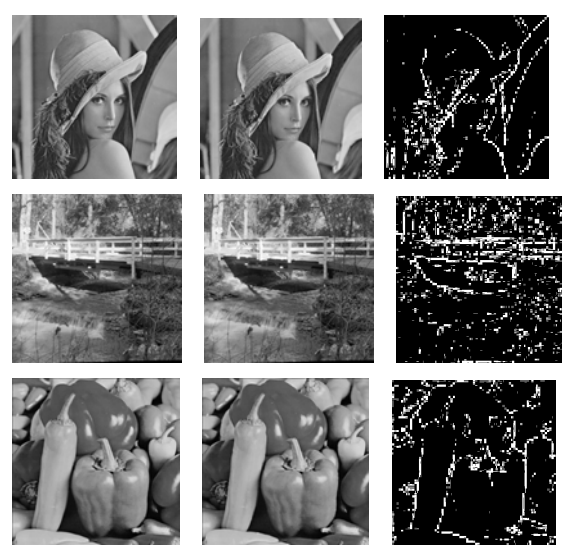
Where TT may be given by experiment and generally TT is 2.

6 Experimental results

We make some experiments by Matlab6.5 to test the algorithm of this paper for some standard grey images. The experimental results are shown in Fig. 3 and Table 1. From top to bottom, the gray images of Fig.3 (a) are Lenna, Bridge and Peppers image. The gray images of Fig.3(b) are those with watermark .The gray images of Fig.3(c) are those which represent site where pixel value are modified. As shown in Fig.3(c), because watermarks are inserted in the strong texture of original image, the visual quality of image achieved from the proposed algorithm is most satisfactory. As shown in Table 1, the stronger the texture is, the bigger the watermark capacity is. According to Fig.4 and 5, we can find that the watermarked image achieved from using wavelet transform method is of inferior visual quality and watermarks are embedded entirely in the edge of original image. As shown in Fig.4, however, because watermarks are inserted in the strong texture of original image, the visual quality of image achieved from the proposed algorithm is improved greatly.

Table 1 Watermark capacity and PSNR

Image	Size	Block number	Capacity	PSNR(dB)
Lenna	511x511x256	5329	2106	43.74
Bridge	511x511x256	5329	4103	42.51
Peppers	511x511x256	5329	1360	44.29



(a) (b) (c)

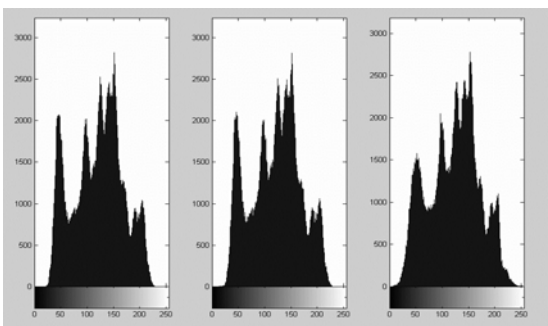
Figure3 (a) The original images. (b) The images with marks (c) The site where pixel value are modified.

(a) (b) (c)

Figure4 (a) The original image. The images after embedding watermarks using (b) wavelet algorithm and (c) ridgelet algorithm.

We generate some anamorphic images attacked by additive 10% salt noise, median filtering, and JPEG compression.

For detecting the existence of watermarks, we randomly generate 500 different watermarks and the 250th watermark is the only true embedding watermark. The detecting results of the watermarked Lenna image are shown in Fig.6, 7 and 8.



(a) Original image. (b) By ridgelet. (c) By wavelet.

Figure5 Histograms of images in Fig.4.

The detecting results for additive 10% salt noise and 3×3 median filtering are shown in Fig.7(a)(b).

In these figures, the response to the correct watermark (i.e. No 250) is much larger than the response to the others and it is higher than the detecting threshold, thus showing the existence of watermark without doubt.

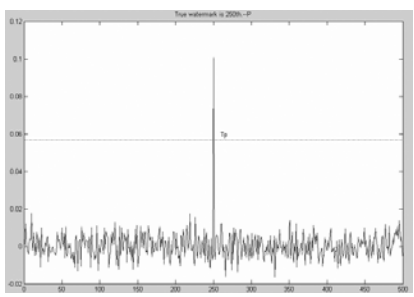
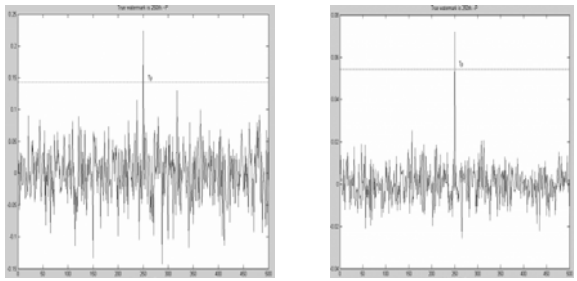
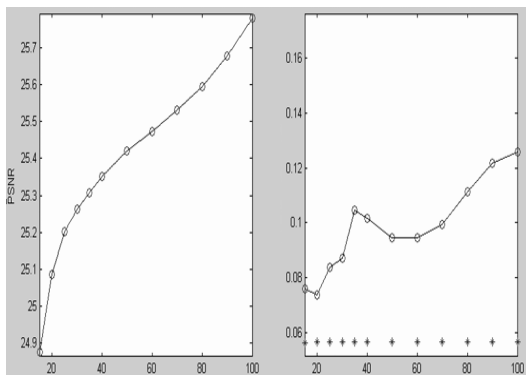


Figure6 Detecting result with no attacking.



(a) Additive 10% salt noise. (b) 3×3 median filtering.

Figure7 Detecting results with salt and median filtering.



(a) PSNR (b) Correlation threshold value.

Figure8 Detecting results by JPEG compression.

7 Conclusions

Based on the principle of multi-channel decomposition of ridgelet transform, we proposed a digital watermarking algorithm based on texture classification in ridgelet transform domain. The watermark embedding locations are adaptively identified based on the ridgelet coefficients relating to the strong texture of original image. The performance of the novel algorithm was very good, and experimental results supported the suitability of ridgelet transform based watermarking scheme for robustly hiding watermarks into images.

References

[1] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, “Secure spread spectrum watermarking for multimedia”,IEEE Trans.Image Processing, Vol.6, Dec. 1997, pp.1673–1687

[2] B.Mauro, P.Alessandro, “Improved wavelet-based watermar king through pixel-wise masking”, IEEE Transactions on image processing, Vol.10, No.5,2001, pp.783- 791

[3] E.J.Candès, D.L. Donoho, "Ridgelets: A key to higher-imensional intermittency?" , Phil. Trans. Royal Society of London A, Vol.357 , 1999, 2495-2509

[4] Xiao Liang, Wei Zhihui,and Wu Huizhong, "Ridgelet-Based Robust and Perceptual Watermarking for Images",Intern ational, Journal of Computer Science and Network Security, Vol.6, No.2B, February 2006, pp.3937 -3942

[5] Guo Lingeng , Shu Zhibiao, "Image Content Based Digital Watermarking Algorithm In Contourlet Transform Domain", DCABES2006 Peoceedings, 2006, 329-332

[6] Minh N. Do and Martin Vetterli, "The Finite Ridgelet Transform for Image Representation", IEEE Transactions on Image Processing, Vol. 12, No. 1, January 2003

[7] E. J. Candes and D. L. Donoho, "Ridgelets: a key to higher-dimensional intermittency", Phil. Trans. R.Soc. Lond. A, 1999, pp. 2495-2509

Object Detection and Localization Based on Image Equalizing and Binaryzation Algorithm

Yi Gao

Criminal Technology Department, China Criminal Police College, Shenyang, Liaoning, China

Email: gao_yi7666@sina.com

Abstract

The key to object detection is separation between the goal and the background, computation of candidate goal characteristic parameter, and making comparison of the standard target characteristic parameter with it, the goal is the one that meets the standard. Using gravity model approach on the centralized positioning based on object detection. This article proposes one kind of method that carries on equalized and the binaryzation target identification under the complex background to the image, it carries on the precise extraction and the localization according to the characteristic parameter. The experimental result indicated that this method could be managed on efficient practice with fast speed, good extraction effect and high recognition rate.

Keywords：Image Equalizing, Binaryzation; Object Detection Localization; Feature Extraction.

The object detection is that separates the goal from the background, this examination algorithm has the direct relations with the image background. The centralized positioning is that determines the physics central place of the goal, it could has direct influence on system examination precision. Because the goal and the background have the similar gradation level, it cannot be separated if the threshold value division is adopted directly, it needs to carry on the histogram equalizing or the gradation to this gradation image stretches, then choose appropriate threshold value to carry on binaryzation processing to the picture target, thus it can achieve the separation of candidate goal and the background. After finishing separation of the goal and the background, it computes candidate goal characteristic quantity, including area, circular, perimeter, etc. Then contrast between these characteristic quantities and characteristic quantity of standard target, the pre-election goal is the one that meets the standard.

1 Image characteristic

The image characteristic includes the area, the perimeter, the circular, etc. The area is the prime number which is contained in the goal object. The perimeter refers to the sum of the goal contour line the picture element distance; the picture element distance has two kinds of situations in the Fig. 1.1, the compound picture element, the compound way may have four directions: the top, bottom, left and right. This kind of compound picture element distance is 1 picture element, like that picture element A and the picture element B's distance is 1, picture element B and the picture element C's distance is also 1. The inclined direction connection's picture element, the inclined direction has four directions: the top left-hand corner, the left bottom, the top right-hand corner, and the right bottom. This kind of inclined direction picture element's distance is $\sqrt{2}$ picture element, as shown in Figure. 1.1. Picture element A and the picture element C's distance is $\sqrt{2}$. When the perimeter survey is carried, the distance is computerized separately according to the picture element connection mode.

	A			
	B	C		

Figure1 .1 Element Distance

The circular is that computing characteristic

quantity of shape complex degree of object (or region), based on the area and the perimeter. For example, the circle and the five-pointed star could be examined. If the area of five-pointed star and circle is equal, then the perimeter of the previous is longer than the one of the latter Therefore, the following parameter could be considered:

$$e = \frac{4\pi \times square}{(perimeter)^2} \quad (1)$$

e is the circular. Regarding radius for the circle, the area equals πr^2, the perimeter equals $2\pi r$, and therefore the circular e is equal to 1. The more its shape approaches to the circle, the bigger e is, and 1 is the most; the more complex the shape is, the smaller e is, and the value of e is between 0 and 1.

2 Gradation histogram equalizing

The histogram equalizing is to transform an image for another that has the equalized histogram image through the gradation transformation, namely the same picture element points process on each gradation level. Suppose the gradation transformation $s = f(r)$ limitedly must to reduce continuously the differentiable function for the slope, the input picture will be transformed into the outgoing picture, the input picture histogram will be $H_A(r)$, outgoing picture's histogram is $H_B(s)$, then their relations may be derived by the following process: according to histogram's meaning, it corresponds small area Yuan equality after undergoing the gradation transformation.

$$H_B(s)ds = H_A(r)dr \quad (2)$$

$$\text{Thus} \quad H_B = \frac{H_A(r)}{ds/dr} = \frac{H_A(r)}{f'(r)} = \frac{H_A(f^{-1}(s))}{f'(f^{-1}(s))} \quad (3)$$

$$\text{And} \quad f' = \frac{df}{dr}, \quad r = f^{-1}(s) \quad (4)$$

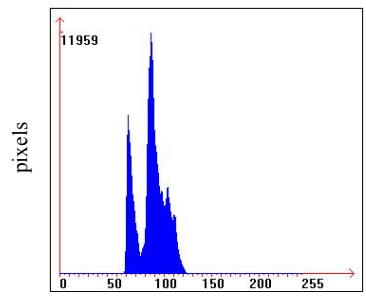
Thus it can be seen, works as in the formula the member and the denominator function only misses a proportionality constant, it is the constant, namely：

$$f'(r) = \frac{R_m}{A_0} H_A(r) \quad (5)$$

$$f(r) = \frac{R_m}{A_0}\int_0^r H_A(r)dr = R_m P(r) \;\; P(r) = \frac{1}{A_0}\int_0^R H_A(r)dr \quad (6)$$

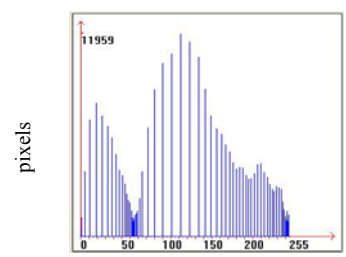
The result of histogram equalizing to the target picture is as shown in Figure2.1

a Primitive Image



b Primitive Image Histogram

c Equalized Image



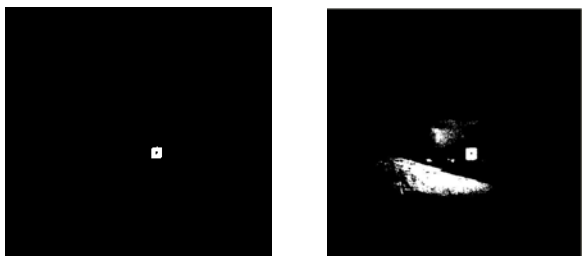
d The Equalizing Image Histogram

Figure2.1 Image Histogram Equalizing Contrast

Through Fig. 2.1, it is can be seen that the equalizing image histogram is evener, the target picture is also clearer; the contrast of the goal and the gradient background is bigger, easier to separate.

3 Separation Between Goal and Background

Choosing the appropriate threshold value to carry on binaryzation processing after the histogram equalizing image, thus we can separate the goal and the background. The so-called image binaryzation is that all picture elements transform the image in the black and white colors in order to extract the goal. Figure. 3.1 is the separation result between the goal and the background.



a. Good Separation Situation b. Bad Separation Situation

Figure3.1 Goal and Background Division Chart

4 Feature extraction

The feature extraction is that computes the candidate goal which the half leaves carries on the characteristic parameter, namely goal area, perimeter, circular parameters and other parameters. We will compare the candidate goal's characteristic parameter with standard target's characteristic parameter, thus we may withdraw the pre-election goal. It can be seen from Figure 4.1 that we may withdraw the goal directly in the good separation situation. This article will illustrate the situation of bad separation situation. According to Figure4.1:

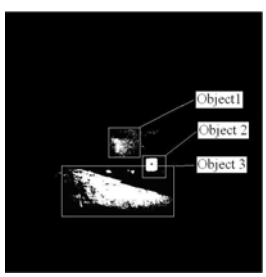


Figure4.1 The Goal Which Separates From the Background

Based on the following table, we may explain how the pre-election goal is separated from the candidate image.

Table 4-1 Feature Contrast Table

	Area	Circular	Perimeter
standard target	150-180	0.650-0.785	50-60
goal1	151	0.410	68
goal2	160	0.752	51
goal3	813	0.426	155

According to the previous table, firstly it can be seen that we may reject the goal 3 through the area feature, secondly we reject the goal 1 through the circular feature, only left the goal 2, and at last we carry on the confirmation to the goal 2 with the perimeter. If it the standard characteristic parameter satisfies what we requested, instead, otherwise this picture has no goal, test result as shown in Figure 4.2.

Figure 4.2 Test Result

We can see that the result of goal test is correct through Fig. 4.2. After examining the goal, we need to carry on physics centralized positioning to the goal.

5 Centralized positioning method

We carry on the centralized positioning based on object detection. The localization algorithm uses the center of gravity algorithm; the two gravity model approach formulas are as followed:

$$X_C = \frac{\sum_{y=1}^{N}\sum_{x=1}^{M} xf(x,y)}{\sum_{y=1}^{N}\sum_{x=1}^{M} f(x,y)} \quad (7)$$

$$Y_C = \frac{\sum_{y=1}^{N}\sum_{x=1}^{M} yf(x,y)}{\sum_{y=1}^{N}\sum_{x=1}^{M} f(x,y)} \qquad (8)$$

In the formula, $f(x,y)$ is the picture element grey level that the image in the point (x,y), M, N respectively is position of examining frame and low-high direction pixels in Fig. 4.2. Because the computation center of gravity's process is the statistical average process, what it figures out the tracking point is not the most luminescent spot individually, but element gradation weighted average position in the image, therefore taking the center of gravity as the tracking point, the track random error is small, and both the anti-jamming ability and the stability are good.

In Figure 4.3, we calculate the center of gravity algorithm result by using the center of target localization. The data in Table 4-2 is tentative data which carry on the batch processing to a series of picture target according to this method of the chapter.

Figure 4.3 Centralized Positioning Result

Table 4-2 Localization of Target Record

	Target localization test result	
frame number	X coordinate	Y coordinate
100	289	299
101	289	299
102	289	298
103	290	298
104	290	297
105	290	298
106	289	297
107	290	298
108	290	298
109	290	298
…	…	…

6 Conclusion

According to tests, we could draw a conclusion that the object detection plan which this article proposed may realize the automatic localization of target, moreover the examination precision is very high and it may satisfy the physical demand.

(1) Successfully separate the goal and background. Because the goal and the background have the similar gradation level, we cannot separate them if we use the threshold value division directly, thus we need to enter the histogram equalizing or the gradation to this gradation image stretches, then choose the appropriate threshold value, carrying on binaryzation processing to the picture target, thus this can achieve the separation of candidate goal and the background.

(2) Successfully withdraw the pre-election goal from the candidate goal. After achieving the separation of the goal and the background, we compute candidate goal feature quantity such as area, circular, perimeter. Then compare these feature quantities with standard target feature quantity, the pre-election goal is the one that meets the standard.

(3) Successfully carry on the center of target localization. After withdrawing the goal, we carry on the centralized positioning with the center of gravity algorithm. The center of gravity computational process is the statistical average process, what it figures out the tracking point is not the most luminescent spot individually, but element gradation weighted average position in the image, therefore taking the center of gravity as the tracking point, the track random error is small, both the anti-jamming ability and the stability are good.

References

[1] A.C. Bovik, T.S. Hvang, D.C. Munson, “The Effect of Median Filtering on Edge Estimation and Detection”, IEEE Transactions On Pattern Analysis and Machine Intelligence, Intelligence, Vol. Pami-9, No. 2, pp. 181~194, March1987.

[2] A.M. Wallace, “An Informaed Strategy for Matching Models Toimages of Fabricated Objects”, Pattern Recognition, Vol. 20, No. 3, 1987.

[3] D.H.L. Sbaiz, S. Sussstrunk, M. Vetterli, "Outlier Modaling in Image Matching", IEEE Transactions On Pattern Analysis and Machine Intelligence, Vol. 25, No.3, pp. 301~315, March 2003.

[4] K. Shan, Q.D. Yao et al, "Applications of L64250 in the real time histogram processing and pixel locating", Journa1 of Zhejiang University (Natural Science), vol.31, No.3, pp.321-328, May 1997.

[5] M.J. Carlotto, "Histogram Analysis Using a Scale-space Approach", IEEE Transactions On Pattern Analysis and Machine Intelligence, Intelligence, Vol. Pami-9, No. 1, January 1987.

[6] S. Wang, J.M. Siskind. "Image Segmentation With Ratio Cut", IEEE Transactions On Pattern Analysis and Machine Intelligence, Intelligence, Vol. 25, NO.6, pp. 675~690, June 2003.

[7] W.C. Luo, W.B. Guo, "Comparison and Analysis of Thresholding Methods for Image Segmentation", Morden Computer, vol. 103, pp. 24-35, November 2000.

[8] Y. Altunbasak, R.M. Mersereau, A.J. Patti, "A Fast Parmetric Motion Estimation Algorithm With Illumination and Lens Distortion Correction", IEEE Transactions On Pattern Analysis and Machine Intelligence, Intelligence, Vol. 12, No.4, pp. 395~408, April 2003.

An Incremental Attribute Reduction Algorithm for Decision Information Systems Based on Rough Set

Hongmei Nie[1] Jiaqing Zhou[2]

1 Department of Information Engineering, Zhejiang Normal University, Jinhua, Zhejiang, China

Email: nhm@zjnu.cn

2 Department of Information, Southwest JiaoTong University, Cheng Du, Sichuan, China

Email:jhzjq@zjnu.cn

Abstract

The attribute reduction is the main subject in the research of the knowledge acquisition, which is based on the rough set theory. The paper divides the rule classes of an information system with decision tables into the set of homogenous rules and the set of non-homogenous rules. Then, we obtain a fast method to determine if the relative positive domain of a conditional attribute subset and the relative positive domain of a whole conditional attribute set are equal. Based on this, we propose an algorithm for incremental attribute reduction. This algorithm can achieve a fast attribute reduction for a dynamically changing information system with decision tables. The effectiveness of the proposed algorithm is verified by simulation results.

Keywords：Rough Set, Conditional Class Set, Attribute Reduction, Incremental Algorithm

1 Introduction

The Rough set (RS) theory, which is proposed by the Prof. Pawlak in 1982, is a mathematic tool to handle fuzzy information and un-deterministic knowledge [1]. In recent years, this theory has been applied to many areas, such as the machine learning, data mining, and decision-support systems.

The attribute reduction is one of the main subjects of the rough set, and, thus, a lot of research has been conducted in this area by many scholars [3-5]. However, most of the research is concentrated on the static data and is not suitable for realistic applications, since the object sets are ceaselessly increased in these

[3] D.H.L. Sbaiz, S. Sussstrunk, M. Vetterli, "Outlier Modaling in Image Matching", IEEE Transactions On Pattern Analysis and Machine Intelligence, Vol. 25, No.3, pp. 301~315, March 2003.

[4] K. Shan, Q.D. Yao et al, "Applications of L64250 in the real time histogram processing and pixel locating", Journa1 of Zhejiang University (Natural Science), vol.31, No.3, pp.321-328, May 1997.

[5] M.J. Carlotto, "Histogram Analysis Using a Scale-space Approach", IEEE Transactions On Pattern Analysis and Machine Intelligence, Intelligence, Vol. Pami-9, No. 1, January 1987.

[6] S. Wang, J.M. Siskind. "Image Segmentation With Ratio Cut", IEEE Transactions On Pattern Analysis and Machine Intelligence, Intelligence, Vol. 25, NO.6, pp. 675~690, June 2003.

[7] W.C. Luo, W.B. Guo, "Comparison and Analysis of Thresholding Methods for Image Segmentation", Morden Computer, vol. 103, pp. 24-35, November 2000.

[8] Y. Altunbasak, R.M. Mersereau, A.J. Patti, "A Fast Parmetric Motion Estimation Algorithm With Illumination and Lens Distortion Correction", IEEE Transactions On Pattern Analysis and Machine Intelligence, Intelligence, Vol. 12, No.4, pp. 395~408, April 2003.

An Incremental Attribute Reduction Algorithm for Decision Information Systems Based on Rough Set

Hongmei Nie[1] Jiaqing Zhou[2]

1 Department of Information Engineering, Zhejiang Normal University, Jinhua, Zhejiang, China

Email: nhm@zjnu.cn

2 Department of Information, Southwest JiaoTong University, Cheng Du, Sichuan, China

Email:jhzjq@zjnu.cn

Abstract

The attribute reduction is the main subject in the research of the knowledge acquisition, which is based on the rough set theory. The paper divides the rule classes of an information system with decision tables into the set of homogenous rules and the set of non-homogenous rules. Then, we obtain a fast method to determine if the relative positive domain of a conditional attribute subset and the relative positive domain of a whole conditional attribute set are equal. Based on this, we propose an algorithm for incremental attribute reduction. This algorithm can achieve a fast attribute reduction for a dynamically changing information system with decision tables. The effectiveness of the proposed algorithm is verified by simulation results.

Keywords：Rough Set, Conditional Class Set, Attribute Reduction, Incremental Algorithm

1 Introduction

The Rough set (RS) theory, which is proposed by the Prof. Pawlak in 1982, is a mathematic tool to handle fuzzy information and un-deterministic knowledge [1]. In recent years, this theory has been applied to many areas, such as the machine learning, data mining, and decision-support systems.

The attribute reduction is one of the main subjects of the rough set, and, thus, a lot of research has been conducted in this area by many scholars [3-5]. However, most of the research is concentrated on the static data and is not suitable for realistic applications, since the object sets are ceaselessly increased in these

applications. Thus, the incremental computation for the attribute reduction is the problem that must be solved for the wide applications of the rough set theory. The research on the algorithms for incremental acquisition of the rule has already been carried out recently [6-8]. However, the subject of this kind of research is the regular set and it does not consider the problem of attribute reduction. In [9], it proposed an algorithm for incremental attribute reduction. However, this algorithm can only deal with the attribute reduction for the information systems without decision attribute, and, thus, is not suitable for the information systems with decision tables. Since in realistic applications, most of the data have decision attribute, the research on the incremental attribute reduction for the information systems with decision tables is more important. In [10-12], they discussed the incremental attribute reduction for the information systems with decision tables by using the decision logic. However, they only discussed the incremental method under the particular situations that may happen after a new record is added to the information systems with decision tables, and did not give a whole scheme to solve the problem of incremental attribute reduction for the information systems with decision tables.

In this paper, we first divide all the rules in an information system with decision tables into the set of homogeneous rules and the set of inhomogeneous rules. Then, we obtain a fast method to determine if the relative positive domain of a conditional attribute subset and the relative positive domain of a whole conditional attribute set are equal. Based on this, we propose an algorithm for incremental attribute reduction. This algorithm can achieve a fast attribute reduction for a dynamically changing information system with decision tables. The effectiveness of the proposed algorithm is verified by simulation results.

2 The basic concept of a rough set

In this section, we give a briefly introduction of some basic concepts related to a Rough set.

Definition 1 (a information system with decision tables [2]).An information system with decision tables S = <U,A,V,f>,where U is the set of objects, and sometimes is called the domain in discussion, $A = C \cup D$ is the set of properties, the subset C and D are called the set of conditional properties and the set of decision properties, respectively, D≠Φ,V is the set of the values of the properties, and $f : U \times A \to V$ is a n information function, which designate the value of a attribute for each object x in U.

For the value of the attribute of the object x, i.e., $x \in U$,and $a \in A$,we have $f(x,a) \in V$.For an arbitrary set $B \subseteq A$,where $IND(B)=\{(x,y):f(x,a)=f(y,a), \forall a \in B\}$,$IND(B)$ forms an equivalent relation on U and is called the indiscernibility relation determined by B on U. The partition on U from it is denoted by $U/IND(B)=\{[x]_B:x \in U\}$, where $[x]_B=\{Y:(x,y) \in IND(B)\}$ is the equivalent class of x related to B.

Definition 2 (Conditional classification and decision classification [2]**)**. Given the decision table $S =<U, C \cup D, V, f>$, where C and D are, respectively, the set of conditional attributes and decision attributes of the decision table, and $U/IND(C)$ and $U/IND(D)$ are, respectively, the partition of the domain U on the attribute set of C and D, the conditional classification is defined as $E_i \in U/IND(C)$ (i=1,…,m, m is the number of conditional classifications), and the decision classification is defined as $D_j \in U/IND(D)$ (j=1,…,n, n is the number of decision classifications).

Definition 3 (The homogeneousness conditional classification and inhomogeneous conditional classification) Given the decision table of $S = <U, C \cup D, V, f>$, and C is the set of conditional attributes in the decision table, for any conditional classification of $E_i \in U/IND(C)$, if all the records in it have the same decision value, E_i is homogenous; otherwise, it is inhomogeneous.

Definition 4 (The *P* positive domain of *Q* [2]**)**. Let U be a domain in discussion, P and Q are equivalent clusters on U. The P positive domain of Q is denoted as $Pos_P(Q)$, where $Pos_P(Q) = \bigcup_{X \in U/Q} \underline{P}(X)$.

Definition 5 (The relative reduction [2]**)**. Let U be a domain in discussion, P and Q are equivalent clusters on U. If, for a subset $S \subset P$, $Pos_S(Q) = Pos_P(Q)$ holds, S is called as the Q reduction of P.

3 Judgment for incremental attribute reduction

From the definition 3, we know that all the conditional classes can be partitioned into the following two parts. One is the homogeneous conditional class set and the other is the inhomogeneous conditional class set.

Definition 6 (Homogenous conditional class set and inhomogeneous conditional class set). Given the information system with decision tables as $S = <U, C \cup D, V, f>$, where $D=\{d\}$, P_C is denoted as the set including all the homogeneous conditional classes in S, and N_C is denoted as the set including all the inhomogeneous conditional classes in S. We call the N_C as the inhomogeneous set related to the conditional attribute set of C.

Definition 7 (Conflict conditional class). Given the information system with decision tables as $S = <U, C \cup D, V, f>$, C and D are, respectively, the conditional attribute set and the decision attribute set of the decision table. Let P_C be the homogeneous conditional set for the conditional attribute set C, and N_C be the inhomogeneous set for C. Given an attribute subset of $B \subseteq C$, for any conditional class $E_i \in P_C$, if one of the following condition is satisfied, E_i is called the conflict conditional class in P_C for the attribute set B; otherwise, it is called the non-conflict conditional class.

(1) There is a conditional class $E_j \in P_C(E_j \neq E_i)$, where E_j and E_i have the same value in the conditional attribute subset B and have the different value in the decision attribute set.

(2) There is a conditional class $E_j \in N_C$, where E_j and E_i have the same value in the conditional attribute subset B.

Property 1 (The monotonic property of the conflict conditional class). Given the information system with decision tables as $S = <U, C \cup D, V, f>$, C and D are, respectively, the conditional attribute set and the decision attribute set of the decision table. For any attribute subset $A, B(B \subseteq A \subseteq C)$, assuming that P_C is the homogeneous conditional class set of conditional attribute set C, and $E_i \in P_C$, the following holds. (1) If E_i is the non-conflict conditional class for the attribute set B, E_i is also the non-conflict conditional class for the attribute set A. (2) If E_i is the conflict conditional class for the attribute set A, E_i is also the conflict conditional class for the attribute set B.

Proof: 1) Assuming that E_i is the conflict conditional class for the attribute set A, from the Definition 7, we know that $\exists E_j \in P_C(E_j \neq E_i)$, which makes E_j and E_i have the same value in the conditional attribute subset A ,and have the different values in the decision attribute set, or there is a conditional class $E_j \in N_C$, which makes E_j and E_i have the same value in the conditional attribute subset A.Since $B \subseteq A \subseteq C$, E_j and E_i also have the same value in the conditional attribute subset B and have the different values in the decision attribute set. Thus, E_i is the conflict conditional class for the attribute set B, which is in contradiction with the condition that E_i is also the non-conflict conditional class for the attribute set A.

2) Assuming that E_i is the non-conflict conditional class for the attribute set B, from 1), we know that E_i is also the non-conflict conditional class for the attribute set A. This is in contradiction with the condition that E_i is the conflict conditional class for the attribute set B. End of the proof.

Theorem 1. Given an information system $S = <U, C \cup D, V, f>$ with decision tables, for any attribute subset $B \subseteq C$, the necessary and sufficient condition for $Pos_B(D) = Pos_C(D)$ is that there is no conflict conditional class for B in the set P_C of homogeneous conditional classes.

Proof: (Sufficiency) In the set P_C of homogeneous conditional classes, there is no conflict conditional class for B, i.e., for $\forall E_i \in P_C$, we have $E_i \in P_B$ (P_B is the

set of homogeneous conditional classes for the attribute set B). From Definition 4, we know that $Pos_B(D) = Pos_C(D)$.

(Necessity) If $\exists E_i \in P_C$ and E_i is the conflict conditional class for the attribute set B, we have the following two situations.

1）If there is a conditional class $E_j \in P_C (E_j \neq E_i)$, which makes E_j and E_i have the same value in the conditional attribute subset *B*, and have different values in the decision attribute set, we have $E_i, E_j \subseteq Pos_C(D)$, and $E_i, E_j \not\subset Pos_B(D)$.

2）If there is a conditional class $E_j \in N_C$, which makes E_j and E_i have the same value in the conditional attribute subset *B*, we have $E_i \subseteq Pos_C(D)$,and $E_i \not\subset Pos_B(D)$.

Since these two situations are in conflict with $Pos_B(D) = Pos_C(D)$, there is no conflict conditional class for B in the set P_C of homogeneous conditional classes. End of the proof.

Theorem 2. Given the information system $S = <U, C \cup D, V, f>$, and $D=\{d\}$ with decision tables, the attribute reductio22n of *S* is *red*. After a new record *r* is added to the information system,

1）if $\exists E_i \in N_C$ or $\exists E_i \in P_C$,$r \in E_i$, or r forms a new conditional class E_{new} ,and E_{new} is the non-conflict conditional class in the attribute set *red*, we have $Pos_{red}(D) = Pos_C(D)$.

2）If r forms a new conditional class E_{new} ,and E_{new} is the conflict conditional class in the attribute set *red*,assuming that E_{new} is in conflict with the conditional class E_k in the attribute set *red*,and $(c_1, c_2, ..., c_k)$ is the attributes in which E_{new} and E_k have different attribute values in the attribute set $C - red$,i.e., $\forall y \in E_k \ (c_1(r) \neq c_1(y) \wedge ... \wedge (c_k(r) \neq c_k(y))$, we have $Pos_{red \cup ci}(D) = Pos_C(D) \quad (1 \le i \le k)$

Proof: 1) All the cases of this theorem under the situation are discussed in the following.

(1) $\exists E_i \in N_C$ （$r \in E_i$）. According to the definition of the conflict conditional class, we can easily see that there is still no conflict conditional class for the attribute set *red* in the set P_C of homogeneous conditional classes. Thus, $Pos_{red}(D) = Pos_C(D)$ still holds.

(2) $\exists E_i \in P_C$（$r \in E_i$）. This situation can be divided into the following two cases.

（2.1）$\forall x \in E_i (d(x) = d(r))$. According to the definition of conflict conditional class, we can easily see that there is still no conflict conditional class for the attribute set *red* in the homogeneous conditional class P_C. Thus $Pos_{red}(D) = Pos_C(D)$ still holds.

（2.2）$\forall x \in E_i (d(x) \neq d(r))$. In this case, the conditional class E_i does not belong to a set of homogeneous conditional classes because of the addition of the record *r*. Let $P_C = P_C - \{E_i\}$,and $N_C = N_C \cup \{E_i\}$. It can be easily seen that after the conditional class E_i is removed, there is still no conflict class for the attribute set *red* in P_C. Thus, $Pos_{red}(D) = Pos_C(D)$ still holds.

（3）*r* forms a new conditional class E_{new} ,and E_{new} is the non-conflict conditional class in the attribute set *red*. It can bee seen that there is no conflict conditional class for the attribute set *red* in P_C. Thus, $Pos_{red}(D) = Pos_C(D)$.

2）From the definition of the conflict conditional class, it can be easily seen that the conditional class E_{new} and E_k are no longer in conflict in the attribute set $red \cup \{c_i\}$（i=1,…,k）. From the monotonic property of the conflict conditional class, E_{new} is still not in conflict with the conditional class, which is originally not in conflict with E_{new} in the attribute set *red*, in the attribute set $red \cup \{c_i\}$ （i=1,…,k）. Thus,there is no conflict conditional class for the attribute set $red \cup \{c_i\}$ in P_C., i.e., $Pos_{red \cup ci}(D) = Pos_C(D) \quad (1 \le i \le k)$. End of the proof.

After a new record r is added to the information system with decision tables, the 1) and 2) in the Theorem 2 contains all the possible situations.

Theorem 3. Given an information system $S = <U, C \cup D, V, f>$, $D=\{d\}$ with decision tables, the result of attribute reduction for *S* is *red*, and the set of the added records is *Add* and $|Add| = m$. For $\forall r_j \in Add \ (1 \le j \le m)$, it satisfies 2) of the Theorem 2. Let us denote $b_j = c_1^j \vee c_2^j \vee ... \vee c_k^j$, where

$(c_1^j, c_2^j, ..., c_k^j)$ is the resulting different attribute values of the conditional class, in conflict with the conditional class formed by the added records, in the attribute set *C-red*. Let $F = b_1 \wedge b_2 \wedge ... \wedge b_j \wedge ... \wedge b_m$. Then convert *F* to a disjunctive normal form and select an element, which has the least attribute combinations, in it. The attributes in this element form a set *A* and we have $Pos_{red \cup A}(D) = Pos_C(D)$.

Proof: we use the induction method to prove this theorem.

1) when s=1,let A=$\{c_i^1\}$ $(1 \le i \le k)$. According to the Theorem 2, we know that $Pos_{red \cup A}(D) = Pos_C(D)$.

2) Assuming that when s=m-1,the conclusion holds, let us investigate the situation when s=m.

Let $F = b_1 \wedge b_2 \wedge ... \wedge b_j \wedge ... \wedge b_{m-1}$ .Convert *F* to a disjunctive normal form and select an element with the least attribute combinations in it. Then, all the attribute values in this element form a set *B*. $\{r_m\}$ forms a new conditional class E_{new}, and E_{new} is the conflict conditional class in $red \cup B$. Let us denote the conditional class in conflict with it as *E*, and denote the conditional attributes, which have different attribute values in the attribute set $C - red \cup B$ for *E* and E_{new}, as $(c_1, c_2, ..., c_k)$. From Theorem 2, we have $Pos_{red \cup B \cup \{c_i\}}(D) = Pos_C(D)$. Let us redefine $F = B \vee (c_1 \wedge ... \wedge c_k)$ and convert it to a disjunctive normal form. Then, select an element in it with the least attribute combinations. The attributes in this element form a set *A*. It can bee easily see that $Pos_{red \cup A}(D) = Pos_C(D)$ still holds.

From Theorem 1 and 2, we know that this theorem holds. End of the proof.

Then, we only need to remove the possible redundant attributes in the attribute set $red \cup A$ to obtain the attribute reduction for the new information system with decision tables.

4 Algorithm for Incremental Attribute Reduction

From the theories given in the last section, we can derive the following incremental algorithm for incremental attribute reduction.

Input: the original decision information system $S = (U, C \cup D, V, f)$, its attribute reduction *red*, its homogeneous conditional class set P_C, its inhomogeneous conditional class set N_C, the added record set Add.

Output: The result *Red* of the attribute reduction for the new decision information system $S_1 = (Add \cup U, C \cup D, V_1, f_1)$.

1.Get the first record in the added record set Add, and denote it as the current record *r*. Let $k = 1$, $Add = Add - \{r\}$.

2.If $\exists E_i \in N_C$ $(r \in E_i)$,go to 5.

3.If $\exists E_i \in P_C$ $(r \in E_i)$

3.1 and $\forall x \in E_i (d(x) = d(r))$,go to 5.

3.2 $\forall x \in E_i (d(x) \neq d(r))$, $P_C = P_C - \{E_i\}$,and $N_C = N_C \cup \{E_i\}$,go to 5.

4.Use the record *r* to form a new conditional class E_{new}.

4.1 for *i*=1 to $|P_C|$ do

1) Let E_i be the *i*-th conditional class in the set P_C of conditional classes.

2) If E_{new} and E_i conflicts in the attribute set *red*, let $c_1, c_2, ..., c_j$,be the different attribute values of E_{new} and E_i on the attribute set $C - red$. Let $b_k = c_1 \vee c_2 \vee ... \vee c_j$,k=k+1

4.2 for *i*=1 to $|N_C|$ do

1) Let E_i be the *i*-th conditional class in the set N_C of conditional classes.

2) If E_{new} and E_i conflicts in the set *red*, let $c_1, c_2, ..., c_j$, be the different attribute values of E_{new} and E_i on the attribute set $C - red$. Let $b_k = c_1 \vee c_2 \vee ... \vee c_j$,k=k+1

4.3 $P_C = P_C \cup \{E_{new}\}$.

5.If $Add = \varnothing$, go to 6; otherwise, get the next record in Add, set it as the current record *r*, let $Add = Add - \{r\}$,and go to 2.

6.Let $F = b_1 \wedge b_2 \wedge ... \wedge b_{k-1}$. Concert F to be a set of elements. Then, arbitrarily select an element in the set which has the least number of attributes, and form a set

A which contains all these attributes.

7.Let $Red = red \cup A$.

8.for for i=1 to $|Red|$ do

1) $P = Red$

2) Let c_i be the i-th attribute in Red

3) $P = P - \{c_i\}$

4) Use Theory 1 to determine if $Pos_P(D)$ equals $Pos_C(D)$. If this conditional holds, $Red = P$.

9. Return Red.

5 Experimental Simulations

In order to verify the effectiveness of the proposed method, experimental simulations are carried out on PC (Intel(R)-Pentium(R) 4,2.4GHz, 256M RAM, Win2000 Server, Delphi 6.0 and SQL Server 2000).The following three algorithms for attribute reduction are selected in our simulations. They are (1) the algorithm based on information entropy [3], (2) the algorithm based on discernible matrixes [4], (3) the algorithm based on the feature selection [5].

Table 1 The data sets in our simulations

Data set	The number of conditional attributes	The number of records
Heart_c_ls	13	303
Pima_India	8	738
Crx_bq_ls	15	690
Liver_disorder	6	1260
Abalone	8	4177

The data sets of Heart_c_ls, Pima_India, Crx_bq_ls, Liver_disorder and Abalone in the UCI database are selected as our test objects. The number of the records and the number of conditional attributes in these sets are listed in the Table 1. From these five data sets, we randomly select 80 percent of the records to form the original information system with decision tables, and the rest 20 percent of the records is treated as the new added data set. Firstly, we apply the three attribute reduction algorithms, mentioned in the last paragraph, to the original decision information system and obtain a result for attribute reduction. After that, based on this result, we apply the proposed incremental attribute reduction method to the added data set and obtain an reduction result for the new information system with decision tables. Then, we make a comparison between this result and the result when the three attribute reduction algorithms are applied directly to the whole information system with decision tables. The result of this comparison is listed in the Table 2, where T stands for the execution time of an algorithm in the unit of a second, and n stands for the number of the conditional attributes in the result of attribute reduction.

Table 2 The simulation result for the UCI database

Data set	The reduction based on feature selection				The reduction based on information entropy				The reduction based on discernible matrixes			
	Non-incremental		Incremental		Non-incremental		Incremental		Non-incremental		Incremental	
	T	n	T	n	T	n	T	n	T	n	T	n
Heart_c_ls	5.804	9	0.1	9	0.256	9	0.01	9	0.287	9	0.016	9
Pima_India	37.922	5	0.1	5	0.416	5	0.016	5	0.6	5	0.017	5
Crx_bq_ls	49.287	13	0.109	6	1.132	6	0.031	6	1.984	6	0.018	6
Liver_disorder	71.84	5	0.1	5	0.178	5	0.016	5	1.459	5	0.17	5
Abalone	118.906	7	1.366	6	17.634	6	0.594	6	42.875	6	0.678	6

From Table 2, we can see that the execution time of the incremental method proposed in this paper is much less than that of the non-incremental method. Furthermore, the simulation results demonstrate that the relative reduction of a dynamic changing information system with decision tables can be speedily obtained, when the proposed method is applied for the incremental procession.

6 Conclusions

Attribute reduction is one of the main subjects in the research of knowledge acquisition based on the rough set theory. Although some research have been

carried out for the static decision information system, such as the minimum reduction algorithm, small reduction algorithm, and the satisfaction reduction algorithm, most decision information systems are dynamic. Due to this, incremental knowledge acquisition algorithms have been put forward. However, the incremental processing objects in these algorithms are regular sets and these algorithms do not consider the problem of attribute reduction. In [9], it proposed an incremental attribute reduction algorithm. But it can not deal with the problem of attribute reduction in the information systems decision tables. In [10-12], the incremental reduction methods for the information systems with decision tables are discussed. However, there is no systematic scheme in these papers to solve the problem of the incremental attribute reduction for the information systems with decision tables. This paper has proposed an incremental attribute reduction algorithm based on the set of conditional classes. This algorithm can speedily obtain the attribute reduction for the dynamically changing information systems with decision tables. Simulation results have verified the effectiveness of the proposed algorithm.

References

[1] Z.Pawlak, "Rough Set",International Journal Of Computer and Information Sciences,No.11, November 1982, pp.341-356.

[2] Guoying Wang, Rough set theory and knowledge acquisition, XiAn:Xian Communication University Press, 2001.

[3] Guoying Wang, Hong Yu, Dachun Yu, "Reduction of Decision tables based on conditional information entropy", Computer Journal,Vol.25,No.7, July 2002,pp.759-766.

[4] Liyun Chang, Guoying Wang, Yu Wu, "A method for rule acquisition and attribute reduction in the rought set theory", Software Journal,Vol.10,No.11,November 1999,pp. 1206-1211.

[5] X. H. Hu, N. Shan, N. Cercone. "A rough set based knowledge discovery systems", 8th International Symposium on Methodologies for Intelligent systems, 1994, pp.386-395.

[6] Z. Zheng, G. Y. Wang, Y. Wu, RRIA: "A Rough Set and Rule Tree Based Incremental knowledge Acquisition Algorithm", Fundamenta Informaticae, Vol.59, 2004, pp.299-313.

[7] Dongjun Yu, Shitong Wang, Jinyu Yang, "An algorithm for the incremental acquisition of rules",Mini and Micro Computer Systems,Vol.25,No.1, January 2004,pp.79-81.

[8] Xuelan Li, Congfu Xu, Weidong Geng, "An algorithm for incremental rule acquisition based on improved differential matrixes", Computer Science,Vol.30,No.5(sepcial issue), May 2003, pp.46-49.

[9] Zongtian Liu, "An incremental algorithm for the minimal attribute reduction",Electronic Journal,Vol.27,No.11, November 1999,pp.96-98.

[10] W. C. Bang Z. N. Bien, "New incremental inductive learning algorithm in the framework of rough set theory", International Journal of Fuzzy Systems,Vol.1,No.1, January 1999,pp.25-36.

[11] Yinxiang Li, Deyu Li, Jufu Zhang, "An incremental algorithm for attribute reduction in the rough set theory", Computer Science,Vol.31,No.10(sepcial issue), October 2004, pp.38-40.

[12] Yinhua Li, Jufu Zhang, Sufang Gao, "An algorithm for the incremental attribute reduction based on the rough logic",Journal for System Simulation,Vol.17,No.2, February 2005,pp313-333.

A Review on the Recognition Methods of License Plate

Hao Peng Dan Liu Haojie Yan

Department of Forensic Science and Technology, China Criminal Police University, Shenyang, Liaoning 110035, China

Email: watchingsky@163.com, dliu@ccpc.edu.cn, jiekezl@163.com

Abstract

The procedure of License plate recognition can be divided into plate location, character segmentation and character recognition. It is widely used in Intelligence Traffic System. In this paper, the recognition methods of license plate are listed and most methods of them are introduced.

Keywords： Recognition of License Plate (RLP); Intelligence Traffic System

1 Introduction

License Plate Recognition means Automated Data Input where Data equals the registration number of the vehicle. License Plate Recognition System is an integrated hardware and software device that reads the vehicles license plate and outputs the license plate number in electronic text to some data processing system. The next step is the character segmentation: after the extracted license plate image is normalized, the individual character has to be distinguished (segmented) from each other. Segmentation becomes difficult when the plate is not clear, the characters are touching each other, there are screws or strong light-effects (like shadows) on the plate, etc [1].

When the characters are properly segmented (separated from each other and precisely localized) there is time to invoke the character recognition algorithm for each individual segmented character image. The output of each character recognition process is the ASCII code of the given character image. By recognizing all characters after each other, the entire plate text is read.

The system of realization for vehicle license plate can divide into three main sections. They are plate location, character segmentation and character realization. Generally speaking, the three procedures have strict order. Every procedure is the basis of the next one. Every procedure has several methods to implement it. And the method used currently to implement will influence the method which will be used in next procedure. So the arrangements of the methods make different kind of procedure to recognize the license plate. These procedures all have their advantage and shortcoming. Some may be efficient but not accuracy, some may be accuracy but very complex and with low speed. According to their quality and the environment will be used, we can chose the most suitable one.

In this paper, some recognition methods of license plate and the principles they have used will be introduced.

2 Preprocessing

The images' quality is very important. It influences the result of the plate location. An image with good quality will make the location successful. In fact, the images are often not acceptable. It is often contaminated by high light and information noise. The angle is another problem for the search. So before searching we should increase the images' quality. This step is called preprocessing.

Preprocessing algorithm for License Plate Recognition (LPR) includes image enhancement and noise reduction. Binary conversion is often included in the preprocessing, but some recognition algorithm use

color image to locate the plate. So in this kind of algorithm the binary conversion is unnecessary.

Remove noises is a key step of the preprocessing. Pepper noise is the most common noise. It will be eliminated by using filers. The simplest filter is arithmetic mean filter. Let S_{xy} represents the set of coordinates in a sub rectangular image windows of size m*n, centered at point(x, y). The arithmetic mean filter computing the average value of the corrupted image $g(x,y)$ in the area defined by S_{xy}. The value of the restored image f at any point(x, y) is simply the arithmetic mean computed using the pixels in the region defined by S_{xy}. Eq. (1) In other words,

$$f(x,y)=\frac{1}{mn}\sum_{(s,t)\in S_{xy}} g(s,t) \qquad (1)$$

This operation can be implemented using a convolution mask in which all coefficients have value 1/mn. A mean filter simply smoothes local variations in an image. Noise is reduced as a result of blurring [2].

Filters make the original image clearer and suitable to be processed. If after preprocessing the image achieves the quality we need. We could begin the step to search in the license plate area.

3 Plate location

Plate location is the first step of the whole procedure. This step will seek the area of the license plate in the image. At this step, the LPR software should determine where the license in the image, and take the area which contain the plate out. After the license plate is located and extracted from the whole image, it has to be transformed into a standardized form:

1)normalized foreground/background colors (normal ized contrast and brightness values),

2)normalized size,

3)uniform plate orientation.

How to locate the license plate will be different according to different methods. Almost all the algorithms have the aim that detecting the border of the area which contains the license plate. There are two kinds of common methods to achieve this aim. The first one is based on gray image. The second one is base on color image.

Methods based on gray image

This kind of image can be expressed by a uniform real number function, for example f(x, y). The value of this function is called the gray-scale or the brightness of the pixel. For gray image is easy to be expressed, and the processing is constraint by memory, the methods based on the gray image is abundant and mature. The color image not only have affluent color information, but also contain light information, so they are very complex and to processed them should use large memory and high speed CPU. For this matter, the methods based on color image are very few. But the color image contains more information than gray image, so recently some scientists also begin to search the method based on color image.

If the original image is gray, the method based on gray image should be used. Generally speaking, we should make binary conversion. In the binary image we could search the license more easily.

Projection method is a good method to detect the license. For using the method, we should use suitable threshold to binary the gray image. After this operation the image will only have gray level values 0 and 1. The prior knowledge of plate is useful to plate location. Because in one image, square border is very imply. The computer may be seeking out several square areas maybe contain license. To use the union characteristic of the plate, the areas which not contain the license can be ruled out.

There are many characteristics of plate. Different methods use different characteristics of locating plate.

The changes of gray scales are very fluent and have own feature. Some methods use this characteristic to search the area which contains the plate.

The horizontal and vertical proportion is also very useful to detect the plate, some method use projection to measure the density of an area and decide whether the area contain a plate.

Wavelets and morphological image processing are also used to locate the plate. The methods which use these mathematical technologies are good at processing those images which have lower quality or noise. So these methods often not need to preprocess image [2].

Methods based on color images

In a long time, people spend lots of time to research the methods based on gray images. But the color image contains more information we can use, because of the enhancement processing ability of the computers, the methods based on the color image are paid more attention. Since the plates have unique color, it can be separated from the background easily. This feature can be used to locate the area which contains the plate [3].

4 Character segmentation

If the plate is located successful, the first step of plate recognition is accomplished. The next step is character segmentation. The task of this step is to divide the word of the plate into some single characters. Character segmentation is an important step in License Plate Recognition system. There are many difficulties in this step, such as the influence of image noise, plate frame, rivet, the space mark, and so on.

Generally speaking, the size of license plate is unity. They all have unity length and width. And there is a big interval between the first two characters and the last five characters. This information is used as the prior knowledge. The methods in this step also have some common principles. Preprocessing is one of them.

The preprocessing of character segmentation often contains three parts. The first is size normalization; the second is determination of plate class and the third is object enhancement [7].

Size normalization: The size of the plate images is an important factor for the accuracy of character segmentation. if we want to make the segmentation successfully, the size of the plate that we use should be unity. But we plate we cut out from the image is also not unity, so we often stretch or shrink methods to modify the plate.

Determination of plate kind: There are three kinds of Chinese license plates: black characters on a yellow back ground, white characters on a Blue background and white characters on a black background [7]. Different plates have different features, so we should use different methods to process the plate, such as binary methods and color based methods.

Object enhancement: There are many factors can influence on the quality of Plate images. Illumination variance and noise make the quality of plate image low, so it difficult for character segmentation. Then some image enhancement should be used to improve the quality of images. For character segmentation, the enhancement should only enhance the character pixels and weakened the background pixels at the same time. So this kind of enhancement called object enhancement. The object enhancement algorithm consists of two steps. Firstly, gray level of all pixels is scaled into the range of 0 to 100 and compared with the original range 0 to 255, the character pixels and the back ground pixels are both weakened. Secondly, all pixels are sorted by gray level in descending order and multiply the gray level of the top 20% pixels by 2.55. Then most characters pixels are enhanced while background pixels keep weakened [7].

There are several classic methods to divide characters.

The first one is to detect the gap between two characters. As we all know, the blank area between each character in plate is equal (except the gap between first two characters). And after the preprocessing, the character and background is optically in gray. So it is easy to calculate the distance between two characters. But this method depends on the size of the image. If the image have been stretched, the result of the method should not be ideal.

Projection method is to use the horizontal projection and the vertical projection to divide the plate. The horizontal projection is the sum of the black pixels in each raw. The vertical projection is the sum of the black pixels in each line. The speed of this method is fast, but it can't process non-regular characters [4, 5, 6, and 7].

The projection method will be introduced in detail.

The Hough Transformation can be used to detect lines in an Image. For each pixel (x_0, y_0) in the image space, using transformation,

$$r = x_0 \cos\theta + y_0 \sin\theta \qquad (2)$$

We get a curve $r = x_0 \cos\theta + y_0 \sin\theta$ in the parameter space (r, θ). Suppose that there are n points in the images pace. After translating them to the parameter space, we obtain n curves in the parameter space. If

these curves cross at a same point (r_0, θ), then the n points in the image space are on a line. So we can find lines in the image space by searching the cross points in the parameter space.

For the plate images with large rotation, it is difficult to obtain horizontal segment lines by horizontal projection analysis. However, for a single character, rotation has little effect on its horizontal projection. It is much easier to analyze the horizontal projection of a single character and find the horizontal segment lines. So the horizontal segmentation algorithm can be realized as follows [7]:

Firstly, find the valleys of the vertical projection and then divide the plate image into many blocks vertically. Secondly, find the horizontal segmentation line for each block by analyzing the horizontal projection of the block. The horizontal segmentation line for a single block is called a subsection line. Thirdly, use Hough transformation on the midpoints of all subsection lines to remove the incorrect subsection lines and combine the correct subsection lines into a whole line. Finally, the vertical segmentation is used the same method as horizontal segmentation. The vertical segmentation algorithm consists of four steps:

1) Find candidates for the vertical segmentation lines. A candidate is assigned for each valley of the vertical projection.

2) Estimate the size of the plate and each character by using the position information of the horizontal segmentation lines and the candidates.

3) Estimate the position of the left and right borders of the big interval by using the prior knowledge of character size. The variance of the pixels gray level along a segmentation line should be small, because a segmentation line should be located in the interval of the plate and the pixels it crosses are the background pixels with similar gray level. On the contrary, when it crosses a character, the gray level variance will be much larger. Based on this fact, the vertical segmentation lines (the left and right borders) for the big interval can be deduced by searching around the estimated positions and finding the best segmentation lines with the minimum variance from the candidates.

4) The other vertical segmentation lines can be located in the same way.

There are some new technologies to be used for the character segmentation. Such as motion analysis.

Color image segmentation is very useful in many applications. From the segmentation results, it is possible to identify the regions of interest and the objects in the scene, which is very beneficial to the subsequent image analysis or a notation. Recent work includes a variety of techniques: for example, stochastic model based approaches; morphological water shed based region growing, energy diffusion, and graph partitioning. Quantitative evaluation methods have also been suggested. However, due to the difficulties, there are few automatic algorithms that can work well.

Image segmentation is difficult because of the image texture. If an image only contains homogeneous color regions, clustering methods in color space can resolve the problem [16].

The difference between the color of the license plate background and that of the characters is very evidently. So we can segment the characters use color image segmentation.

5 Character recognition

In order to recognize each character in the plate one by one, we change the image into ASCII code or Unicode, and then store these codes into the database.

The core principle of character recognition technologies is recognizing different character according to the feature they have. Every character has its own shape and features, and the different font also have their own features. So the methods are different according to the font and the main features they have.

Optical character recognition (OCR) has been a topic of interest since possibly the late 1940's when Jacob Rabinow started his work in the field. The earliest OCR machines were primitive mechanical devices with fairly high failure rates. But they quickly gave way to computer-based OCR devices that could outperform them both in terms of speed and reliability [14].

Today there are many OCR devices in use based on

a plenty of different algorithms. All of the popular algorithms have high accuracy and high speed, but still many suffer from a fairly simple flaw, the mistakes are often very unnatural to the human point of view. E.g. mistaking a "5" for an "S" just as some people do and seemed reasonable, but mistaking a "5" for an "M" that seemed counter-intuitive and unexpected. Algorithms make such mistakes because they generally operate on a different set of features than humans for computational reasons. These methods do work and are often computationally efficient, but they make the computer see letters through a decidedly non-human set of eyes [8, 9].

There are varied methods for character recognition, there are three main classes: the methods based on template matching, the methods based on neural net and the methods based on component analysis.

The template matching methods create the vector of character by the feature of character. Then match characters by comparing the template and active character. This method should be useful only if has a method which can precisely depict the characters' feature. Common feature includes pixel gray level value, character border linked list, strokes, yielding point and so on. These features can precisely depict the characters, but they are easy to be influenced by the noise, so have low robust.

Neural net is the most popular techniques using in character recognition. Neural net can't be suitable to the images which the noise changed very frequently. And the result made by this method is firmly depended on the neural net's structure and its convergence. The methods based on neural net are not very stable.

Component analysis is often used with categorizer.

6 Chinese character recognition

There is an important character of Chinese license plate. The Chinese license plates contain Chinese character. So the Chinese character recognition is needed to be taking into account.

The People's Republic of China created the simplified standard for its own use. The traditional character set is still used in Taiwan, Hong Kong, Macau, and in overseas Chinese publications. Also in common use is four main font styles: songti, fangsongti, kaiti, and heiti [15].

The character use in the license plate is a variant of heiti or very similar to it. So it can be recognized by using the heiti reorganization methods.

The characters are used in the plate is limited. So some people put up a method base on projection. This method can recognize Chinese characters effectively by detecting the strokes amount, character structure and long or short horizontal amount in the projection image.

There are many Chinese feature analysis methods, such as Syntactic pattern methods, extraction Chinese characters' skeleton methods, characters' statistics features based methods and shape features based methods. The methods which use stroke features are very common, because this method is easy to be implemented. The number of characters used in the plate is very few. If military vehicle isn't considered, there are only thirty-four characters used in the plate. If observing these characters, it can be easily found that these characters have very different stroke features. Some characters have many strokes, like: 冀, some ones have few strokes, like 京; some ones is right-left structure, like 沪, some ones have long horizontals, like 京,吉 and so on. The characters can be recognized by the statistics of structures based methods and shape features based methods. In this case, a few features of characters can be picked-up, and an analysis tree is built. Then the character in a small character set will be analyzed [10, 11, 12, 13].

Projecting the characters to the x and y axis, analysis the points according to the wave peak and wave valley. We can detect the long strokes in the character by these methods too. These methods have poor anti-disturbance, so before using this methods, we should preprocess the image.

7 Final process

After the three processing steps mentioned above, the image contains the license plate will be changed to the character code that computer can process.

The character code (should be Unicode) can be stored in the database, compare the code stored in the database to achieve other tasks.

8 Summary

In this paper, the procedure of the plate recognition is introduced. The plate recognition can be divided into three main steps. They are plate location, character segmentation and character recognition. Every step is based on the former one. And every step has many methods. Some methods are classical, and some use the advanced technologies. But because the backgrounds of plate are very complex and the feature of plate is various. There is not a common method that can be used to recognize all kinds of plate images successfully. Every method has its focus.

The Chinese vehicle licenses have their own features. The most mark is the Chinese licenses contain Chinese characters. It is harder to recognize the Chinese characters than to recognize the English characters. But because the set of Chinese character is very small and the features of them are different, we can search the Chinese characters in this small set, the efficiency should be mended.

Acknowledgments

The authors would like thank to all the colleagues for their former work in this research, especially, the authors of reference [7].

References

[1] YING Hong-wei, SONG Jia-tao, YANG Zhong-xiu, REN Xiao-bo, Research on Vehicle License Plate Segmentation Algorithm, 2007. (in Chinese)

[2] Gonzalez and Woods, Digital Image Processing, Prentice Hall, 2002.

[3] Zhao Shu guang. Principles, Development and Applications of Programmable Analog Devices M. XiAn Electronic Science and Technology University Press, 2002. (in Chinese)

[4] Feng Yang, Zheng Ma, Mei XieA, Novel Approach for License Plate Character Segmentation, Industrial Electronics and Applications, 2006 1ST IEEE Conference, 2006.

[5] ZHANG Yu, MA Si2liang, HAN Xiao, ZHANG Zhongbo, Image Extraction and Segment Arithmetic of License Plates Recognition, 2006. (in Chinese)

[6] LIU Xing, JIANG Tianfa, Research and application of separating technique of vehicle plate character image, 2006. (in Chinese)

[7] Yungang Zhang, Changshui Zhang, New Algorithm for Character Segmentation of License Plate, 2002.

[8] ZHANG Xie-hua, ZHANG Shen, Research of Algorithm for Recognition of Characters on Vehicle License Plate, 2007. (in Chinese)

[9] Eric W. Brown, Character Recognition by Feature Point Extraction, 2004.

[10] R Romero, R Berger, R Thibadeau, D Touretzky, Neural Network Classifiers for Optical Chinese Character Recognition, 1995.

[11] T Natio, T Tsukada, Yamada, et al, Robust license plate recognition method for passing vehicles under outside environment J.IEEE Transaction on Vehicular Technology, 2000.

[12] PARIST R. Car plate recognition by neural networks and image processing [C]//Proc.of IEEE International Symposium on Circuits and Systems.[S.l.]:IEEE Press, 2000.

[13] WANG Lei, REN Hong, Synthesis narration of Chinese character recognition method in license plate, 2007. (in Chinese)

[14] V. K. Govindan, A. P. Shivaprasad, Character recognition — A review, Pattern Recognition, 1990.

[15] R Romero, R Berger, R Thibadeau, D Touretzky, Neural Network Classifiers for Optical Chinese Character Recognition, Pattern Recognition, 1995.

[16] Y Deng, BS Manjunath, H Shin, Color image segmentation, Computer Vision and Pattern Recognition, 1999.

Low-quality Fingerprint Image Enhancement and Fragmentary Fingerprint Image Reconstruction

Haojie Yan Dan Liu Hao Peng

Department of Forensic Science and Technology, China Criminal Police University, Shenyang, Liaoning 110035, China

Email: jiekezl@163.com, dliu@ccpc.edu.cn, watchingsky@163.com

Abstract

A fingerprint from real criminal cases is very reliable forensic evidence. The most problematic fingerprint is still much more reliable than eye witness testimony. At present, the technology of image processing is widely used to deal with low-quality fingerprints. Using different filters to improve the quality of the fingerprint image makes the information provided by the fingerprint is closed to the real information much than before. Moreover, the technology of fragmentary fingerprint image reconstruction is still not mature currently, needing to do the further research on the issues involved.

Keywords: Low-quality fingerprint image enhancement; Fragmentary fingerprint image

1 Introduction

Generally speaking, image enhancement is carried on by the digital image processing methods, such as image smoothing, filter, converting the image to a binary image, normalization etc. While operated practically, the fingerprint image enhancement is generally realized as follows: normalization, pattern estimation, frequency analysis, template generation, and filter [1]. Only if a fingerprint image is normalized, the mean and square of this image can be controlled in the given range for the later processing. The purpose of the fingerprint image normalization is to reduce the square of the gray image. The methods of fingerprint image enhancement can be divided roughly as follows: the gray level transformation enhancement, the time domain filter enhancement and the frequency domain filter enhancement [3]. Currently, methods of image enhancement that based on pattern estimation filter enhancement and Gradient vector filter enhancement are much more popular. Pattern estimation filter enhancement algorithm is very sensitive to the width of fingerprint ridgelines. The size of the pattern window influences the processing result. If the window is too large, the maximum of the fingerprint central curvature will be mistaken.

2 The generation of fingerprint's characteristics and the noise in the fingerprint image

The fingerprint image is combined by the ridgelines and valleys of finger's surface. Everybody's fingerprint has its own characters, and the particularity is decided by the characteristics of region lines and their relationship. The definitions of fingerprint image characteristics are various [2]. ASCII bureau puts forward four kinds of characteristics which are used for the details of fingerprint matching: the terminal point of ridgelines, crunodes, compound characteristics (three furcations or intersections) and undefined details. But the most common definition of detail is the particular model that is put forward by American FBI currently. It divides the most significant characteristics of a fingerprint image into terminal points of ridgelines and crunodes. Each clear fingerprint generally has 40~100 such particular points. With these characteristics and their relationship of ridgelines in some part of regions,

the owner's identity can be confirmed [2]. The function of particular characteristic extraction algorithm seriously depends on the quality of the input fingerprints image. However, the input fingerprint image is always different from the original fingerprint image, which is caused by distortion of fingerprint image. Much more of this distortion and aberrance is caused by the collecting process of the fingerprint form. The fingerprint image that we've gotten is a planar image which forms a three-dimensional finger. Fingerprint images collected will have some distortion because of the difference of collecting pressure. Generally speaking, this kind of process is very difficult to be controlled.

In general, if a finger gets in touch with collective equipment completely, the ridgelines structure information of the finger will be collected completely. But some factors will cause the fingerprint images unauthentic, such as dryness degree of skin of the finger, sweat-stain, dirty-stain and skin disease. For example, some parts of ridgelines structure information can't touch with collective plate, some false information will be collected, and both the artificial collecting fingerprints and the wounded finger would change the particular information of finger permanently or temporarily, so the false information will be collected. In the meantime, the collecting equipment itself will also cause some noise interference, which will cause that fingerprint image that need to be analyzed brings amount of suspicious characteristics and real characteristics in great quantities to be neglected, leading into a great deal of false information. To ensure the function of the detail characteristic algorithm, fingerprint image needs to be done image enhancement.

In a fingerprint image, the region in which ridgelines and valleys are clear and can be extracted the feature points rightly is called well region. The region in which the information of ridgelines and valleys is broken by a few of folding lines or dirty-stains, but it still can be seen, and enough information of ridgelines and valleys can be provided by the regions nearby, this kinds of region are called destructive but restorable region. If the information of ridgelines and valleys is destroyed by a great deal of noise, and the neighbor regions can't provide enough information of the ridgelines and valleys, it is called irrecoverable region. The two former regions generally can be called restorable regions. And the purpose of fingerprint image enhancement is to improve the clearness of ridgelines' information in the restorable regions, and to delete the irrecoverable regions.

3 Fingerprint image enhancement

There are two kinds of image enhancement methods. They are spatial domain enhancement method and frequency domain enhancement method. For the methods in spatial domain, the Gray level value of the image pixel is operated directly [8]. For example, calculating the average gray level value of the pixels in a small region, which includes a interested point, then using the average value instead of the original gray level value of this point. This method is called smooth processing. The theory of image enhancement methods in spatial domain can be described with Fig 1 and Eq.(1):

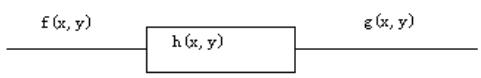


Figure1 Spatial domain enhancement.

$$g(x,y) = f(x,y)h(x,y) \qquad (1)$$

In the Eq.(1), $f(x,y)$ is the image before enhancement, and $g(x,y)$ is the image after enhancement, $h(x,y)$ is the space operation function.

Image enhancement methods in frequency domain are that operating with the value of image transformation in some transformation field. For example, processing the image by Fourier transform, after that revising the spectrum of the image with some method (for example, filter etc), and then inverse transforming the revised transformation value to the spatial domain, finally we gain the enhanced image. This is an indirect method. The theory of the image enhancement methods in frequency domain can be described with Fig 2, Eq.(2) and Eq.(3):

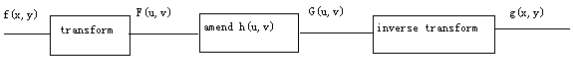


Figure 2 Frequency domain enhancement.

$$G(u,v) = F(u,v)h(u,v) \quad (2)$$

$$g(x,y) = F^{-1}\{G(u,v)\} \quad (3)$$

$f(x,y)$ is the image before being processed, $g(x,y)$ is the image after being processing. $F(u,v)$ is the Fourier spectrum of the image before being processed, and $G(u,v)$ is the Fourier spectrum of the image after being processed. $h(u,v)$ is the revised filter function [4,5,6].

An important feature of the fingerprint image is that in the fingerprint image many ridgelines are parallel. Using this feature, even if the current ridgeline is not continuous, we can also get the direction of this ridgeline through observing the direction of the neighbor ridgelines in a small region. But in this small region, those ridgelines which directions are different from the direction of the current ridgeline are always added noise. So using this feature, auto-adaptive matching filter can be designed. For each point in a fingerprint image, according to the information of the points in its neighbor field, using filter on the point can enhance the points which directions are the same as the directions of the ridgelines, and weaken the points which directions are different from the direction of the ridgelines. So the noise of the fingerprint image can be restrained and the purpose of fingerprint image enhancement can be achieved. On the other hand, because the directions of every small region are always different, filters with different parameters should be adopted on the basis of the direction of every small region. Auto-adaptive filter is usually adopted practically [7]. Many literatures discuss with the fingerprint image enhancement. Coetzee etc. processes the input gray image with Marr-Hildreth edge operator to gain the image of the ridgeline edges, and enhances the image with Convolution Operation template. Randolph etc. put forward a kind of method that uses a group of direction filter to enhance the input Binary Image. O'Gorman elicits k k template modulus through the direction of ridgelines in local regions, and discusses the design of the filter in detail. Sherlock enhances the fingerprint image with Fourier filter, and obtains pretty well effect [3]. But in the case of the input fingerprint image being not good, the directions of some local regions can't be obtained accurately. Hon puts forward the Gabor filter method which both with frequency selection and direction selection to enhance the fingerprint image. This kind of method obtains the filtering images through processing the input fingerprint image with a series of Gabor filters, and estimates the image of directions by these filtering images [9]. This kind of method can obtain a very good effect when the quality of the input fingerprint image is low. But the computing consumption of calculating the local directions image is very high, so that it's difficult to be run in the network system. Hong etc. proposed an advanced algorithm to lessen the computing load.

4 Fragmentary fingerprint image recon struction

While picking up the fingerprint lines, the information of directions must be used [2]. Only if the picked up information of directions is accurate, appropriate template can be constructed, and the Convolution Operation with fingerprint image can be made, eventually, smooth the image along direction of tangent of the fingerprint ridgelines to connect the discontinuous of the fingerprint ridgelines [10]. This process can reserve the characteristic information of the fingerprint image, and reduce the appearance of false characteristic points, and preserve the real characteristic points. Due to a fingerprint image is often contaminated by information noise and information losing of the edge structure and so on, the directions of fingerprints got by different kinds of methods are not always exact. Because the fragmentary fingerprint image contains a mass of blurry regions, it's difficult to calculate the right direction of ridgelines. According to analyze the direction of ridgelines, it can be seen that except some special regions, the change of the directions of ridgelines are generally gentle. As mentioned above, the fingerprint image has many parallel ridgelines. Using

this feature, even if ridgeline is not continuous, the direction of this ridgeline also can be elicited through observing the direction of the ridgeline in a small neighbor region.

5 Conclusions

Obviously, there are about five kinds of image reconstruction methods:

(1) simultaneous equations method (also called Matrix method);

(2) Inverse Projection Method;

(3) Fourier Transformation Method;

(4) Filter Method --- Inverse projection(also called Convolution Operation method)

(5) Gradient Method [1].

Fragmentary fingerprint image reconstruction is still a problem to be solved. The directions of ridgelines aren't grasped well and it is easy to bring forward false characteristic information, which makes it difficult to accomplish the fingerprint identification. Further research in the field of partial fingerprint reconstruction technique will be done in the future.

References

[1] Kenneth R.Castleman, Digital Image Processing, Publishing House of Electronics Industroy, 2002.

[2] Liu Shaocong,Geng Qingjie, Dactylography, Publishing House of Police Officer Education,1994(in Chinese).

[3] O'Gorman L, Ninckerson V,An Approach to Fingerprint Filter Desige, Pattern Recognition,1989.

[4] Chen Jing and Luo Bin, A New Method of Fingerprint image Enhancement, Journal of Fuyan Teachers College (Natural Science),vol 24,Number 3 Sep.2007(in Chinese).

[5] WANG Feng,LI Ji-gui,Fingerprint image enhancement algorithm research[J],Modern Computer, 2003(in chinese).

[6] CHEN Da-hai,GUO Lei,LI Hai, Gabor filter enhancement algorithm base on binarization fingerprint image, Computer Engineering and Application, 2007(in chinese).

[7] ZHANG Ming, Chen W U.CHEN Na,An algorithm based on orientation filter for fingerprint image enhancement[J],Microcomputer Development, 2005(in Chinese).

[8] Hong L,Wan Y,Jain A.,Fingerprint image enhancement, IEEE Transaction on Pattern Analysis and Machine Intelligence, 1998(in Chinese).

[9] A.B. Smith, C.D. Jones, and E.F. Roberts, "Article Title," *Journal*, Publisher, Location, pp. 1-10, Date.

[10] Z.M.Kovacs, R.Rovatti, M.Frazzoni, Fingerprint ridge distance computation methodologies, Pattern Recognition, 2000.

Design and Implement of Information Extraction System Based on XML

Yanyan Xuan[1] Yan Hu[2]

Dept. Computer Science & Technology, Wuhan University of Technology, Wuhan, 430070, China

Email:1 yy.xuan@163.com; 2 huyan168@sina.com

Abstract

By studying the structure of HTML documents, this paper solves the problem of web information extraction through the standard XML technology and poses an information extraction method based on XML: construct HTMLDOM tree to implement Web cleaning and generate XHTML documents by analyzing HTML web, then analyze the XHTML files through the Xerces-J's DOM methods and construct an XPath generation algorithm; use the advantages of XSLT and XPath technology in the aspects of data location and conversion to automatically learn and generate the information extraction rules and implement the Web information extraction according to the generated XPath.

Keywords: Information Extraction; XML; XPath; XSLT; Extraction Rule

1 Introduction

With the rapid development of the Internet, the Web has become the main source of information. The mount of Web information has increased explosively. How to obtain the necessary information efficiently from the Web becomes an issue requiring urgent solution. At present the majority of Web information publishes in HTML form. The Web vector which is semi-structured and unstructured focuses on not the description but the display of data. It is difficult to analyze through program and it can not provide the structured query language for the users to query efficiently. Therefore information extraction [1] technology becomes necessary. The essence of information extraction technology is extracting factual information from natural language text and describing the information in the structured form for information query, deep excavation of text, automatically answering questions and so on.

The program for information extraction is called wrapper. The main task of constructing wrapper is to compile extraction rules. Therefore, how to compile flexible and efficient extraction rules becomes the research focus in the field of information extraction. At present the research of Web information extraction is mainly about information extraction based on natural language processing, machine learning [2], ontology [3], the HTML structure analysis and Web query. The paper proposes and implements an information extraction method on the basis of studying the technology which has been proposed .It can extract useful information from the Web efficiently for the relational database storage.

2 Information Extraction Principles and System Flow Design

Information Extraction (Information Extraction, or IE) is a process of extracting related data from page sets. Web information extraction is information extraction based on Web pages. Formal description statements are as follows: For a given group of Web page S, define a mapping W. W maps the objects in S to a data structure D which is more structured with a clearer semantics (such as relational database). And the mapping W has the same function with the Web pages set S' which is

similar with S in semantics and structure. Thus, the key to information extraction is defining the mapping W, the extraction rules.

The information extraction system described in this paper gets data from the Web. It processes the raw data through data acquisition and data preprocessing, gets the Web pages in the specific field and stores them in the domain knowledge base as the data input for information extraction phase. Construct HTMLDOM tree by using NekoHTML [5] to analyze HTML documents. Clean the page by traversing the nodes of the tree. Construct HTML documents correspondent to XML grammars. Analyze the XHTML document generated in way of DOM. Use the JTree based XPath generation algorithm implemented in the system to obtain XPath path expressions of the information nodes. Compile the extraction rules according to the expressions and XSLT technology. Implement the automatic generation of information extraction rules on the basis of rules learning. Use the templates to extract information nodes from XHTML documents. And finish the process of extraction. Work flow is shown in Figure 1.

3 Realization Process of Information Extraction

3.1 Brief Introduction to XML, XPath and XSLT

XML (extensible Markup Language) language is a meta-markup language defined by W3C. It is the standard of describing data content and structure on the network. Before XML was published, the development of the Internet was constrained by the following issues:

① HTML could not describe data content which is the very necessity of data retrieval and electronic commerce.

② HTML's ability to describe data presentation was very inadequate. For example, HTML could not describe the objects of vector graphics, scientific symbols and so on. At present, they can only be presented by images.

③ The status of HTML example markup language can not meet the development demand of the new mark needs at all.

The appearance of XML makes all the above issues well resolved. XML contains three elements: DTD document type definitions or XML Schema outline, XSL extensible style language and XLink extensible link language. DTD and XML Schema provide the logical structure of XML documents and define the elements in XML documents, elements' attributes as well as the relationship between the elements and the elements attributes. XSL is the language to provide the presentation style of XML documents. XLink will further extend the existing simple links on the Web.

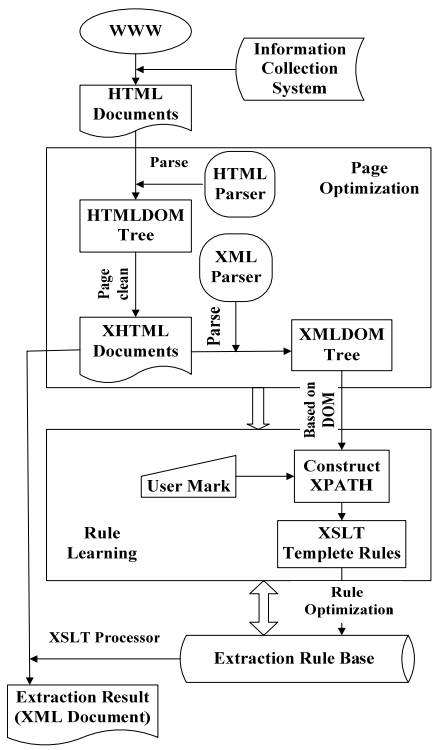


Figure1 Information Extraction Work Flow

XPath is the fourth generation of declaration language used to locate the nodes in XML documents. XPath's location path designates which nodes in the documents are needed but does not designate the algorithm to search these nodes. As long as an XPath sentence is transferred to the method, XPath engine will decide how to search all the nodes satisfying the expression. XPath's basic grammar is composed of expressions. The paper uses XPath expressions to compile XSLT template and maps the data to XML. XPath expressions designate the path from root <html>

elements to information nodes. Suppose the TAG sequence of an XHTML is "/html[1]/body[1]/table[2]/tr[1]/td[4]/text()", it means the content of Table 2, Line 1, rank 4 in thedocument. This XPath is just the path expression of required data.

XSLT, drafted and produced by W3C, is a language used to transform XML documents structure. It is a branch of the XSL and a standard widely used at present. It can be used for transformation between different data formats. In the process of transformation, XSLT processor reads XML documents and XSLT style sheet, transforms documents based on the instruction in XSLT style sheet, and then output the transformed document. Because XML is a complete tree structure document, while using XSLT to transform XML documents, we need to process some of these information nodes. XPath language is especially used to locate every part of XML documents which can help XSLT to search information nodes quickly and efficiently.

3.2 Analysis and cleaning of HTML pages

HTMLDOM tree is a description style of Web pages. It is a hierarchy tree structure constructed according to the meaning of HTML tags in Web pages where each node is a separate HTML element. HTMLDOM presents HTML document as a tree structure (tree node) with elements, attributes and text which is shown in Figure 2.

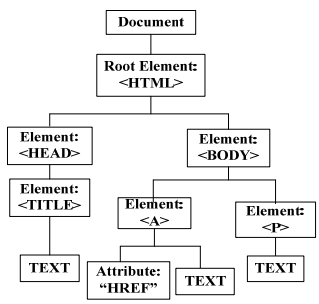


Figure2 HTMLDOM Tree

NekoHTML can analyze HTML documents and use the standard XML interface to access the information. The process is implemented mainly through interfaces provided by Org.w3c.dom (nekohtml and xerces [6] provides the implement of these interfaces). The interfaces in the package are: Node, Document, Element, Text and so on. Node is the root interface of all the nodes in DOM tree. Using these interfaces, we can visit all the nodes in the HTMLDOM tree conveniently.

Steps of transforming HTML documents to XHTML documents are as follows:

① Use NekoHTML to analyze HTML document into HTMLDOM tree.

② Traverse this nodes tree through Node Interface and get all the corresponding nodes of HTML document.

③ Judge through Element interface according to the different type of node elements and remove the non-thematic elements in the page such as SCRIPT, LINK, META, IMG and so on.

④ Encode the TEXT nodes in the page (for HTML tags in the pages).

⑤ Use the processed nodes to compile XHTML documents.

3.3 Extraction Rules

XSLT treat the XML document to be processed as a node tree known as the source tree. It also treats the transforming results as a node tree known as the result tree. The result tree is separated from the source tree. The structure of the result tree can be the same with or different from that of the source tree. While constructing the result tree, the elements from source tree can be filtered and re-sequenced, and any structure also can be added to it. The basic constraints are in agreement with XPath. In XSLT, the transformation is called stylesheet. The stylesheet defines a set of rules to transform source tree to result tree known as template rules. Stylesheet is defined by xsl:stylesheet element in XSLT document and is usually composed of several template rules. A template rule is divided into two parts: patterns used to match the elements in source tree which are defined by XPath syntax, templates used to show how to construct a part of result tree. Template rules are defined by xsl:template element. Its match attribute is a pattern

used to match the elements in source tree, and its content is just the template. Because XSLT can extract data from an XML document and present it as a new XML document, XSLT document is just the extraction rules from the angle of information extraction.

XPath is the main component of XSLT. XPath path expression of the information node to be extracted in the document is necessarily needed while compiling XSLT template. However, it is not an easy task to compile XPath by locating information nodes in the documents. This paper does some research into this problem, and constructs an XPath generation algorithm based on JTree.

Steps are described as follows:

① Based on the XHTML documents after the web page cleaning, use Xerces-J's DOM methods to analyze the XHTML documents.

② Construct a DOM parser to analyze the XHTML files into Document objects.

③ Traverse the methods of DOM tree to find all the Node nodes of this Document Object.

④ Decode the text node in accord with the transformation step④ from HTML documents to XHTML document.

⑤ Traverse the source DOM tree's Node nodes to generate the TreeNode nodes of JTree.

⑥ Use TreeNode nodes to construct XPath expressions.

This paper develops a graphical interface to finish this process. Users can mark the information nodes to be extracted based on their individual needs. And XPath expressions can be obtained automatically. Because of the space problem, the concrete realization of the algorithm is expounded in another article of the author's. XSLT template can be compiled according to the XPath path expression of each information node.

The following is an HTML template compiled by using XPath of information nodes after HTML sample pages are cleaned and location information nodes generate XPath expression.

```
<?xml version="1.0" encoding="gb2312"?>
<xsl:stylesheet version="2.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
<xsl:output encoding="gb2312" indent="yes"/>
<xsl:template match="/HTML[1]/BODY[2]/TABLE[1]/
TBODY[1]/TR[1]">
<content>
  <caption>
<xsl:value-of select="TD[2]/SPAN[1]/text()"/>
</caption>
  <key1>
<xsl:value-of select="TD[3]/text()"/>
</key1>
<key2>
<xsl:value-of select="/TD[4]/text()"/>
</key2>
</content>
</xsl:template>
```

This XSLT style defines the rules for transforming the source XHTML data to the target XML data. Based on these rules, use XSLT processor to implement the transformation and target XML document is obtained. Store the extraction rules obtained through HTML sample pages learning into extraction rules training base. Use inductive learning methods to optimize the extraction rules produced in the sample training phase. When users request to extract the information from pages, use the extraction rules in rules base to extract information from the HTML pages. If there are no corresponding rules in the rules base, study the HTML pages to be extracted to find the page structure, then obtain the new rules and store them into the rules base. This paper develops an application program, a complicated process of using XPath path obtained by users' marked information nodes to compile the templates. While the single information blocks can generate templates quickly and automatically, template rules of multi-information blocks are processed at present by the general XSLT template compiled by the author after obtaining XPath.

4 Conclusions

This paper proposes a solution of XML-based Web

page information extraction. This system can extract the useful information from the HTML pages, express it as structured and expansive XML documents and realize the transformation from semi-structured data to structured data. The information extraction system implemented in this paper is a general Web information extraction system that can be applied to extract data from all the Web pages with a relatively high recall rate and accuracy rate. While analyzing source HTML documents, this paper used JTidy to organize HTML pages and found the fault-tolerant features of JTidy relatively poor. After transforming, it is relatively difficult to analyze the XHTML documents and to carry out the following work. On this issue, this paper studies NekoHTML. It focuses on Parser itself, while XML application is generally constructed around the Parser. So the author uses NekoHTML to analyze HTML pages, access node tree and clean it up, and then compile programs to match node and generate HTML documents. This XHTML completely fits the XML standard and is easy to analyze. This system uses semi-automatic manual marking method when constructing XPath, so it can locate information nodes accurately. The further research direction is to optimize the process to automatically obtain the path expression after studying a few samples. When using the obtained XPath expressions to compile the template rules, it realizes the automatic generation of single-information block's rules. The automatic generation of multi-information block's rules is also the focus of further research.

References

[1] Laender H F, Ribeiro-Neto B A, A S da Silva et al, "A Brief Survey of Web Data Extraction Tools", SIGMOD Record, Vol.31, No.2, 2002, pp. 84~93.

[2] R D Doorenbos, O Etzioni, D S Weld, "A scalable comparason-shopping agent for the World-Wide Web", In: ACM Agents'97. New York: ACM Press, 1997, pp.39~48.

[3] D W Embley, "Conceptual-model-based data extraction frommultiple-record Web Pages", Data and Knowledge Engineering, Vol.31, No.3, 1999, pp.227~251.

[4] Bray T，Paoli J, Sperberg-McQueen C M.Extensible Markup Language (XML)1.0, W3C recommendation, http://www.w3.org/TR/1998/REC-xml-19802108, December 1997.

[5] CyberNeko HTML Parser, http://people.apache.org/~andyc/neko/doc/html/.

[6] http://xerces.apache.org/xerces-j/index.html.

Image Segmentation Using Improved PCNN

Feng Xu[1] Li Guo[2] Daguo Shan[1] Hongchen Yang[1]

1 Department of Forensic Science & Technology, China Criminal Police University, Shenyang 110035, Liaoning, China

2 College of Information Science & Engineering, Northeastern University, Shenyang 110004, Liaoning, China

Email: sunshan-xu@sohu.com

Abstract

The methods proposed in this paper improved the classical Pulse-coupled neural network (PCNN), which is an important tool on image segmentation. Through the external input grows linearly as the iterative numbers increasing, the synchronous bursts of non-linking neurons with different input were generated in order to realize the multi-object segmentation. The approach of region merging is used after segmentation. Region merging was carried out according to region area and region homogeneity. It was proved that the method was valid and more accurate than other common algorithms of image segmentation.

Keywords: Pulse-coupled neural network (PCNN); image segmentation; multi-object segmentation, region merging

1 Introduction

As image segmentation is a crucial process of range image analysis, a number of range image segmentation techniques have been proposed in the literatures. They can be roughly classified into four categories: threshold based, edge based, region based and hybrid techniques. Thresholding is one of the old, simple and popular techniques of segmenting images consisting of bright objects against dark backgrounds or vice versa. Edge-based segmentation approaches firstly obtain step edges and roof edges respectively by locating the points with depth discontinuities and normal inconsistencies, then link them to divide image into different regions. Region-based segmentation approaches include region-growing methods, clustering methods, etc. Hybrid segmentation approaches use edge detection techniques to estimate the number of regions and to initialize region-growing or clustering algorithm of region-based segmentation, which improves the accuracy of locating boundaries [1].

Pulse-coupled neural network (PCNN) is a simplified model for describing the synchronization behavior observed experimentally form the wide range of brain cortex of cats [2,3], and as a processing tool, has been widely applicable to the research work on image segmentation [4], image fusion [5], image enhancement [6], image target detection [7], shortest path solution [8], and image feature extraction [9], due to its specific characteristic of grouping neurons according to spatial proximity and intensity similarity. The network parameters of PCNN must be selected for different image segmentation in order to obtain satisfactory result, which has seriously affected application of PCNN. In this paper, we present an algorithm that combines improved PCNN and region growing technique to partition the image into several meaningful components.

2 Model of improved PCNN

2.1 Improved PCNN

The Pulsed Couple neural network (PCNN) is a digital simulation of the visual cortex of the mammals [10]. Since the model is highly nonlinear and complex interactions, it is difficult to use mathematical methods

to control and interpretation of results of neurons. For this reason, Indblad and Kinser proposed the improved the PCNN, and get the following model as Fig.1. The PCNN neuron consists of three parts: dendritic tree, linking modulation, and pulse generator, as shown in Figure. 1.

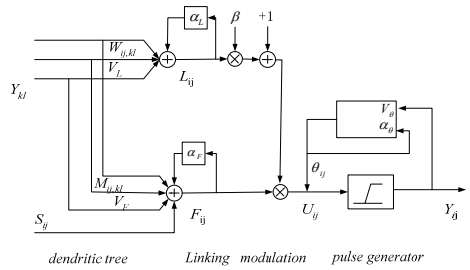


Figure1 the Model of PCNN

Where S_{ij} is the input signal, and every pixel corresponds to a neuron, F_{ij} is the feedback input, L_{ij} is the link input, U_{ij} is the internal activity, Y_{ij} is the pulse output, θ_{ij} is the dynamic threshold, the weight matrix W is the local interconnection, β is the link constant, V_θ is the decay constant of the dynamic threshold. α_θ is time constant α_F and α_L are the time constants; generally, $\alpha_F < \alpha_L$.

To improve the performance of the shortest paths search, an improved optimal search algorithm based on PCNN is proposed. The following expressions mathematically describe its model.

$$F_{ij}(t) = S_{ij} \tag{1}$$

$$L_{ij} = V_l \sum_{k,l} w_{ij,kl} Y_{kl}(n-1) \tag{2}$$

$$U_{ij}(n) = F_{ij}(n)[1 + \beta L_{ij}(n)] \tag{3}$$

$$\theta_{ij}(n) = \begin{cases} V_\theta & t = t_1 \\ V_\theta e^{-aT(t-t_1)} & t_1 < t < t_2 \\ V_\theta & t = t_2 \end{cases} \tag{4}$$

$$Y_{ij}(n) = \begin{cases} 1, & U_{ij}(n) > \theta_{ij}(n) \\ 0, & other \end{cases} \tag{5}$$

For the improved model of PCNN, the external input grows linearly as the iterative numbers increasing. And between the neurons each path has its own threshold θ_{ij}. After the fire, θ_{ij} makes step change with the path change. The same neighborhood of the neurons mutual links through θ_i.. Only the fired neurons can influence the outside. Clearly, the improved model PCNN retain the characteristics of the original model, and better suited to describe the shortest path problem. Based on PCNN characteristics, the shortest path can be found by the search algorithms with excellent performance.

2.2 Improved PCNN algorithm design steps

1) To meet the direction of auto-wave transmission is consistent with the shortest path, the designed coupled character of the neurons is:

a) If only neurons *i* and *j* connected, their weight is w_{ij}, neuron *i* fire at the first time, the neuron *j* will be delayed fire at t_i+w_{ij} , according to Eq. (1).

b) There has *m* neurons i_1 i_2 i_3……i_m, connected with neuron *j*, and its linking weight respectively is w_{i1j} w_{i2j}……w_{imj}. when One of the neurons i_1 i_2 i_3……i_m, fired before neuron *j* fired, Only the autowave who first reach the neurons *j* can continue to spread, according to Eq. (2).

c) We define the each neuron fired only once in the process of optimization, according to Eq. (3).

2) The shortest path algorithm steps:

a) Initialization. Let $y_j(0)=0$ and $\theta_j(0)=0$. for $\forall j \in G(V, E)$,

b) Fire the start neuron, i.e.$\theta_{start}(1)=-\varepsilon \rightarrow y_{start}(1)=1$,and keep the states of the rest neurons unchanged (where $\varepsilon > 0$).

c) Autowave travel. $\forall j \in G(V, E)$, If $w_{ij} \neq 0$,calculate:

$$\theta_k = \begin{cases} V_\theta - \dfrac{V_\theta}{w_{ij}}\Delta t \\ \theta_k(k-1) - \dfrac{V_\theta}{w_{ij}}\Delta t \\ 0 \end{cases} \tag{6}$$

$$\theta_j(k) = \min\left\{\theta_{ij}(k) : w_{ij} \neq 0\right\} \tag{7}$$

$$\theta_k = \begin{cases} 1 \\ step(-\theta_j(k)) \end{cases} \tag{8}$$

Where Δt is the iteration step.13671211750

d) Record the route.

$$b_{ij} = \begin{cases} 1 \\ 0 \end{cases} \quad (9)$$

e) Repeat the step c) and d) until all destination neurons in the network are in fire state.

The flowchart of the improved algorithm based on PCNN is shows as Figure2.

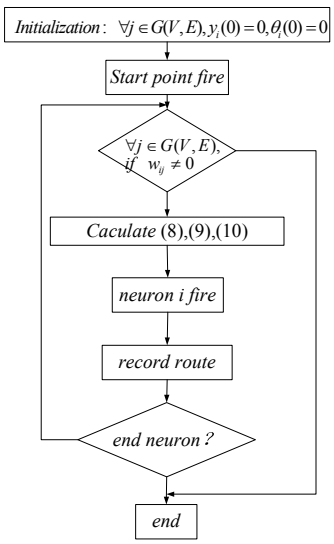


Figure 2 Flowchart on improved PCNNNetwork searching path

2.3 set V_θ

In order to effectively control the nature of the autowave in PCNN. If the every one of neurons is stimulated, it must be stimulated only once in a stimulating period, then network will be reset, the network will automatically repeat wave behavior. This situation can be determined by V_θ. Here X_{max} and X_{min} respectively are the largest and the smallest stimulation in the network. The X_{max} and X_{min} as their stimulate neurons, which naturally stimulate at time $t=T(X_{max})$ and $t=T(X_{min})$ (If X_{min} can naturally stimulate). From the time $t=T(X_{max})$ to $t=T(X_{min})$, every neuron at least stimulates once. Now we reset the network at time $t=T(X_{min})$, i.e. each of initial value of the network is reset to be zeros. So the neuron stimulate at time X_{min}, stimulate time meet conditions:

$$T(X_{\min}) \le 2T(X_{\max}) \quad (10)$$

Where $T(X_{\max}) = \psi_\theta \ln(V_\theta / X_{\max})$ $T(X_{max})$ meet $X_{\min}[1+\beta L(T(X_{\min}))] \ge V_\theta \exp(-T(X_{\min})/\tau_\theta)$ So we can get

$$V_\theta \ge \frac{X^2{}_{\max}}{X_{\min}[1+\beta L(T(X_{\min}))]} \quad (11)$$

We can obtain the shortest path from the last fired neuron to the first by the route record matrix. All of these shortest paths are obtained simultaneously, automatically and parallelly.

Generally, the above-mentioned algorithms followed by improved PCNN produce meaningful image segmentations. Some applications, however, may require further merging of some regions. For region merging, a number of similarity (between regions) criteria have been proposed, each of which has its specific applications. In addition to the above over-segmentation reduction method, there still remain neighboring regions that could by merging yield a meaningful segmentation, on the principal that each region is homogeneous and sufficiently different from its neighbors. Some criteria are based on area of the regions. If the area is very small, it must be redundancy. The objective cost function used in this paper is the square error of the piecewise constant approximation of the observed image, which yields a measure of the approximation accuracy and is defined over the space of partitions.

3 Experiment results

We present a few of the results obtained by the application of the proposed algorithm on standard images. It can be seen from Table 1, we proposed segmentation algorithm is significantly less than the other methods, Segmentation is obviously better than the watershed and Gaussian gradient operator. And the number of the segmenting regions is smaller than traditional methods. From the Table 1, we also can see the time using proposed algorithm is shorter than other method. Some experimental results are shown in Fig. 3. Experiments on the segmentation based on traditional watershed transform and morphological gradient method are conducted for comparison. The original images are showed in Figure 3(a). The experimental results in Figure 3(b) indicate that the watershed approach, which

Table 1 Comparision between traditional algorithm andproposed method

Method	Segment region	Time (ms)
Watershed	29005	165
Guassian gradient	5421	172
Morphological gradient	7486	122
Proposed method	1475	118

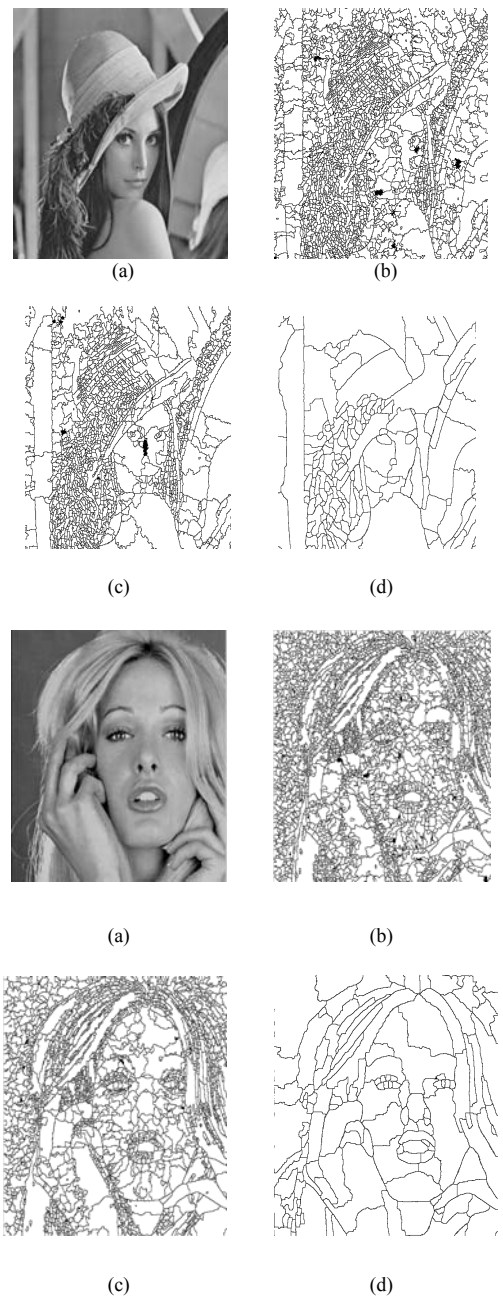


Figure3 Comparison of segmentation results

(a) original image; (b) watershed result;
(c) multiscale gradient result; (d) improved PCNN result

is existed over-segmentation. In this approach, only used the watershed transform to segmentation and region merging.

Fig. 3(c) is the final result of segmentation in which morphological gradient image is applied as referenced images, the phenomenon of over-segmentation is serious. Fig. 3(d) is the results of improved PCNN approach which are brought forward in this paper are used. Using the proposed algorithm, a significant decrease over-segmentation and segmentation results is clear.

4 Conclusion

The image segmentation based on improved PCNN is proposed in this paper, the external input grows linearly as the iterative numbers increasing. And between the neurons each path has its own threshold. After the fire, threshold makes step change with the path change. The same neighborhood of the neurons mutual links through threshold Only the fired neurons can influence the outside. Region merging is carried out according to region area and region homogeneity after segmentation. Experimental results show that the proposed algorithms can significantly reduce the computational cost of image segmentation while efficiently improving segmentation accuracy.

References

[1] Cheng HD, Jiang XH, Sun Y et al, "Color image segmentation: advance and prospects," Pattern Recognition, Vol.34, No.12, 2001, pp. 2259-2281.

[2] R Eckhorn, H J Reitboeck, M Arndt, P Dicke, "Feature liking via synchronization among distributed assemblies: Simulations of results from cat visual cortex ," Neural Computer, Vol.2, No.3, 1990, pp. 293-307.

[3] Eckhorn, R., Reiboeck, H.J., Arndt, M. et al, A neural networks for feature linking via synchronous activity: Results from cat visual cortex and from simulations, In Model of Brain Function(ed. Cotterill, R.M J), Cambridge University Press, Cambridge, 1989.

[4] Johnson J L, Padgett M L, "PCNN medels and applications," IEEE Trans on Neural Networks, Vol.10, No.3, 1999, pp. 480-498.

[5] Eckhorn, P., "Neural mechanisms of scene segmentation: recording from the visual cortex suggest basic circuits linking field models," IEEE Trans. Neural Network, Vol.10, No.3, 1999, pp. 464-479.

[6] Zhang J Y, Lu Z J, Shi L, et al, "Filtering images contaminated with pep and salt type noise with pulse-coupled neural networks," Science in china Ser. F: Information Sciences, Vol.48, No.3, 2005, pp. 322-334.

[7] Yamaoka D, Ogawa Y, Ishimura K, "Motion segmentation using pulse-coupled neural network," In: SICE kmual Conference, in Fukui, 2003, pp. 2778-2783.

[8] Broussard R P, Rogers S K, Oxley M E, et al, "Physiologically motivated image fusion for object detection using a pulse-coupled neural network," IEEE Trans Neural Network, Vol.10, No.3, 1990, pp. 554-563.

[9] Caufield H J, Kinser J M, "Finding the shortest path in the short time using PCNN," IEEE Trans on Neural Networks, Vol.10, No.3, 1999, pp. 604-606.

[10] R Eckhorn, H J Reitboeck, M Arndt,P Dicke, "Feature liking via synchronization among distributed assemblies: Simulations of results from cat visual cortex," Neural Computer, Vol.2, No.3, 1990, pp. 293-307.

Evolving an Image Comparison Matrix Using Genetic Programming

Xiaofei Wu

Department of Computer Crime Investigation, China Police University for Criminal Investigation, Shenyang 110035, Liaoning, China

Email: chenou2002@163.com

Abstract

This paper is to apply the genetic programming technique to evolve a matrix which provides a quantitative indicator to detect subtle difference between two apparently similar pictures. The images here refer to the PPM format images.

The GP environment I used is the LIL-GP which is a C language system for developing genetic programming applications. To train the LIL-GP to generate the target matrix, I defined some terminals and functions composing the GP programs. I also defined the fitness function which was very important to my problem. The fitness case was an array of PPM images composed of one original image followed by some lower quality version images.

The ideal of the fitness function was to take the standard image comparison matrix as the reference which can guide the GP to find the target matrix by compute the distance between the reference one and the GP program. The target matrix should perform better than the reference one.

Keywords：Genetic Programming; Image Comparison.

1 Important information

There have been some attempts in developing an image comparison matrix, but the results have not been satisfactory and competitive yet. Artificial Intelligence techniques have been proven highly successful at the problems of navigation, task prioritization, and obstacle avoidance.

GP technology naturally adapt to the problem of my project. The matrix of image comparison may be composes of pixel values and some operators all of which can be the components of the GP as the function and terminal sets. And by some selection operations, the matrix, in other word, the GP program can evolve towards the target matrix if the fitness function is defined properly.

2 Knowledge and Techniques

2.1 PPM Image

The image format I used in the paper is PPM. The main reason is that it is very easy to write and analyze programs to process this format. We can use subroutine libraries to read and interpret the format conveniently and accurately.

2.2 Genetic Programming

Genetic programming is an optimization technique based on the concepts of Darwinian evolution. A population of individuals, each of which representing a potential solution to the problem to be optimized, undergoes a process analogous to biological evolution in order to derive an optimal or at least near-optimal solution. The solution offered by each individual is assigned a fitness which is a single numerical value indicating how well that solution performs. New individuals are generated by procedures analogous to

biological reproduction, parents of which chosen from the existing population with a probability proportional to their fitness. The new individuals may replace less fit members of the population, so the overall population fitness improves with each generation.

2.3 LIL-GP

LIL-GP evolves trees whose nodes are C function pointers. So tree evaluation is done entirely with complied code which gives us a manifold speed increase and allows us to handle much large problem(bigger populations , more generations), and is portable to a wide variety of platforms.

A tree is stored as an array of type Inode. An Inode is a union which can be a pointer to a function structure, an ephemeral random constant (ERC) structure, or an integer.

LIL-GP's core is a tree representation that is compact yet amenable to extremely fast evaluation. Trees are stored as preorder expressions, with special symbols inserted to indicate ephemeral random constants and conditionally evaluated sub-trees. The compactness of the representation allows larger populations to be stored in real memory, and the fact that trees are stored as contiguous blocks of memory minimizes the impact of paging on performance. The selection methods and genetic operators are implemented in an object-oriented fashion, allowing new routines to be added easily. Checkpoint files save the entire state of the run, allowing a promising run to be extended, or an interrupted run to be completed. A portable random number generator provides consistent results across platforms.

To create a program in LIL-GP, the user must provide one C function for each GP function and terminal, in addition to a fitness evaluation function. Several callback hooks are provided for initialization, customized output, and reading and writing user state information to checkpoint files. Once the user code is written, the executable needs only to be compiled once. All parameters and operators are available to switch between single and multiple population runs just by modifying the parameter file.

3 Implementing Problems

3.1 Statement of the goal

The fundamental problem is to discover a matrix which can quantitatively indicate how broken an image is compared with the original image. If the matrix is good enough, for the input vector of image pairs that has an increase trend of the quality, the output should be an array of values with a decrease trend.

As the figure1 illustrates, X stands for taking a pair of images, one of which has 100 percents quantity and another one with X percents of quantity. The output values should inverse proportional to the input.

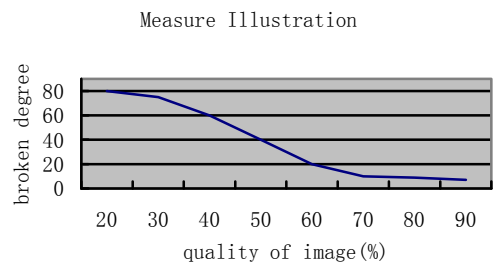


Figure1 Broken degree as a function of quality of image field.

3.2 Implementation

General: To implement the problem in LIL-GP, there are five files had to be written to embed in the system kernel. These files are app.c, app.h, function.c, function.h and appdef.h. These codes can be divided into two categories: C functions implementing functions and terminals, and user callbacks.

The user callbacks are placed in a file named app.c. It does some application specific tasks like function sets initializations, calculation of fitness, etc. The other group of C functions, usually placed in function.c is the codes called by the kernel during tree evaluation.

There are two defined constants that the kernel of LIL-GP needs in apped.h. One is MAXARGS which is the maximum number of arguments (children) for any function. The other is DATATYPE which is the C data type returned by all functions and terminals.

Appdef.h is also a place to put application-specific defines that we may need. So as not to conflict with any kernel defines, all application defines I did was prefixed with APP_.

Functions and Terminals: two of the user callbacks are required to create the function sets. All the others must be present, but they can be just stubs if we don't want to make use of them.

The first user callback is used to create tables for each function set. There may be more than one function set when individuals are represented by multiple trees, since each tree can have its own function set. Each function set is an array of type function. The following tables show, for each type of node, what the eight fields of the corresponding function structure should be.

Table 1 Ordinary Function

Code	the c function implementing the function
Ephem_gen	NULL
Ephem_str	NULL
Arity	the arity of the function (greater than zero)
String	the name of the functions
Type	func_data or func_expr, as appropriate
Evaltree	-1
Index	0

Table 2 Ordinary Terminal

Code	the C function implementing the terminal
Ephem_gen	NULL
Ephem_str	NULL
Arity	0
String	the name of the terminal
Type	TERM_NORM
Evaltree	-1
Index	0

Table 3 Ephemeral Random Constant Terminal

Code	NULL
Ephem_gen	C function to generate new random values
Ephem_str	the C function to print values to a string
Arity	0
String	the generic name of the terminal .(printed trees will almost always have the string representing the value of the terminal , rather than this name.)
Type	TERM_ERC
Evaltree	-1
Index	0

Table 4 Evaluation Function/Terminal

Code	NULL
Ephem_gen	NULL
Ephem_str	NULL
Arity	-1. (the kernel will determine the arity by looking at the argument terminals in the target tree)
String	the name of this function/terminal
Type	EVAL_DATA(EVAL_EXPR) appropriate
Evaltree	the number of the tree to evaluate when this function is hit
Index	0

Table 5 Argument Terminal

Code	NULL
Ephem_gen	NULL
Ephem_str	NULL
Arity	0
String	the name of this terminal
Type	TERM_ARG
Evaltree	the argument number (which child of the corresponding evaluation function this terminal represents)
Index	0

After the function table is created, a list of function sets needs to be created that reference it. An array of type function set should be created with one member for each function set. The size field should be set to the number of functions and terminals in it, and the Cset field should point to the function table.

There are 6 terminals in my program. They respectively stand for the red, blue, and green value of a pixel in the input image1 and image2. And the four operator functions work just as the normal mathematic operators. For example, the "add" function may take the red value of the first pixel of image1 and the red value of the first pixel of the image2 and add them as the result value. For every ordinary function and terminal in the problem, there is a C function to implement the action of that node.

Each C function is passed two arguments and what it does with these arguments depends on it implementing a function or a terminal, and if it is a function, what type of function. All these C functions return the user-defined type DATATYPE. There are two

types of functions, referred to in LIL-GP as type "DATA" and "EXPR". If the function is of type DATA, when it is found in a tree, all its children will be evaluated and their return values passed to the user code implementing the function.

If the LIL-GP function is of type EXPR, the user code passes pointers to its children, which it can then ask the kernel to evaluate if needed. It can evaluate each child as many times as appropriate, or not at all.

If the function is of type DATA, it can ignore the integer passed to it. The argument will be an array of arguments, one element for each child. The C function should reference the d field of each element to get that child's value. For instance, consider the two-argument addition function in my code.

```
DATATYPE f_add(int tree, farg *args)
{
      Return args[0].d+args[1].d;
}
```

When this function occurs in evaluating a tree, the LIL-GP kernel will evaluate the children, store their values in an array, and call the C function. For type EXPR functions, the t field of each array element should be accessed. It is a pointer to the corresponding child. This pointer can be passed to the evaluate_tree() C function to actually do the evaluation. Evaluate_tree() also needs to be passed the integer argument(called tree in this case). C functions implementing terminals ignore both arguments passed to them. A simple example is the independent variable terminal red1 which stands for the red value of a pixel of the first image.

```
DATATYPE f_red1(int tree, farg *args)
{
      Return g.r1;
}
```

This function just returns the value of the red value of the pixel which has previously been stored in a global variable by the application fitness evaluation function.

To create a terminal that acts as an ephemeral random constant, two C functions need to be written. One will generate a new constant, and the other will print its value to a string. The first is passed a pointer to a DATATYPE, it generates a new value and places it in the pointer.

```
Void f_erc_gen(DATATYPE *r)
{
  *r=(random_double()*10.0;)
}
```

This function generates a random real number in the

interval [0;10) (assuming that DATATYPE is defined to be double or some compatible type). The second function is used when printing out individuals. It is passed a DATATYPE value. It should create a string representing that value and return it. Typically this will print the value into a buffer and return the buffer's address. The buffer should be declared static. It should not be dynamically allocated (as there is no code to free it). Below is an example.

```
Char *f_erc_print(DATATYPE d)
{
   Static char buffer[20];
   Sprintf(buffer, "%.5f",d );
       Return buffer;
}
```

Assuming again that DATATYPE is double or something compatible; this will print the value to five decimal places.

Evaluation and Argument Functions: No user code needs to be written to support the ADF functions or corresponding argument terminals. Special entries are made in the function table for them, and the kernel handles the evaluation internally.

Evaluation function with arguments has type DATA or EXPR, just like ordinary function. If the type is DATA, when the evaluation function is hit, each child is evaluated once, and the return values are made available via the argument terminals in the evaluated

tree. If the type is EXPR, then the children are evaluated only when the evaluation of the target tree hits the appropriate argument terminal (and if the same argument terminal is hit multiple times, the child is revalued each time).

Application Initialization: There are two functions provided for application-specific initialization. Function One is passed an integer flag indicating whether the run is starting from a checkpoint or not. It should return 0 to indicate success, or anything else to abort the run. This function should do jobs such as memory allocation and reading parameters.

I also do two initial works in this function. One thing is to store the training images. There are total 20 images used in my work. They come from 2 initial images with PPM type. Each initial image will be processed by the CJEPG script to produce difference versions with lower quality. The first 10 images are used as the training images to train the LIL-GP to produce the target matrix. And the next 10 images will be used to test the matrix found by GP. The other thing to do in this function is to set the reference matrix which will be used in the eval_fitness() function. In this paper I use matrix ABS(image1-image2) as the reference matrix and It means the subtraction operator applying to the every pixel of image1 and image2.

Function Two is called at the end of the run, and used to do things like free memory.

Fitness Evaluation Function: This function is called whenever an individual is evaluated. It is passed a pointer to an individual structure. It should fill in these fields:

The function can evaluate trees of the individual by calling evaluate_tree(), passing it a pointer to the tree data and the tree number. Typically the function will iterate over all the fitness cases. The global variable g, which is a user-defined structure, is used to pass information between app_eval_fitness() and the functions and terminals. In my code, for example, g.r1 is set to the red value of the pixel of one PPM image. When the evaluation reaches the independent variable terminal r1, the implementing function simply reads this value and returns it.

This function consists of two parts. One is to evaluate the tree. The other is to evaluate the fitness of each program. For the fitness evaluation here, the idea is to use the reference matrix mentioned before to guide the GP. And let the GP to try to find the matrix.

Custom Output: After every evaluation of each generation, lil-gp calls the function app_end_of_evaluation(). It is passed a generation number, a pointer to the entire population, statistics for the run and generation, and a flag indicating whether a new best-of-run individual has been found or not. It should return a 1 or 0, indicating whether the user termination criterion has been met and the run should stop.

One thing to do is to determine if the user-defined condition is met after every generation. In my case, I check if the hits of the individual are equal to the fitness_cases.

The other purpose of this function is to plot the result matrix compared with the reference matrix and also the result of applying the found matrix with a new image. To do so, it will first save the 9 output value in diff_matrix_found. And save the new image result in diff_new_image. Since we already have the reference value stored during the function app_initialize(). We can now write them to an output file and there should be 4 columns of values ready for the GNUPLOT utility.

In versions of LIL-GP prior to 0.99b, it was an undocumented feature that by modifying the parameter database, the breeding parameters could be altered dynamically during the run. If you take advantage of this, you must now call rebuild_breeding_table() after modifying the parameters, and pass it the multi-pop pointer passed to you. If you do not, your changes to the parameter database will have no effect. This ability is now considered a bona fide feature of lil-gp, and will be supported in future release.

Changers to the subpopulation exchange topology parameters underwent a similar change. If you change the parameters during the run, you should call a function named rebuild_exchange_topology() after making changes in order for them to have any effects.

Some kernel operations (for instance, restarting

from a checkpoint file) imply rebuilding the breeding and topology tables from the parameter database. You should only make changes to these parameters when you intend to immediately call the appropriate rebuilding functions, otherwise unpredictable things will occur.

Another user callback app_end_of_breeding() is called after the new population is created each generation. This is passed the generation number and the population structure, just as in the end of evaluation callback, but no statistics information.

Results

Now let's examine the result of the GP. At the beginning I defined my fitness function as just requiring a decreasing trend, not an as close as possible match of the original values. The result output file is:

"mydata_file"

10	2862.54331	5269571.000000	79.105620
20	1946.328679	3505962.000000	199.113099
30	1566.393507	2796144.000000	58.359304
40	1225.504587	2467333.000000	53.514532
50	1092.254370	2173382.000000	31.829865
60	1003.350405	1948194.000000	69.910032
70	926.457033	1645235.000000	61.003783
80	776.187434	1210116.000000	41.924350
90	523.485847	769067.000000	42.311684

The four columns in the data file stand for respectively the quality of the image, the found matrix, the reference matrix and the found matrix applying to a new image. I then use the GNUPLOT utility to give a more intuitive view. The gnuplot script is:

Set terminal postscript eps

Set output "myplot.eps"

Plot "mydata_file" using1:2, "mydata_file"using 1:3 with lines, "mydata_file" using 1:4 with lines

And the matrix found by GP is described below.

```
===BEST-OF-RUN===
        Current generation :0
                Generation:0
                        nodes:63
                        depth:5
                        hits:1
TOP INDIVIDUAL:
--#1—
                        Hits:1
                    Raw fitness:8.0000
                Standardized fitness:8.0000
                    Adjusted fitness:8.0000
TREE:
(/(*(*(/(/ R1    R2)
          (+R2    R2)
        (/(*-0.55776      R2)
          (/G2      0.28254)))
(/(+(/0.56379      B2)
     (-G2            G1))
  (+(+B2            G1
     (-0.81836      R1))))
   (*(*(-(* G2              0.26121)
          (+G1              B1))
          (/(*R2            R2)
          (-B1          -0.59134)))
        (+(+(*R1            B1)
           (/ 0.59520       G2))
          (*(/B2            G1)
            (/G1          0.74798)))))
```

To run the GP, type the command line : Matrix –f input.file (matrix is the name of the executable, input.file is the name of the parameter file.)

The content of the input.file is

```
Pop_size=30
Max_generations=25
Random_seed=1234567890
Output.basename=matrix
Output.detail=80
Output.stt_interval=1

#how to generate the initial population
Init.method=half_and_half
Init.depth=2-6

#limits on the tree size
Max_depth=5
```

```
##breeding parameters
Breed_phases=2
Breed[1].operator=crossover, select=fitness
Breed[1].rate=0.9

Breed[2].operator=reproduction, select=fitness
Breed[2].rate=0.1

#
#multiple population
#
#multiple.subpops=8
#multiple.exch_gen=2
#==================================
#    ring topology one way
Exch[1].from=1
Exch[1].fromselect=best
Exch[1].to=2
Exch[1].toselect=worst
Exch[1].count=3

Exch[2].from=2
Exch[2].fromselect=best
Exch[2].to=3
Exch[2].toselect=worst
Exch[2].count=3

Exch[3].from=3
Exch[3].fromselect=best
Exch[3].to=4
Exch[3].toselect=worst
Exch[3].count=3

Exch[4].from=4
Exch[4].fromselect=best
Exch[4].to=5
Exch[4].toselect=worst
Exch[4].count=3

Exch[5].from=5
Exch[5].fromselect=best
Exch[5].to=6
Exch[5].toselect=worst
Exch[5].count=3

Exch[6].from=6
Exch[6].fromselect=best
Exch[6].to=7
Exch[6].toselect=worst
Exch[6].count=3

Exch[7].from=7
Exch[7].fromselect=best
Exch[7].to=8
Exch[7].toselect=worst
Exch[7].count=3

Exch[8].from=8
Exch[8].fromselect=best
Exch[8].to=1
Exch[8].toselect=worst
Exch[8].count=3

Multiple.exchanges=8
```

And here is a segment of the actual GP run:

```
[lil-gp Genetic Programming System.
[Original version create at the Michigan State University.
[kernel version 1.1; September 1998.

Initialization:
    Parameter database.
    Ephemeral random constants.
    Generation spaces.
Building function set(s):
Set 0: * / + - r1 r2 b1 b2 g1 g2 R
Tree 0 uses function set 0.
Function set complete.
Seeding random number generator with 1234567890
Creating initial population(s):
    37 trees were generated to fill the population of 30 (30 trees).
    Initial population(s) complete.
    Tree=0.18249
    [<><><><><><><><><>] hits=0 fitness=0.000000
    Tree=(--0.90758    b2)
```

```
[<><><><><><><><><>] hits=0 fitness=3.000000
Tree=(-( / r2    g2(+ b1    g1))
 [<><><><><><><><><>] hits=1 fitness=7.000000
Tree=(+(-(/r1 b1)(/b2 r1))(*(+ -0.17417   r1) (*r2   r2)))
[<><><><><><><><><>] hits=0 fitness=1.000000
Tree=(* g1    b1)
[<><><><><><><><><>] hits=0 fitness=0.000000
Tree=r1
[<><><><><><><><><>] hits=0 fitness=0.000000
Tree=r2
[<><><><><><><><><>] hits=0 fitness=3.000000
Tree=(-(+(+(/ 0.06248   b2)(+ -0.67523   r1))(-(+ g2   r1)(-
-0.58993    r1)))(/(-(-   -0.19889    0.62967)(/   r1    b1))(-(+b1
g1)(+b1   r2))))
[<><><><><><><><><>] hits=0 fitness=4.000000
Tree=(/r1   g1)
[<><><><><><><><><>] hits=0 fitness=0.000000
Tree=(+r2   -0.04464)
[<><><><><><><><><>] hits=0 fitness=3.000000
Tree=b2
[<><><><><><><><><>] hits=0 fitness=5.000000
STATEMENT STATISTICS
--------memory--------
        Allocated : 28944
        Freed     :28931
        Not freed : 13
        Max allocated:28183
        Malloc'ed blocks: 251
        Realloc'ed blocks:9
        Free'ed blocks :248
------- time---
        Overall:     434s wall
        Evaluation: 429s wall
        Breeding: 0s wall
--------generation spaces-------
        Space 0 size   :200
        Space 1 size   :100
--------ephemeral random constants-------
        Used: 66
        Freed:66
        Allocated:1000
        Blocks:1
```

So I have to say that the beginning matrix found by GP based on the initial fitness function is not good enough since the fact that it does not perform well on new images.

To improve that, I try two things. One is to define a more strict fitness function which measures the distance between the matrix found by GP and the reference matrix to make them a closer match. The other is to limit the max size of the programs. As we know from the machine learning theory the most general solutions (programs here) are usually the shortest one.

After several times of the GP runs, I find it is still very difficult to find the target matrix and value of the hits was always 0, so I reduces the number of the pictures and defined a new function which is max(x,y).

```
Double Max(double x, double y)
{
  If (x>y)
    Return x;
  Else
    Return y;
}
```

The target matrix should also plot the news image as a decreasing line. But for the stricter fitness function, I still have some problem with it.

4 Conclusions

I find the fitness function is very important. At the beginning of the project I used a relative simple fitness function and the GP can found one matrix quickly, but the performance of the matrix was not well on the new image. When changing it to a more strict one, the fact became that it was not easy for GP to find the target matrix.

5 Future works

Image comparison is still a challenging domain nowadays. The matrix found by GP still needs an inspection. To improve the performance of the matrix, I could define a new function or terminal set to make the

GP more efficient. And I should also pay more attend to the fitness function since it is a very important concept through doing my project. I may define a more strict fitness function.

References

[1] Koza J R, Automatic Creation of Human-Competitive Pro--grams and Controuers by Means of Genetic Programming[J] .Genetic Programming and Evolvable Machines, 2000

[2] GOLDBERG DE, Genetic Algorithms in search, Optimization and Machine learning [M], Addison-Wesley ,1989

[3] Chambers,L, The practical handbook of Genetic Algorithms, chapman & Hall/CRC, 2000

[4] ZUO J, TANG CJ, ZHANG TQ, Mining Predicate Association Rule by Gene Expression Programming [J], Berling Heidelberg:Springer Ver2 lag,, Lecture Notes In Computer science, 2002, 2419: 92-103

[5] Koza J R, Genetic programming: On the programming of computers by means of natural selection, Gambrige, Mass.:MIT Press, 1992

[6] Genetic Algorithms Research and Application Group (GARAGe), Michigan State Universityhttp: //garage.cps.msu.edu/software/ lil-gp/lilgp-index.html

Multispectral Imaging for Fingerprint Detection Using Computer Projector

Qingzhi Feng[1] Haobo Li[2] Chunbing Zhou[1]

1 Department of Forensic Science and Technology, China Criminal Police College, Shenyang 110035, Liaoning, China

Email: zhouchunbing@china.com.cn

2 Industry Park District Branch, Suzhou Police Bureau Suzhou 215123, Jiangsu, China

Email: tenghaoyang@163.com

Abstract

The use of computer projectors as controlled light sources and digital cameras as image detectors (the so--called computer projector photo-assisted system, CPPS) is an ubiquitous alternative for the examination of indefinite fingerprints in forensic identification laboratories. The performance of CPPS for such examinations depends on several factors, from which the most relevant are the conformation of fingerprint substances and the composition of illuminating sequences. In this paper, with the aid of a chrominance model, the correlation of spectral fingerprints with illuminating colors is investigated, allowing one to determine the redundancy and limitation with respect to visible spectroscopy displayed by the computer projector. Difference spectral analysis and ratio spectral analysis are introduced and theirs properties compared with standard procedures.

Keywords: Multispectral imaging; Chrominance model, Fingerprint detection; CPPS

1 Introduction

Naturally, the image is a kind of visual substantiality to express the object, which can describe numerous characteristics including spatial vectors, temporal vectors, spectral vectors and physical properties of the object. In the case of both spatial resolution and temporal resolution can be assured, it is inevitable that the spectral imaging is presented and developed. The technique of multispectral imaging is employed to capture and process the several waveband image information of a certain object, afterwards, a synthesized image can be obtained by means of multispectral image fusion, and more characteristics of the object can also be extracted from the synthesized image. Consequently，it is broadly applied to such research fields as the space remote sensing, precision farming, geography investigating, biology medicine and forensic identification etc. In fact, a multispectral imaging system is usually equipped with more than 50 spectral channels (spectral resolution $\Delta\lambda/\lambda = 0.1$ in each channel), which is composed of the computer, photoelectrical detector and dispersive device. Fig. 1 shows the principle of a multispectral imaging system applied to the fingerprint detection. The computer controls the dispersive device to change spectral channels sequentially so that fingerprints can be illumined by different waveband light, in the meantime, the photoelectrical detector captures relevant waveband images of fingerprints and transfers them to the computer.[1]

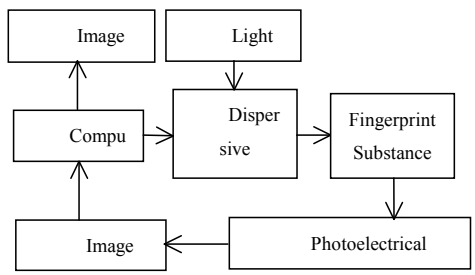


Figure1 A multispectral imaging system for fingerprint detection

2 Computer projector photo-assisted system

As an important component in the multispectral imaging system, the dispersive device was developed from the dispersive prism to the diffraction raster. At present, acousto-optic tunable filter and liquid crystal tunable filter were successively produced and utilized. Sometimes, in order to obtain different waveband light, the computer projector is selected as the substitute of the dispersive device. CPPS utilizes the casual combination of familiar devices such as regular computer sets and digital cameras for multispectral imaging of objects.[2,3] The ubiquity of these elements makes it an attractive alternative for the examination and identification of fingerprints in forensic identification laboratories throughout the country.

CPPS is compatible with numerous sensing principles such as cell viability test, fluorescent indication, colorimetric assay, chemical evaluation, physical evidence photography, etc. In a CPPS measurement, the examined object is illuminated by a color sequence displayed by the computer projector, while the image of the examined object is recorded by the digital camera in synchronism with the illumination. The result is a image stream from which spectral signals from selected regions of interest can be extracted.

The effect of CPPS recording fingerprints depends on how the spectral features of examined fingerprints are evaluated and compensated. The way that these CPPS signals are composed has a main impact on the ability to distinguish examined fingerprints and consequently to identify them. The amount of colors used for illumination and the selection of these colors in relation to the examined fingerprint sets are also relevant. All these aspects have been acknowledged in a number of studies; however, a deeper understanding of the limitations and possibilities of this method requires a detailed analysis of the operating conditions and the way that the CPPS signals are composed. In this paper, the performance and advantage of CPPS on the fingerprint identification is investigated with the aid of a chrominance model. The correlation of illuminating colors with the signal features is established, allowing one to assess the virtues and weaknesses of the different alternatives in the fingerprint photography. The difference spectral analysis and ratio spectral analysis are introduced and compared with standard procedures and with the spectroscopic performances for different operating conditions.

3 Chrominance model

A computer projector is a light source composed of multiple single elements with specific emission distributions that constitute a collective source. Also multiple elements are integrated of liquid crystal display units made based on the interference of polarization light. When rays coming from the light source transmit the liquid crystal display unit, rays will be decomposed into two bundles of polarized light with the same oscillation direction and the definite phase difference. These two bundles of light will cause the interference of light, and form the new light of which wavelength is determined by the phase difference. By adjusting the voltage charged on the liquid crystal display unit, the different waveband light will be obtained. Therefore one can conveniently implement multispectral image scanning by computer programming.

In this context, a color c_i displayed by the computer projector, and specified by the triplet (r_i, g_i, b_i) that indicates the modulation of the projector primary spectral radiances $[R]$, $[G]$ and $[B]$ produces an intensity in the digital camera that can be written as:

$$\left.\begin{aligned} Ir_i &= \int_{\lambda} c_i(\lambda)\cdot F_r(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)\, d\lambda \\ Ig_i &= \int_{\lambda} c_i(\lambda)\cdot F_g(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)\, d\lambda \\ Ib_i &= \int_{\lambda} c_i(\lambda)\cdot F_b(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)\, d\lambda \end{aligned}\right\} \qquad (1)$$

Where $F_r(\lambda)$, $F_g(\lambda)$ and $F_b(\lambda)$ are the filters of the digital camera, $\rho(\lambda)$ is the reflectivity of the examined substance and $D(\lambda)$ is the spectral response of the camera detector. Also the color c_i can be written

as the linear combination of the projector primaries:

$$c_i(\lambda) = r_i \cdot [R] + g_i \cdot [G] + b_i \cdot [B] \qquad (2)$$

Where r_i, g_i and b_i are numbers within [0, 1]. Introducing eqn (2) to eqn (1) and rearranging:

$$\left.\begin{aligned} Ir_i &= r_i \int_\lambda [R]\cdot F_r(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)d\lambda + g_i \int_\lambda [G]\cdot F_r(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)d\lambda \\ &\quad + b_i \int_\lambda [B]\cdot F_r(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)d\lambda \\ Ig_i &= r_i \int_\lambda [R]\cdot F_g(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)d\lambda + g_i \int_\lambda [G]\cdot F_g(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)d\lambda \\ &\quad + b_i \int_\lambda [B]\cdot F_g(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)d\lambda \\ Ib_i &= r_i \int_\lambda [R]\cdot F_b(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)d\lambda + g_i \int_\lambda [G]\cdot F_b(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)d\lambda \\ &\quad + b_i \int_\lambda [B]\cdot F_b(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)d\lambda \end{aligned}\right\} \qquad (3)$$

$$\begin{bmatrix} Ir_i \\ Ig_i \\ Ib_i \end{bmatrix} = \begin{bmatrix} r_R & r_G & r_B \\ g_R & g_G & g_B \\ b_R & b_G & b_B \end{bmatrix} \times \begin{bmatrix} r_i \\ g_i \\ b_i \end{bmatrix} = S \times \begin{bmatrix} r_i \\ g_i \\ b_i \end{bmatrix} \qquad (4)$$

or briefly

Where the terms of the substance matrix S are functions of $\rho(\lambda)$, and retain up to nine spectral windows of the substance reflectivity, captured through these combinations of projector radiances and camera filters. [4,5]

The intensities measured in this way contain the spectral information that would allow spectral reconstruction, but also the spectral characteristics of the platform on which the examined fingerprint embedded. So far, two main approaches have been followed for highlighting the substance features: difference spectral analysis and ratio spectral analysis depending on how a reference measurement for a sample of the platform with $\rho^0(\lambda)$ is used. For example, the spectral signal from the examined fingerprint is highlighted by subtracting the spectral signal from the platform in the difference spectral analysis.

Eqn (5) and (6) describes this procedure.

$$\begin{bmatrix} \Delta Ir_i \\ \Delta Ig_i \\ \Delta Ib_i \end{bmatrix} = \begin{bmatrix} Ir_i - Ir_i^0 \\ Ig_i - Ig_i^0 \\ Ib_i - Ib_i^0 \end{bmatrix} = \begin{bmatrix} r_R - r_R^0 & r_G - r_G^0 & r_B - r_B^0 \\ g_R - g_R^0 & g_G - g_G^0 & g_B - g_B^0 \\ b_R - b_R^0 & b_G - b_G^0 & b_B - b_B^0 \end{bmatrix} \times \begin{bmatrix} r_i \\ g_i \\ b_i \end{bmatrix} = \Delta S \times \begin{bmatrix} r_i \\ g_i \\ b_i \end{bmatrix} \qquad (5)$$

Where the elements of ΔS are differences in intensities, e.g.

$$\begin{aligned} \Delta S_{1,1} &= \int_\lambda [R]\cdot F_r(\lambda)\cdot D(\lambda)\cdot \rho(\lambda)d\lambda - \int_\lambda [R]\cdot F_r(\lambda)\cdot D(\lambda)\cdot \rho^0(\lambda)d\lambda \\ &= \int_\lambda [R]\cdot F_r(\lambda)\cdot D(\lambda)\cdot [\rho(\lambda) - \rho^0(\lambda)]d\lambda \end{aligned} \qquad (6)$$

Although imperfect, these two methods require less illuminating colors to evaluate approximately spectral characteristics of the examined fingerprint, and provide a clear distinction between predominant emission (positive values of ΔS) and absorption (negative values of ΔS).

Considering a three primary color illuminating sequence that maps the terms of S in the intensity profile, the matrix S can be calculated as below:

$$\begin{bmatrix} Ir_1 \\ Ig_1 \\ Ib_1 \\ Ir_2 \\ Ig_2 \\ Ib_2 \\ Ir_3 \\ Ig_3 \\ Ib_3 \end{bmatrix} = \begin{bmatrix} r_R & r_G & r_B & 0 & 0 & 0 & 0 & 0 & 0 \\ g_R & g_G & g_B & 0 & 0 & 0 & 0 & 0 & 0 \\ b_R & b_G & b_B & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & r_R & r_G & r_B & 0 & 0 & 0 \\ 0 & 0 & 0 & g_R & g_G & g_B & 0 & 0 & 0 \\ 0 & 0 & 0 & b_R & b_G & b_B & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & r_R & r_G & r_B \\ 0 & 0 & 0 & 0 & 0 & 0 & g_R & g_G & g_B \\ 0 & 0 & 0 & 0 & 0 & 0 & b_R & b_G & b_B \end{bmatrix} \times \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} r_R \\ g_R \\ b_R \\ r_G \\ g_G \\ b_G \\ r_B \\ g_B \\ b_B \end{bmatrix} \qquad (7)$$

In the experiments of the physical evidence photography, the intensities from the three camera channels for a particular illumination i are composed as:

Where r_i, g_i and b_i can be selected to obtain the particular illumination i by computer programming. [6]

$$\begin{aligned} I_i = Ir_i + Ig_i + Ib_i = r_i(r_R + g_R + b_R) + \\ g_i(r_G + g_G + b_G) + b_i(r_B + g_B + b_B) \end{aligned} \qquad (8)$$

4 Application on fingerprint detection

As is known to all, the fingerprint detection is an important issue in course of the criminal scene investigation. To evaluate accurately the classification of examined fingerprints is the basis to optimize the fingerprint detection. Typical CPPS fingerprint detection strategies, their particular characteristics and their effects on the fingerprint classification properties have been analyzed with the aid of a chrominance model. From this analysis, it can be concluded that for well designed light-filter sets, CPPS can provide the spectrally richest fingerprints with the shortest illuminating sequence, whereas difference spectral analysis can be complementarily used to enhance fingerprint images. In the case of difference spectral analysis, the same illuminating sequence produces three times more intensities,

determined by the combination of illuminating radiances and camera filters. Furthermore, a subset of relevant projector-camera configurations can be associated with individual fingerprints, allowing a more detailed customization of measuring conditions, and an optimized management of the available information by dismissing non-significant variables. This pattern is partially allowed by the composition of the considered set of fingerprints that enable one to exploit all the possible projector-camera combinations. The result shows the effect of CPPS detecting fingerprints for a particular illumination composed of the three projector primaries. The different magnitude of the spectral windows can be calculated, since the information is concentrated along the terms corresponding to the main diagonal of S, which formed by blue light—blue filter, green light—green filter, etc. For this reason, blue light—blue filter and green light—green filter are effective strategies employed on multispectral imaging for fingerprint detection.

Since CPPS is a casual assembly of components rather than a strictly defined instrument, the fingerprint constitution becomes vital to improve the robustness of the fingerprint detection. A commonly used approach in CPPS is to illuminate with sequences more than three colors, which in principle contain redundant colors. By increasing the number of illuminating colors to 7, the added characteristics of the examined fingerprint rather increase. Although formed by a combination of projector primaries, the considered color sequence can retain, up to certain extent, part of the 9 spectral bands. It is also illustrative to corroborate that not any color sequence contributes additional discrimination. Using a nine color illumination formed by the original three primary illuminations, the achieved performances are the same as with the three color illumination. Beyond the fact that repeated sequences are valuable for canceling random noise, the present result also indicates that the length of the illuminating sequence does not guarantee a better performance of the evaluation by itself. On the other hand, sensibly longer sequences do not harm, and offer a valuable degree of freedom in operating scenarios where the optimum sequence for a particular target set of fingerprints is ignored. It must also be noticed that the present model does not consider the noisy illuminating, which also implies a certain distribution of optimum conditions that can be supplied by a longer collection of slightly different illuminations.

5 Conclusions

In general, the constitution of the examined fingerprint might involve any collection of color substances, and frequently with only subtle color differences, for which the larger the number of considered spectral terms, the more robust the result and the greater the flexibility to choose the best operating conditions. Interestingly, white light is one of the less significant contributions, due to its completely unspecific characteristics. White light is always absorbed by any substance, leaving behind just small modulations compared with a specific illuminating radiance maximally absorbed by a target substance. Summarizing, Difference spectral analysis and ratio spectral analysis can be complementary exploited. Difference spectral analysis can provide the best performance for fingerprint detection with substantial differences, requiring shorter illuminating sequences. On the other hand, the ratio spectral analysis allows better fingerprint disentanglement from its platform substances, which becomes a dominating feature for the examination of indefinite fingerprints.

References

[1] Exline D L, Wallace C, Roux C, et al. Forensic application of chemical imaging: Latent fingerprint detection using visible absorption and luminescence [J]. Forensic Sci. 2003, 48(5): 1047-1053

[2] Hardeberg J Y, Schmitt F, Brettel H. Multispectral image capture using a tunable filter[J]. Proc. SPIE, Bellingham, WA, 2000, 3963

[3] Connah D, et al. Recovering spectral information using digital camera systems[J]. Color Technology, 2001, 117:309

[4] S. Westland and C. Ripamonti. Computational colour science, Willey, 2004, ch. 10

[5] Maloney L T. Evaluation of linear models of surface spectral reflectance with small numbers of parameters[J]. Journal of the Optical Society of America, 1986, 3:673-1683

[6] R. Johnson and D. Wichern. Applied multivariate statistical analysis, Pearson Education Ltd., 2002

Look into Speculation Behavior in Real Estate Market through Cryptic Cost

Ruichao Du[1] Shujun Ye[2]

School of Economic and Business, Beijing Jiao tong University, Beijing, P.R.China

Email:1 durc006@163.com; 2 shjye@bjtu.edu.cn

Abstract

The paper makes a study on the effects on government housing policies which was expected to arrest speculation behavior in excess, such as raise the rate of capital gains tax. The result shows that the effects are inconspicuous because of cryptic price in real estate market and indirection of tax, making the additive capital gains tax become a part of price, as cryptic cost, the customers absorb the additive cost instead of speculators at last, what's more, it will generate an erroneous signal that the expected price is going up. The concept of new tax which is named purchase tax was given purposing in turning cryptic additive tax to surface, in order to arresting speculative demand.

Keywords：Speculative demand; cryptic cost; purchase tax; real estate market; government housing policies

1 Introducing

The real estate business is pool of funds, over centralization of capital will lead to speculative demand. The project which was named "Shenhaimingyuan" in the center of shanghai, opening price was 6000yuan/sq.m in 2001, to everyone's astonishment, the price had been going up to 24000yuan/sq.m in 2005,The price was three times as much as before in three years. As a parallel case, the increment speed on real estate price got head of CPI and household consumption ability:

2 The subject of price rose

Basic on the study of price growth in real estate market, the reasonable growth result from the contradiction that the accommodation decided by scarcity of real estate and absolute advance in social needs. In the period, raising demand attracts accession and price rose. To speak up, reasonable price rose will be priming of economy, nonetheless, geographical, imbalance of supply and demand, appreciated value and scarcity of land decide the real estate market possess speculation. What's more, the price rose intensifies speculative motive, attracting more speculators enter into the real estate market, creating the increment speed on real estate price get ahead of CPI and household consumption ability.

Table 1 The real estate price and growth rate of CPI

Year	2004	2005	2006	2007
The real estate price	8000	8952	10473	12044
Growth rate of real estate price(%)		11.90	17.00	15.00
Growth rate of CPI(%)	2.40	1.60	2.80	6.50

By the way, cryptic price in real estate market is also an important subject. The real estate price is made up of land cost, architectural engineering cost, public facility cost, tax and others. As a rule, the customers can't know about the composition of the price, just analyze price, creating blind investment. As a seller's market, real estate market attracts more speculators and picks up movement rate. High price attracts the speculative demand instead of the consumer demand because of expectation of price rose.

3 The control law in real estate market

In the 20th,the real estate price was 3 times as much as before in South Korea because of speculators. The South Korea's government is based on the premise that satisfied the lodging demand of low-income consumers took measures as foot-in-the-door effect and heavy duty to crack down the speculations.

Aim at continuous rising price, the government in China promulgates a series of decrees such as raising the rate of land tax.

4 Mechanis of purchase tax

The market is made up of need market and supply market. The measures which government took were all aimed at cost of sellers, looking at the picture1 as followed, the real estate market is over demand, the gap is AB, the speculation behavior stretches price rising up to P_0, the additive capital gains tax means additive cost, but customers absorb the additive cost instead of speculators because of cryptic price and indirection of tax. What's more, it will jack up the price from P_0 to P_2. Now take another thread, demand expected reduction will make demand from D_0 to D_1, creating a drop in price from P_0 to P_1.

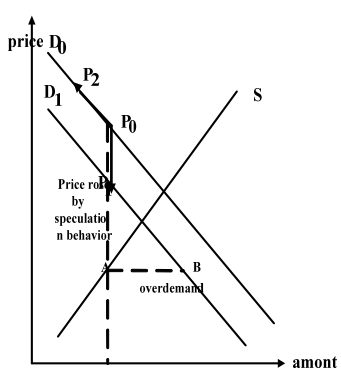


Figure1 supply and demand in real market

Take raising the rate of capital gains tax for example. The real estate price would go up from P1 to P2 if taking the measures. The part between two curves demonstrates the additive tax.

In a few words, set $P_1=(aX^2+c)$,then $P_2=(aX^2+c)(1+\theta)$, x as the anticipation, a as cost parameter, θ as rate of capital gains tax, the capital tax is $P_3=\theta(aX^2+c)$.

The subject of taxation of capital gains tax is sales income, but by the reason of indirection of tax, the additive tax becomes a part of cryptic cost to raise the real estate price. The carriers will be customers at last, what's more, the carriers generate an erroneous signal that the expected price is going up, causing more speculative motive, the speculators sack an enormous profit without any loss.

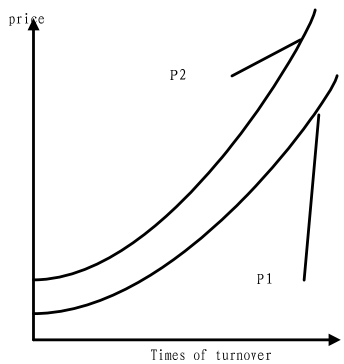


Figure 2 the change of price after raise the rate of capital gains tax

5 Conclusions

The actual state in real estate market is that the measures as impose tax to bate earning made parasitically no impression on speculators, on the contrary, it caused price rose and more speculation behavior, forming vicious cycle. Point to the actual state as indicated above, the concept of new tax which is named purchase tax was given, there will be no or less value variance before and the carriers of new purchase tax is also customers at last. But the new purchase tax purposes in turning cryptic additive tax to surface, in order to arresting speculative demand, as Fig.1 indicated above, requirement goes up form P0 to P1, creating a drop in price to P2, the speculation behavior would be arrested learn by experience.

The rate of purchase tax as a off-price tax commensurate to the number of door, in order to arrest speculative demand, achieve soft landing of real estate price, fulfill direct custody. The purchase tax whose

subject is cost of more than 2 houses is not equal to added value tax, according with goal of arresting speculation behavior and protecting normal investment. Considering the transaction cost in China is 5% of price and the experience in South Korea, set rate of purchase tax as followed:

Table2 Rate of purchase tax

Number of houses	0	1	2	3	＞3
Rate of purchase tax(%)	0	10	30	50	70

The speculation demand will be partly arrested by establishment of purchase tax, the requirement measures as followed:

- The statistic is “door”, containing parents and celibate children, married child is another door;
- Ameliorate purchasing systems, the applicant should resort personal ID card and register of household;
- Ameliorate selling systems, the decorate should not listing without the blessing of government.

References

[1] J·E·Mcnulty.Overbuilding,Real-EstateLeading Decisions, and the Regional Economics Base[M].Journal of Real Estate Financeand Economics ,1995

[2] D·Dipasqua1.Why Don t We Know More aboutHousing Supply.Joumal of Real Estate Finance and Economics,1999

[3] D·Denise,C·W·Willian.Urban Economics and Real Estate Markets[M].Prentice-Hall,1996

[4] Guo Xin ru.Analysis on China's Real Estate Bubble[J].ScienceTechnologyandIndustry,V01.7,NO.5,October 2007,PP45~46

[5] Gu Zhi ming, ZHAO Hai-fen. Containing Speculative Purchase of House: Necessity and Strategy[J]. J.of Wuhan Uni.of Sci.&Tech. (Social Science Edition, Vo1.8,No.4,May 2006,PP24~24

[6] ZonghuaBao. Revelation though South Korea.Focus, Vol.1,No.12,June 2005, pp.71~86

Recognition of Notice Marking before Pedestrian Crossing

Ning Zhang[1] Tiejun HE[2] Zhaohui Gao[3] Hui Chen[4]

ITS Research Center of Ministry of Education, Southeast University Nanjing, Jiangsu, 210096, China

Email:1 ningzhang1972@yahoo.com.cn; 2 hetiejun56826@yahoo.com.cn;

3 zhaohuigao2008@gmail.com; 4 terry0363202@126.com

Abstract

An approach to recognize the notice marking before pedestrian crossing with the shape of white rhombus is proposed. The target was extracted through color segmentation, and Hough transformation was utilized to detect the edge lines of the target. Harris corner detection algorithm was further used to acquire the two corners with the maximum change of target angle, by which to examine whether or not lines detected by Hough transformation are edge lines of the target and to eliminate the pseudo edge lines. At last, the intersectant points of the detected edge lines were utilized to judge whether the target is a rhombus. 20 images including rhombus notice marking in the scenes and another 20 images having no rhombus were tested. Only two images were detected in errors and the detection rate reaches 90 percents. Non-rhombus images can be judged accurately in the scene that there exist no rhombus targets. The results show that the method is feasible and can be used in traffic signs recognition.

Keywords：Notice marking before pedestrian crossing; Rhombus detection; Hough transformation; Harris corner detection

1 Introduction

Traffic signs and road marking are critical components of urban transportation systems, which transfer information using symbols and letters to guarantee safe and efficiency of the traffic in cities. With the wide applications of computer vision, traffic signs recognition (TSR), which is an important field of intelligent transportation system (ITS), has already received great attentions in some research institutes all over the world. TSR is an important research topic, in which image processing methods are utilized to automatically detect and recognize traffic signs and markings in the vehicle-based vision systems. It can provide real time road conditions to drivers, help them conduct vehicles and improve the safety of driving [1-2].

Most of the researches on TSR were focused on the detection and identification of cycles, triangles, lanes markings [3-5]. In recent years, some changes have taken place on the road traffic markings. For instance, a white rhombus road marking, which is called notice marking, is added 30-50 meters before pedestrian crossing on urban roads. However, we do not find reports or papers about using image processing to detect it. This paper does address the detection and recognition of the notice marking with the shape of white rhombus, which can remind drivers in time that they will arrive at pedestrian crossings and have enough time to decrease the travel speed, and improve the traffic safety of vehicles and pedestrians.

Notice marking before pedestrian crossing is generally close to the road crossings in the city, and its logo is rhombus with the white lines, covering an area of about one square meter. By extracting white color the target can be obtained from the road scene and then Hough transform and Harris corner detection were used respectively to get the two characteristics of the target: edge lines and the two corners with the maximum angle changing. And then utilized the two characteristics to determine whether the target is rhombus or not.

2 Location of notice marking before pedestrian crossing

In this paper, data was acquired by erecting a camera in the vehicle, and the target was located from images captured by the camera. Figure 1 shown as following is a frame of images collected by the camera displaying the road scene with the white rhombus (640×480 pixels). The shape of the target, extracted from the camera image, has great relationship with the angle of the camera and the interesting area of the image. A rhombic target in 3D world is mapped into planar image by camera. According to the principle of perspective, the appearance of the rhombic will change in shape [6]. As a result, we have to predetermine the interesting target area and then recognize the rhombus in that area, which can minimize the inaccuracy and promote the ratio of detection. In this experiment, we set the target area in the center of the picture which is 100 pixels offset between the up and down border. First, the image of the scene was binarized. Considering the target is white rhombus, a threshold value of T (the value is 190 in this case) was set in all color subspace of R, G, B. Only in the case that all the pixel values are larger than T, then the current pixel value was set to 1, or else 0. The result is shown in figure 2.

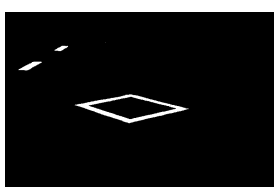


Figure1 Image collection Figure2 Extraction of white target

3 Feature extraction of Notice marking before pedestrian crossing

After extracting the white road marking target, the next step is feature extraction. The main feature of Notice marking before pedestrian crossing is edges and corners, so we utilize Hough transform to detect target edge lines and Harris corner detection to detect target corners.

3.1 Straight line detection with Hough Transform

Hough transform is widely used in the detection of straight line. By transforming spatial domain of the image into parameter space and describing the curves in the image with points that controlled by special parameters, then useful information could be calculated with statistics means. Hough transform is also a useful method in disadvantageous surroundings such as the condition of noise, shape distortion or shape impairment. As a result, it is widely used in actual engineering. Considering the case when the slope is infinite, detection of straight lines in actual use often employ parameter equation as below [7]:

$$\rho = x\cos(\theta) + y\sin(\theta) \qquad (1)$$

Where ρ is the distance between the straight line to origin of coordinate, and θ is the angle between X axis and normal.

In this way, any point in the image can be mapped into a curve in parameter $\rho-\theta$ space by the Hough transform. It is easy to conclude that the curves mapping by the points in a line is sure to cross in a special point in $\rho-\theta$ space. By summing up the points of the cumulative value in the parameter space, the points with a maximum accumulated amount are mapped to the parameter of the straight lines.

3.2 Harris corner detection principle

Harris operator is proposed by C.Harris and J.Stephens which is used to extract the feature of corner point based on the point feature of the signal [8]. The algorithm is easy to calculate and the corners detected by Harris are well-distributed, as well as the feature points can be extracted quantitatively. The proceeding is list as followed:

$$E(x,y) = \sum_{u,v} w(u,v)[I(x+u,x+v) - I(x,y)]^2 = [u,v]M\begin{bmatrix} u \\ v \end{bmatrix} \qquad (2)$$

$$M = \sum_{u,v} w(u,v)\begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} \qquad (3)$$

Formula (2) and (3) are the pixel correlation functions of Harris corner detection, where, $I(x,y)$ is the grey value of the image; I_x and I_y are the grads in two different direction of every pixel which are calculated by difference operator; w is Gaussian noise smoothing window, and (u,v) is coordinate migration. If the two eigenvalues of M matrix have the local maximum value, the point is accepted as feature point. In order to avoid the complicated calculation of eigenvalues, the assessing function is adopted like the form below:

$$R = \det(M) - k \times tr^2(M)3 \qquad (4)$$

Where, tr is the trace of matrix, and det is the determinant value of matrix, k is 0.04-0.06 taken by experience. When R is positive, it is taken as corner. Generally a threshold of T is set, when R > T, it could be detected as corner. Obviously, less feature points could be recognized when T is set larger.

4 Recognition algorithm of notice marking before pedestrian crossing

At present, rhombus target identification had not a precedent that we can draw comparison. After extracting the road marking on the second part, and utilizing Hough transform and Harris corner detection to obtain the edge lines and corners characteristics, this paper propose an algorithm of notice marking before pedestrian crossing in accordance with the general geometry feature extraction method. The steps are as followed:

1) Target Location: Target location is to obtain the target and exclude most things that are not interested. Because the rhombus target is white, this feature is different from most scenery in the scene. As a result, the initial target can be extracted with color. The methods have already mentioned in the second part.

2) Target edge detection with Hough transform: In order to reduce the number of edge, the target skeleton extraction is carried out, and then detected with Hough transform. According to Hough transform, the intersection of curves in parameters domain is corresponding to the line in spatial domain.

3) Target corner detection: The aim of Hough transform is extracting the edge of target, and the Harris corner detection is used to extract the corner point. In the process of Harris corner detection, the corner response function of non-maxima restrain is carried out, and the number of the corner extraction depends on the threshold.

4) Removal of pseudo-edge: As the edge of target detected by Hough transform is not always the edge of the rhombus, so Harris corner detection is used to test the distance between the corner points of the target and edge line to remove the pseudo-edge.

5) Rhombus discrimination: According to the rhombus property, adjacent edges are equal and opposite edges are parallel. After removing the pseudo-edges, judge the lines whether accord with the property of the rhombus, and then do a judgment.

For the actual scene of rhombus road marking, as the weather or time change cause the change of light, the acquisition of image brightness is affected. Because the road background color is humdrum, extracting target through color segmentation is robust.

5 ExperimentS and analysis

The algorithm for recognition of a white rhombus road marking was simulated by computer. As shown in figure 2, the white target was extracted through color segmentation, and then the target skeleton was extracted in order to reduce the number of edge, which shown as figure 3. Figure 4 is the result of Hough transform, and 5 lines were detected in this experiment, as shown in figure 5. So excluding the pseudo-edge was still necessary. Use Harris corner detection to extract the two corners with the maximum angle changing of the target by a larger threshold, as shown by "+" symbol in figure 6. Calculate the distance d_i from the corner point (x_i, y_i), which was detected through Harris operator, to the edge lines $A_i x + B_i y + C_i = 0$. If $d_i < \varepsilon$ (ε equal to length of two pixels), then the edge line is the goal line of the target. At last judging the lines whether accord with the property of the rhombus, and then conclude the detection target is a

rhombus or not.

This paper has two groups of experimental data to test the algorithm, a group data of 20 images including rhombus road marking in the scenes, and the other group without rhombus road marking. The experimental results are shown in table 1. According to the first group data, only two images were detected with errors and the detection rate reaches 90 percents. Non-rhombus images can be accurately judged.

Table.1 Experimental Results

Scene image (20 frames)	Numbers of correctly detected	Numbers ofmistakes	Detection ratio
Image include rhombus	18	2	90%
Image without rhombus	20	0	100%

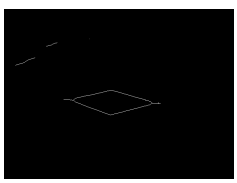


Figure3 Extraction of target skeleton

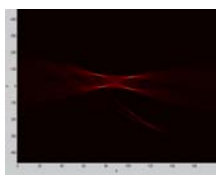


Figure4 Hough transform

Figure5 Target edge lines detection

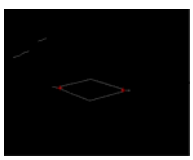


Figure6 Harris corner detection

6 Conclusion

This paper proposes an approach of recognizing the notice marking before pedestrian crossing with the shape of rhombus by using the color segmentation, edge detection and corner extraction. Firstly use R, G, B color channel intergraded in RGB color space to extract the white target in the image areas. Secondly detect the target edge with Hough transform, extract the two corners with the maximum change of target angle, and then remove the pseudo-edge according to the relation of edges and corner points. Finally judge the lines whether they are accord with the property of the rhombus, and then conclude if the detection target is a rhombus or not.

Although people had come up with some intelligent algorithms in the fields of traffic signs recognition, they were limited to a certain specific geometric shape. The algorithm mentioned above is about recognition of notice marking before pedestrian crossing, and has gained a good effect on the bases of experiments. At the same time the geometric shape of most traffic signs, such as triangle, rectangle and so on, are with characteristics of corners and edges. So the means of corner and edge extraction mentioned in this paper can also use to recognize other traffic signs.

References

[1] Wang Rong-ben, Zhao Yi-bing, Li Lin-hui, “Approach Review of obstacle detection for intelligent vehicle”, Journal of Highway and Transportation Research and Development. Vol.24, No.11, November 2007, pp.109~113

[2] Zuo Xiao-qing, Li Qing-quan, Xie Zhi-ying, “Lane-based road data model”, Journal of Chang’an University: Natural Science Edition, Vol.24, No.2, march 2004, pp.73~76

[3] J.C. McCall, O. Achler, M.M. Trivedi, “Design of an instrumented vehicle test bed for developing a human centered driver support system”, IEEE Intelligent Vehicles Symposium, Parma. Italy, pp. 483-488, Jun 2004

[4] F. Moutarde, A. Bargeton, A. Herbin, L. Chanussot, “Robust on-vehicle Real-time Visual Detection of American and European Speed Limit Signs, With a Modular Traffic Signs Recognition System”. IEEE Intelligent Vehicles Symposium, Istanbul, pp.1122-1126, June 2007

[5] Xu You-chun, Wang Rong-ben, Li Ke-qiang. “A Linear Model Based Road Identification Algorithm”, Journal of Imageand Graphics, Vol.9, No.7, July 2004, pp.858~864

[6] D. Murray, J.J. Little, “Using Real-Time Stereo Vision for Mobile Robot Navigation”, Autonomous Robots, Vol.8, No.2, April 2000, pp.161~171

[7] Hao Ting, Meng Zheng-da. “Torch Recognition of Robot in Complex Environment”, Journal of southeast university:Natural Science Edition, Vol.35, No.s2, November 2005, pp.151~154

[8] C. Harris, M. Stephens. “A Combined Corner and Edge Detector”, Proceeding of the 4th ALVEY Vision Conference, Manchester, pp.147-151, Augest 1988.

Jointing Images by Digital Image Processing Techniques for Crime Scene Investigation

Dan Liu[1] Yu Huang[2] Chunbing Zhou[1] Min Gao[2]

1 Department of Forensic Science & Technology, China Criminal Police University, Shenyang 110035, Liaoning, China

2 The public Security Bureau of Longquan District, Chengdu 610100, Sichuan, China

Email: dliu@ccpc.edu.cn, huangyu_leo@yahoo.com.cn, zhouchunbing@hotmail.com

Abstract

With the wide use of the digital equipments, such as high pixel digital camera, high resolution scanner, etc., the digital images are being gradually used in almost every aspects of people's life. Due to its unique features, along with the development and application of computer equipments and software, the digital image processing technique is growing up to a new subject, applied to practice. And together with the traditional methods, their own particular advantages are exploited for mutual benefit and enhancement. Image mosaic method is one of the image-based rendering methods; it is based on the ability to align different views of a scene (overlapped) into a large image and then to seamlessly blend the image together. A method of jointing images to make panoramas for crime scene investigation is proposed. It can effectively reduce the possibility of mismatch; it does not have intensive computing involved. It works well for almost all pictures with common characteristics.

Keywords: image-based rendering; mosaic of multi-viewpoint image; virtual environment; crime scene investigation

1 Introduction

Traditionally, images are made by taking analog pictures with sensitive films, after that processing the negative film, and then developing them, or directly developing the digital images. Panoramas must be taken in a comparatively long distance and at high places, which always has its limitations because of the objective conditions, such as topography. This method may lose more detail features of the target objects, which makes it failure to get the original intentions of photographing [1].

For crime scene investigation, the panorama photo-taking method is consecutively to photograph a plurality of scenes, either in a horizontal or vertical direction. The developed images are jointed by hand. This kind of method has fairly higher requirements of photographing skills, such as the depth of field, exposure value, the setting of equipments, such as the fixed point, aperture and shutter and objective conditions, such as time and light. The most important thing is that the subjective factors play a main role during the image jointing process, which makes the result is not objective and the clear seams between the neighbor images can be seen.

Compared with the traditional image-joining method, the digital image processing technique has its own advantages and characteristics [2, 3].

2 The comparison between the digital image mosaic system and the conventional jointing method

Using the digital image mosaic system to joint images has higher objectivity and precision. The means of image adjustment in the digital image mosaic system is more different from the traditional ones. The

* This paper is supported by Fok Ying-tong Education Foundation (91077).

individual images can be pre-processed before they are to be jointed together by computer. Many partial or entire modifications can be done to get the best effect. This process is different from that of adjusting the camera's physical parameters to achieve the same effect. The entire adjustments can be done to the jointed image in order to maintain a relative integrity among the physical parameters of the picture, such as hue, lightness and saturation.

Images have experienced both the partial adjustments in the early period and the entire adjustments in the later period.The digital image mosaic system can preserve the detail features of the original images. Traditionally, a panoramic image is made of single ones requiring long shooting distance, which may lose more details of the original objects or scenes, thus the objects being small and unclear. The traditional jointing method is to digitalize the developed images, and then to eliminate the noises in them by computers. During this process, dozens of transformations from analog images to digital ones or from digital images to analog ones are needed, so it may fail to preserve the original features and details. The digital image processing technique can preserve more details of the original objects because it minimizes the possibility of data losing during the mode shifts [4, 5, 6, and 7].

3 The digital image mosaic technique

1) Photograph consecutively a set of images, either in a horizontal or vertical direction; The shooting quality and mosaic effect are very important for the 3D scene display. The shooting quality of a real world picture requires the right location of the camera (shooting station) in the course of shooting. In many mosaic systems, the images to be stitched together are selected by hand. They are a series of photos shot from the scene with definite intervals between two adjacent shooting stations. The whole shooting area must be covered with the photos [8].

In order to use real world image as original texture to satisfy the requirement of texture mapping, the shot photos should have definite overlap degree.

Definition1. Let $P\%$ be the percentage that the repetitively shot part P occupies the area of the whole photo along the normal [9, 10] direction of the image plane (See Figure 1), then $P\%$ is called the overlap degree,

$$P\% = P/L \qquad (1)$$

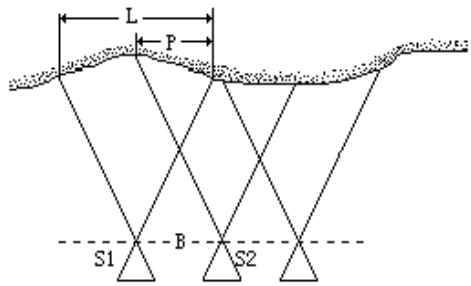


Figure1 Overlap photos

In the course of taking real world picture, if we shoot a photo every other definite time, there should be a certain overlap region in the adjacent photos. Because of the influence of many factors such as the lens decomposing ability of camera, projection error, leaning error and negative planishing et al. in a photo, the image distortion and the point displacement of the edges are bigger than those of the centre. For this reason, the definitions of the edges are worse than that of the centre. The overlap degree of the two adjacent photos cannot be too small. It is commonly bigger than 15% at least. In fact, when we are shooting, in order to ensure and make it convenient for image mosaic, we should shoot a reference object in the overlap part of the two adjacent photos and overlap the reference object when mosaic is processed.

The overlap degree $P\%$ can be controlled by controlling the distance between the two shooting stations in the course of shooting. For example, B is the distance, let S_1 and S_2 be the two shooting stations (see Figure 1), i.e.

$$B = S_1S_2 = L \cdot (1 - P\%) = 2H \cdot tg\alpha/2 \cdot (1 - P\%) \qquad (2)$$

Here, H is the average distance from the camera to the shot objects and α is the view field angle of camera.

Generally, to construct a panoramic image, the primitive images should be shot in almost the same

environment, which has not big difference in neighbor images. For example, there are two shot images with one overlap region (a building) in them, we require the situation of the building is the same, that is to say, it has no window open in one image, should have no window open in another image. We can do this by selecting the shooting time or doing digital image processing on it.

2) Convert the images into the digital ones with high pixel. (The analog pictures should be dealt with by the high resolution scanner);

3) Set up the image mosaic model.

We consider that the original images are ordered in the sequence that they captured. Through image registration, we estimate the relative transformation between two images of the sequence.

In order to develop the mosaic algorithm; we must construct a general model. Generally, we have two original images with one same overlap region. After processing and merging the overlap region, we get a destination image, which covers the scene of the two original images.

Definition2 [11]. As displayed in Figure 2, let $I1$ be the overlap region of the left original image O1, $I2$ is the overlap region of the right original image O2, and $I3$ be the overlap region of the destination image.

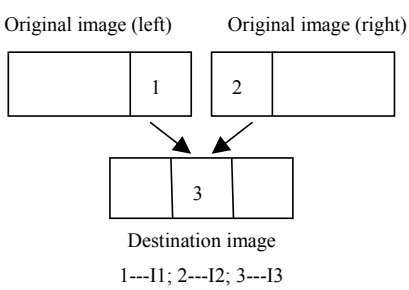


Figure2 Overlap images

We give a model as follows:

$$I3 = \frac{\omega_1(1-\alpha)I1 + \omega_2\alpha I2}{\omega_1(1-\alpha) + \omega_2\alpha} \qquad (3)$$

Where, ω_1 and ω_2 are power coefficients. They are determined by the effect of the two original images after the initiatory original images processing. When the effect of one original image is better than the other one, its power coefficient will be set much larger than that of the other (e.g. 50:1); the grey level of $I3$ will close to the grey level of this original image. For example, if $\omega_1 \rangle\rangle \omega_2$, then the grey level of $I3$ closes to the grey level of $I1$.

Simple proof: for $0 < \alpha < 1-\alpha < 1$, $I1$ and $I2$ is limited, $\omega_1 \rangle\rangle \omega_2$, so we can ignore the values of $\omega_2\alpha I2$ in numerator and $\omega_2\alpha$ in denominator of formula (3). Thus $I3 = \dfrac{\omega_1(1-\alpha)I1}{\omega_1(1-\alpha)} = I1$.

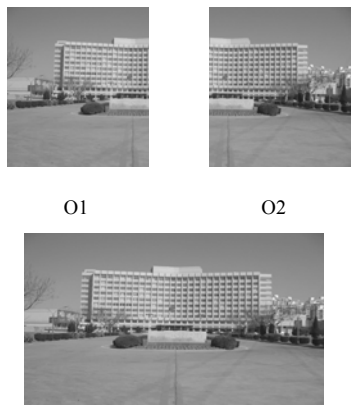


Figure3 Destination image of seamless image mosaicing, $\omega_1 = 50, \omega_2 = 1$

In fact, we want the left edge of the overlap part is exactly the left image, and the right edge of the overlap part is exactly the right image, the algorithm can also fulfill it. When $\alpha = 0$, $I3 = I1$ and when $\alpha = 1$, $I3 = I2$. That is to say, the gray level of the left edge of the overlap region equals to that of $I1$, and the gray level of the right edge of the overlap region equals to that of $I2$. So, we can get smooth transition in the overlap region. This method can avoid the blur and seam in the edges of the overlap region.

4 Problems occurred in the jointing of images by digital image processing technique for crime scene investigation

With their clearness, convenience and easily modifying, digital images are made the effective supplements for the analog images, they also have disadvantages. E.g. they are easy to be modified. Having been processed by computer or software, digital pictures are hard to be restored back into the exact original images. So it is difficult to distinguish the Authenticity of the digital images. The Huanan Tiger Incident is an Anti-Case to the legal validity of digital images and the

digital image-mosaic technique. The validity of digital images has become a heat issue in the field of law.

During the later period of image-jointing, there are certain difficulties existing in the partial adjustments of images.

5 Conclusions

Creating image-based virtual environment has many potential applications because they can provide realistic details of visual information. Once we visualize the real environment in terms of image mosaics or panoramic movies, we can overlay geometrical graphics objects or real objects onto scene.

The whole automatic process of our algorithm is made software named by Magic Image mosaic. The test images are taken from real crime scene or public area [12]. The evaluation is carried on an Ultra SPARC 1 workstation. The two results of image mosaics of two images are shown in Figure 4. From the destination image, we can see that there is not any obvious mosaic gap.

Through shooting some parts of one scene, we constructed a large field of view panoramic image. An experimental result of the seamless image mosaic algorithm is given in Figure 5.

(a) Crime scene 1, China

(b) Crime scene 2, China

Figure4 Construction of seamless mosaic for multi-viewpoint overlap images

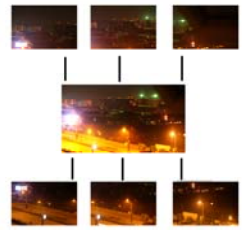


Figure5 The overall scene of crime investigation

References

[1] MEIER J B, Painterly rendering for animation[A], Computer Graphics Proceedings[C], 1996, pp. 477-484

[2] L.Teodosio, W. Bender, "Bender. Salient video stills: Content and context preserved", Proc. ACM Multimedia'93, 1993, pp.33-46

[3] R. Szeliski,"Video mosaics for virtual environments", IEEE CG&A,1996. pp. 22-30

[4] M. Irani, P. Anandan, S. Hsu, "Mosaic based representations of video sequences and their applications", IEEE Conference on Computer Vision and Pattern Recognition, 1995

[5] B. Rousso, S. Peleg and I. Finci, "Mosaicing with generalized strips", DARPA-Image Understanding Workshop, 1997

[6] L. McMillan, G. Bishop,"Plenoptica modeling: An image-based rendering system", Proc. SIGGRAPH'95, 1995. pp.39-46

[7] Szeliski R, "Video mosaics for virtual environments", Computer Graphics and Applications, Mar 1996, pp. 22-30

[8] Zhiling Yang, Kai Wang et al, Digital image fetching, processing and applications, Beijing: Posts & Telecommunications Press, 2003

[9] Szeliski R,"Image Mosaicing for Tele-reality Applications", IEEE Computer Graphics and Applications, No.6,1994, pp. 44-53

[10] Dan Liu, The mathematical foundation of digital image processing, Beijing: Defence Industry Press, 2005

[11] Dan Liu, "Novel seamless image mosaicing algorithm For virtual environment", International Journal of Computer Mathematics, Vol.84, No.2, February 2007, pp. 219-229

[12] Dan Liu, Xiangjian He, "Seamless image mosaic for multi-viewpoint overlapping pictures", International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'04),2004, pp. 395-398

Research on Medical Image Visualization and Interactive Virtual Cutting*

Jian Wu Xiaoping Sun Guangming Zhang Zhiming Cui Jing Xu Jie Xia

The Institute of Intelligent Information Processing and Application, Soochow University, JS 215006 China

Email: szjianwu@163.com

Abstract

Medical imaging visualization is an important field of scientific computing visualization. Three-dimensional segmentation plays a very important role in three-dimensional reconstruction of human tissues and organs and region of interest. In order to better observe information of human internal tissues and organs and provide reliable basis for clinical diagnosis, pathological analysis and treatment, it is indispensable to do cutting on volume data. Virtual cutting usually processes in the user's interactive operations after three-dimensional reconstruction. This paper does research on algorithm of cutting volume data, and proposes a set of solutions including plane cutting, bounding box cutting and complex graphics cutting. Experiment results indicate that algorithm in this paper is practical and efficient, which can reach favorable virtual cutting effect. By this meaning, doctors can know the internal information of volume data more clearly and reduce the amount of computation at the same time.

Keywords: Visualization; Three-dimensional Segmentation; Interactive Cutting; Virtual Cutting; Medical Image

1 Introduction

With the rapid development of medical imaging, medical imaging technologies such as CT and MRI have become important means of modern medical clinics, while scientific visualization technology provides more helps for doctors in applications of precise diagnosis and treat, operations plan, and visual operation [1]. Virtual cutting is an important interactive operation in the course of 3D visualization and plays an increasingly important role in medical-aided diagnosis and treatment. After three-dimensional reconstruction of medical imaging, many important doctors' region of interest (such as organ lesions and surrounding tissues) are within volume data, so that it only can be observed after resecting external structures. At present, plenty of image segmentation algorithms have been proposed [2,3], but these algorithms are difficult to achieve real-time interactive operation and difficult to do discretionary resecting in accordance with doctors' wishes. Cutting operation of the volume data can meet the needs of doctors' data-processing efficiency and interactive requirement, and provide more reliable data for clinic diagnosis, pathological analysis and treatment, so it gets more and more concern.

This paper does research on interactive virtual cutting of medical data, and a set of practical virtual cutting algorithm is proposed based on medical imaging visualization, which can retain the part of volume data doctors are interested in and be used for better quantitative and qualitative analysis.

2 Summary of Volume DATA

2.1 Volume Data Definition

Usually, a biopsy sequence for a part of the body is

This research was partially supported by JiangSu Province Colleges Industry Promotion Project: the development and industrialization of universal medical image diagnosis and interpretation system (JHB06-26) and Soochow Technological Breakthrough Project(Industry): medical image computer-aided diagnosis system based on the 3-D reconstruction and visualization(SGR0703).

gained from CT, MRI and other medical imaging equipment. To achieve three-dimensional cutting, we must first set these slices into three-dimensional data. Suppose there are n piece of slices in the sequence, the resolution at the direction of x-axis and y-axis are I and J separately in each slice(Usually I= J). Sz (x,y) expresses the zth slice, $0 \le x \le I-1$, $0 \le y \le J-1$, $0 \le z \le N-1$. By heaping up slice sequence in the z-axis direction, the 3-D data structure can be formed.

$V(x,y,z)=Sz(x,y)\ x \in [0..I-1]\ y \in [0..J-1]\ z \in [0..N-1]$ （1）

To the construction of three-dimensional data structures of sequence slice, it is necessary to know the numbers, the adjacent pixels space in the image, the location of Z-axis. Such information has records in a DICOM document. Table 1 shows the three-dimensional reconstruction information of the DICOM document.

Table 1 Document Meta-information about 3D Position

Tag	Name	The value
(0008,0060)	Image type	CT
(0020,0011)	Sequence number	7
(0020,0013)	Document number	53
(0028,0030)	The distance between adjacent pixels	0.390625;0.390625
(0020,1041)	The position of fault image	-127.6

Through anatomizing all image documents in series, we can obtain document's meta-information, and then construct 3D volume data. As shown in Figure 1 and Table 1, this DICOM document is No. 98 image in No. 7 series shot for patient. Through all serial images' meta-information, the distance between adjacent pixels and the position of every slice image would be gotten, we acquire 3D space of this volume data is: 0.390625;0.390625;2.799999. So each pixel area in those images is 0.390625×0.390625 mm2, and the distance between adjacent slice images is 2.799999 mm. The constructed volume data model is shown as follows.

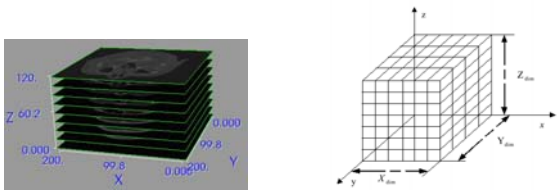


Figure1 Data structure model

The data structure is a three-dimensional grid structure, V(x, y, z) on behalf of the gray value of the voxel (x, y, z). $(\Delta x, \Delta y, \Delta z)$ shows the space in the X,Y and Z-axis directions. The size of every volume pixel (voxel) in volume data is $0.390625 \times 0.390625 \times 2.799999$ mm3.

2.2 Volume Data Visualization Technology

Visualization in scientific computing is to study how to convert scientific data into visualized data, which can help scientists understand information calculation. Medical image visualization is an important application of visualization in scientific computing technology in the medical field, which is well used for clinical application. The required data of 3D medical image visualization is a series of faults Image Sequences produced by imaging equipments, such as CT, MRI. After discrete the sequence data, three-dimensional structure of the data field can be generated. Through three-dimensional data field visualization technology, 3D object model can be reconstructed, and the visual images that accurately reflect the organizations of the human body can be created, in order to take full advantage of the human visual system to display characteristics of the three-dimensional shape of human organs for better quantitative qualitative analysis.

From the 80s, many visualization of 3-D reconstruction methods have been proposed and successfully applied to the field of medical imaging. 3-D visualization technology in general can be divided into: Indirect Volume Rendering，IVR and Direct Volume Rendering，DVR [4].

RCA(Ray Casting Algorithm) is a classical algorithm of the volume rendering [5,6]. It casts parallel rays from each pixel on the screen according to the direction of view throughout the 3D volume data sets. The sampling points are chosen by the same step along the rays. And the color intensity and opacity of the sampling pixel is determined by the eight pixels which are the nearest to it using tri-linear interpolation. Finally, image is generated on the screen which pixels are fitted from the sampling points in that rays by the forward to

back order or reverse. In this paper, RCA was used to render the cutting volume data.

3 Virtual Cutting Algorithm

The principle of virtual cutting is defined as follows. The voxels of volume data will be rendered if they are in the region. On the contrary, the voxels are regard as null. The region for cutting can be closed or not. When using a plane surface to cut volume data, the region is not closed, but it is closed when using a cube. In this paper, three virtual cutting algorithms are proposed after the fully study of the characteristics of data, including plane cutting algorithm, bounding box cutting algorithm and complex graphics cutting algorithm.

In this paper, the experimental data are CT image sequences, which were acquired from the First Affiliated Hospital of Soochow University. The number of images about human head with DICOM 3.0 standard is 120, and image resolution is 512*512.

3.1 Plane Cutting Algorithm

If the plane defined in three dimensional space intersects with the volume data, rendering processing on the volume data is only done on one side of the plane. The slice images can also be gotten by using two parallel plane to cut volume data.

Suppose plane equation in the three-dimensional space is A×x+B×y+C×z+D=0(x, y, z are the coordinates of volume data). When A×x+B×y+C×z+D>0, the volume point (x,y,z) will be rendered, and when A×x+B×y+C×z+D≤ 0, the volume point (x,y,z) is set null.

As known by the point determine polynome equation of plane, as long as one volume point P(x0,y0,z0) and normal vector of a plane n={a1,a2,a3} are determined, the equation of the plane is a1(x-x0) +a2(y-y0) + a3(z-z0) =0. Therefore, if using a plane to cut the volume data, we only need to identify the normal vector of the plane and a point on the plane [7].

(1) Normal vector definition

In 3D space, unit normal vector of Arbitrary plane can be identified by α, β, γ which are angles between the three coordinate axis and the plane, Where $0 \le \alpha \le \pi$, $0 \le \beta \le \pi$, $0 \le \gamma \le \pi$. Just as Figure 2 shows.

Here cos α, cos β, cos γ is called direction cosine of point O to point P and their values can be calculated by the following formula.

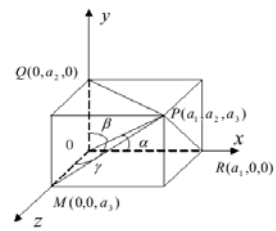


Figure2 Diagram of normal vector

$$\cos\alpha = \frac{a_1}{|OP|} = \frac{a_1}{\sqrt{a_1^2 + a_2^2 + a_3^2}} \quad (2)$$

$$\cos\beta = \frac{a_2}{|OP|} = \frac{a_2}{\sqrt{a_1^2 + a_2^2 + a_3^2}} \quad (3)$$

$$\cos\gamma = \frac{a_3}{|OP|} = \frac{a_3}{\sqrt{a_1^2 + a_2^2 + a_3^2}} \quad (4)$$

The direction cosines meet the following equation.

$$\cos2\alpha + \cos2\beta + \cos2\gamma = 1 \quad (5)$$

Therefore, normal vector will be determined as long as any two angle values and the third angle's quadrant are given.

(2) Choice of slice plane point

Suppose P is one point in the slice plane. In order to reduce computational complexity, point P is initialized to be the center of volume data 0(M/2,N/2,K/2), Where M, N and K indicate respectively the number of sampling points along X, Y and Z. The principle is shown as the following.

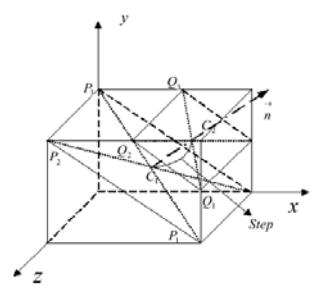


Figure3 Diagram of plane cutting algorithm

Suppose Step is the moving distance along normal vector n, Step can be negative and positive. When Step is negative, it means Step's direction is opposite to normal vector n. The P'(x', y', z') expresses the corresponding point to P(x, y, z) after the slice moves, and the coordinate relation between P' and P is as follows.

$$\begin{cases} x' = x + Step * \cos\alpha \\ y' = y + Step * \cos\beta \\ z' = z + Step * \cos\gamma \end{cases} \quad (6)$$

Figure 4 the result of the experimental of plane cutting algorithm.

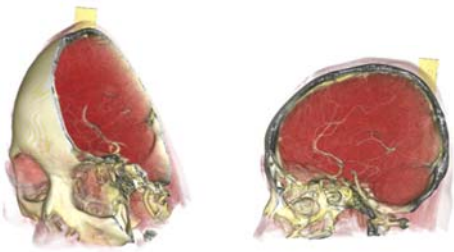


Figure4 Rendering effect of plane cutting algorithm

3.2 ZBounding Box Cutting Algorithm

Bounding box cutting limits the volume data in a rectangular bounding box, which is the smallest rectangular of the effective part of the volume data. Then in accordance with the bounding box, some effectively cutting is done to reduce the size of the data field. Suppose the spatial coordinates of image is *OXYZ*, the object space coordinates is *OUVW*. Traversaling the volume data to determine the minimum bounding box, the length, width and height is respectively as the following.

$$(x_{\max} - x_{\min}) \times \Delta x, (y_{\max} - y_{\min}) \times \Delta y, (z_{\max} - z_{\min}) \times \Delta z$$

Where $\Delta x, \Delta y, \Delta z$ are respectively the interval steps along the three coordinate axis. If voxel point is outside the bounding box, voxel is set null.

When rendering the volume data, volume data must be transformed from the space coordinates to image coordinates, and this procedure is called volume data transformation. In order to reduce unnecessary voxel transformation, we only project minimum bounding box to the imaging plane, only consider the projection light from bounding box, and ensure that every projection light intersects with volume data.

Figure 5 is the result of the experimental of bounding box cutting algorithm.

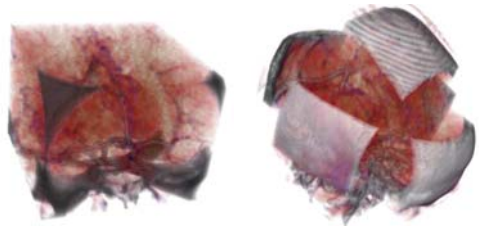
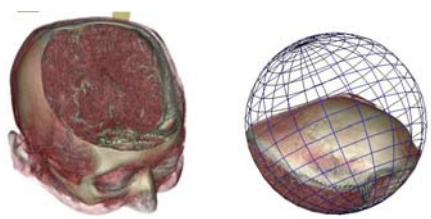


Figure5 Rendering effect of bounding box cutting algorithm

3.3 Ball Cutting Algorithm

With the continuous development of 3D medical image reconstruction technology, virtual surgery gradually becomes a hot research. Virtual surgery use computer to simulate the surgical operation, which is a new technology combining medicine and information science in recent years. Virtual surgery uses three-dimensional model to simulate the arbitrary cutting on human tissues and organs. Traditional approach of "a clean cut" is not suitable for simulating the operation, so a number of complex graphics cutting method are demanded. Briefly, the arbitrary space surface or a graphic of closed region all can be used for virtual cutting operation. In this paper, ball cutting algorithm is proposed.

In the process of ball cutting, user must determine a point in space as spherical center firstly, and then set a length value as the spherical radius. When rendering the 3D objective, different treatments are done by judging whether the volex is located in the internal sphere. Figure 6 is the result of the experiment of ball cutting algorithm.



(a) Image reconstruction after cutting (b) The cut part

Figure6 Rendering effect of ball cutting algorithm

Similar with ball cutting, space surface can be used for volume data cutting. The key point of this method is that it requires the user to make choice several vertices to generate the quadratic surface. Further research about this method are expected.

4 Conclusions

This paper proposes a set of interactive virtual cutting algorithm including plane cutting, bounding box cutting and complex graphics cutting based on medical imaging three-dimensional reconstruction research. Experiment results present that algorithm in this paper is practical and efficient, which can reach favorable virtual cutting effect, and doctors can observe the internal information of human tissues and organs more clearly by this way. By simple manual operation, users can do virtual cutting at any direction and location, and can do interactive operation such as moving, zooming, rotating, locating operation to the cut reconstruction image. Our research is not only limited to interactive virtual-cutting of human tissues and organs, further research will be focused on the physical model of soft-tissue research, to ensure realistic of virtual operation by simulating physical response to external force on tissues and organs.

References

[1] ShengZe Tang. 3-D Data Field Visualization[M]. Beijing: Tsinghua University Press, 1999. 12

[2] Yujin Zhang. Image Segmentation[M]. Beijing: Science Press, 2001. 2

[3] Xiping Luo, Jie Tian, etc. Survey of Image Segmentation Method[J]. Pattern Recognition and Artificial Intelligence, 1999, 12(3): 300-312

[4] Jie Tian, Shanglian Bao, Mingquan Zhou. Medical Image Processing and Analysis. [M]. Beijing: Publishing House of Electronics Industry, 2003. 5

[5] Levoy M.. Efficient Ray Tracing of Volume Data[J]. ACM Transaction on Graphics, 1990, 9(3):245-261

[6] Rudiger Westermann, Bernd Sevenich. Accelerated Volume Ray-Casting using Texture Mapping[A]. l2th' IEEE Visualization Conference,2001,24-26

[7] Haibo Zhang. 3-D Segmentation and Visualization of Computed Tomography[D]. Shandong Normal University, Masteral Dissertation, 2005

The Measurement of Investment Risks in Listed Open-ended Fund

Cui Lu[1] Shujun Ye[2]

Economics and Management School, Beijing Jiaotong University, Beijing, China, 100044

Email:1 cuilu1985@126.com; 2 shjye@bjtu.edu.cn

Abstract

This paper attempts to research on the measurement of the investment risk of the listed open-ended fund by means of variance-covariance method. By considering the investment weight of stocks and funds, the investment risk of the stocks and funds can be measured and the original model of VAR can be changed into the model of the measurement of investment risk. By empirical study, the conclusion is: the investment risk of bond fund is lower than that of stock fund; the proportion of investment risk in per net worth is big; the investment risk and per net worth of the listed open-ended fund is being related. Investors can use the conclusion to make the right decision; Fund operators can also use it to pay attention on the investment risks promptly.

Keywords：listed open-ended fund, stock investment risks, bond investment risks, VAR model, variance-covariance method

1 The Problem of the Measurement of Investment Risks in Listed Open-ended Fund

The invest risk of the listed open-ended fund results from by the appearance and operation of the listed fund brings the investors uncertainty about gain and loss. This paper builds up the investing risk measurement model of the listed open-ended fund based on the VAR model, which helps reveal the risk of investing and help investor make wise investing decisions by adjusting the fund position ratio structure and lower the risk.

1 Variance-covariance method of VAR

VAR is the biggest potential loss value of a certain kind of financial asset in a certain period in the future under a certain confidence level. It can be represented as:

$$\Pr ob(\Delta P < VAR) = 1 - c \qquad (1\text{-}1)$$

In the formula above, ΔP is the amount of loss in the carrying period Δt, **c** is confidence level. VAR is the risk value under the confidence level **c**. From this formula, we can get the probability for the loss less than VAR in the carrying period is 1-**c**. [1]

Three fundamental elements are required in the calculating of the VAR. They are the Carrying Period, the Confidence Interval and the Distribution Characteristics of Future Asset Value. A measurement is usually needed, too.

Major methods of calculating VAR are Parametric Method, Historical Stimulating Method, Pressure Testing Method, Monte Carlo Method, etc. [2]

Parametric Method needs to foreknow or presuppose the probability distribution of the return on assets. And one the major parametric methods is Variance-covariance Method. In variance-covariance method we presuppose that the return on assets fulfill the normal distribution, then estimate the VAR value by calculating the variance and covariance matrix of all the returns on an assembling of assets.

Suppose the initial value of the asset is P_0, the least return ratio in the carrying period is R^*, the lowest value of the assets is P^*, the average return ratio is μ, we can get

$$P^* = P_0(1 + R^*) . \qquad (1\text{-}2)$$

So the VAR can be presented as

$$VAR = E(P) - P^* = P_0(\mu - R^*) \quad (1\text{-}3)$$

In a normal distribution,

$$-\alpha = \frac{-|R^*| - \mu}{\sigma}, (\alpha > 0) \quad (1\text{-}4)$$

$$1 - c = \int_{-\infty}^{P^*} f(p)dp = \int_{-\infty}^{|R^*|} f(r)dr = \int_{-\infty}^{-\alpha} \Phi(\varepsilon)d\varepsilon \quad (1\text{-}5)$$

Given the c value, we can find deviation α in the Statistics Table to fit with formula above. So we can get

$$R^* = -\alpha\sigma + \mu \quad (1\text{-}6)$$

Suppose μ and σ are on the same day, the VAR in a carrying period Δt is:

$$VAR = P^0(\mu - R^*) = P_0\alpha\sigma\sqrt{\Delta t} \quad (1\text{-}7)$$

In the preceding formula, α is the quantiles according to the confidence in the normal distribution. For example, 2.23 is the quantiles going with the confidence level 99%, and 1.65 goes with 95%.[4]

2 Risk Testing Model on Listed Open-ended Fund based on VAR Model

The investment risk on the listed open-ended fund is mainly the risk on stocks and bonds. So to estimate the risk of the listed open-ended fund ought to estimate the risk on stocks and bonds by the modified VAR model put forward by this paper.

Step 1: Calculate the biggest loss on the STOCK investment under the standard VAR model.

$$VAR = \omega P_0^1 \alpha^1 \sigma^1 \sqrt{\Delta t} \quad (2\text{-}1)$$

Step 2: Calculate the biggest loss on the BOND investment under the standard VAR model.

$$VAR = \upsilon P_0^2 \alpha^2 \sigma^2 \sqrt{\Delta t} \quad (2\text{-}2)$$

Step 3: Put the risk on Listed Open-ended Fund into consideration:

$$VAR = VAR = \omega P_0^1 \alpha^1 \sigma^1 \sqrt{\Delta t} + VAR = \upsilon P_0^2 \alpha^2 \sigma^2 \sqrt{\Delta t} \quad (2\text{-}3)$$

In the above formula, P_0^1 is the initial value of the bond, put the confidence level at 95%, so we can get

$$\alpha^1 = \alpha^2 = 1.65 \quad (2\text{-}4).$$

σ^1 is the standard deviation of the return ratio on stocks, as σ^2 is the bonds'. ω is the investment ratio to the stocks by the fund, as υ is what to the bonds. And Δt is the carrying period.

3 Case study

3.1 Sample Selecting

The data of the funds and the fund assembling in this paper comes from the Fund of "http://www.china finan-ceonline.com/". We can get the basic research data by recording the trading data of the stocks and bonds in each single trading day.

In order to make the data more comprehensive and researchable, this paper picked up four equity funds and four bond funds as samples by random sampling. See as following:

Table 1 Bond Names

Equity Fund		Bond Fund	
No.	Name	No.	Name
000001	HUAXIAZENGZHANG	070005	JIASHIZHAIQUAN
070002	JIASHIZENGZHANG	217003	ZHAOSHANGZHAIQUAN
162201	HEYINZENGZHANG	001001	HUAXIAZHAIQUAN A/B
160603	PENGHUASHOUYI	121001	GUOTOURONGHUA

As stocks and bonds are the main fields in which funds are invested, we assume funds are invested in these two fields exclusively. The data below are stocks ranking the top ten and bonds in the position of funds ranking the top five (or top four) at the period between 30 June, 2007 and 30 September, 2007.

3.2 The Measurement of the Investment Risk of Stocks

Since the top ten stocks are the biggest part in which the funds are investing, we just make the top ten stocks as representatives to analyze the stock investment risk

$$(V_1 = \omega P_0^1 \alpha^1 \sigma^1 \sqrt{\Delta t}) \quad (3\text{-}1)$$

The following graph is a line chart according to the stock investment risk of the picked eight funds.

Table 2 Position Clarities

HUAXIAZENGZHANG	JIASHIZENGZHANG	HEYINZENGZHANG	PENGHUASHOUYI	JIASHIZHAIQUAN	ZHAOSHANGZHAIQUAN	HUAXIAZHAIQUANA/B	GUOTOURONGHUA
Detail Position (stock)							
SUNING	QIANJIN PHARMACY	GEHUA CATV	CMBChina	cmpd	FOODS1	COSCO	Yantai Wanhua
CMBChina	lolo	cj-elec	POLY REAL ESTATE	SANFANGXIANG	COSCO	cmpd	CIB
LUZHOULAOJIAO	sunward	Hundsun Technologies	COSCO	COSCO	SDB A	XNTG	cs ecitic
zjgold	CIB	chinaship	Vanke A	PING AN	ZHUZHOU SMELTER	goldeneagle	cmpd
saicgroup	wangfujing	reht	chinaship	bankcomm	XISHAN COAL	SHANDONG HAIHUA	CNOOC ENGINEERING
zjpark	DONG-E E-JIAO	GREE	OCTHOLDING A	lzhg	cs ecitic	Sea Star	CHINA YANGTZE POWER
yypaper	Gemdale	changan	Gemdale	tjcep	SPD BANK	dngroup	COSCO
Gemdale	ctvro	XINHUA MEDIA	DONGFANG ELECTRIC	DAQIN RAILWAY	OCTHOLDING A	Topband	wisco
HONGYUAN SECURITIES	CNOOC ENGINEERING	XJ ELECTRIC	Nonfemet	CHINA CITY BANK	ZHEJIANG MEDICINE	Sunlord	Gemdale
cmpd	CMBChina	Aisino	WULIANGYE	nwti	cncm	Annada	wangfujing
Detail Position (bond)							
06 Central bank bill 68	06 Central bank bill 76	05 national debts 14	21 national debts (15)	99 national debts 8	06 cdbank bills32	05 national debts (14)	07eximbank bills 09
SHAOGANG convertible bonds	20 national debts (10)	06 cdbank bills 31	02 national debts (10)	02 national debts (10)	03 cdbank bills 28	07 Central bank bill 18	02 national debts (14)
99 national debts (8)	21 national debts (15)	07 adbank bills 01	07 Central bank bill 84	07 Central bank bill 83	GUIGUAN convertible bonds	07 national debts 02	05 adbank bills 16
GUIGUAN convertible bonds	03 eximbank bills 01	07 Central bank bill 15	07 Central bank bill 91	21 national debts (15)	02 cdbank bills 08	07 national debts 14	07 Central bank bill 81
20 national debts (10)	02 national debts (14)	21 national debts (15)	02 national debts (14)	06 Central bank bill 76	21 national debts (15)		

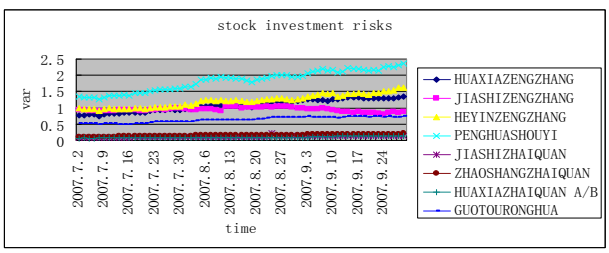


Figure1 Stock Investment Risks

From the chart, the line representing the stock investment risk always goes below the line of bond's, which comes the conclusion that investment risk of bond funds are obviously lower than that of stock's. This is decided by the investment field and position ratio of different type of funds.

3.3 Measurement of the Investment Risk of Bonds

Since the top five (or top four) bonds are the major part in which the funds are investing, we just make the top five (or top four) bonds as representatives to analyze the bond investment risk

$$V_1 = \upsilon P_0^2 \alpha^2 \sigma^2 \sqrt{\Delta t} \qquad (3\text{-}2)$$

The following graph is a line chart according to the bond investment risk of the picked eight funds.

From the chart, the line representing the bond

investment risk always goes above the line of stock'. This is because the investment of the bond funds relies on all kinds of bonds. Although the risk of bonds is relatively lower than the stocks', while too much quantity of capital flowing into bonds may bring about more risks than the stock investment. At the same time, we can find that the fluctuation of the line of the bond investment is smaller than the stock's, which results from the nature of bond and stock.

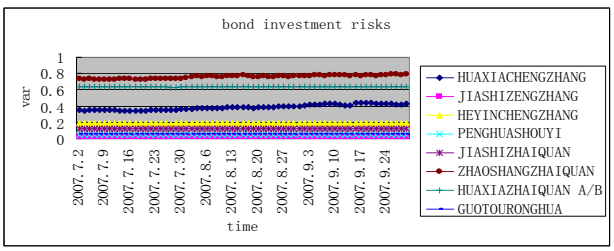


Figure2 Bond Investment Risks

3.4 Measurement of Investment Risk

According to the formula :

$$VAR = V_1 = \omega P_0^1 \alpha^1 \sigma^1 \sqrt{\Delta t} + V_1 = \upsilon P_0^2 \alpha^2 \sigma^2 \sqrt{\Delta t} \quad (3\text{-}3)$$

which gives out the measurement method for investment risk, we can find that the investment risk for funds is the sum of stock risk adding bond risk. The following chart demonstrates more clearly the risks among the eight funds.

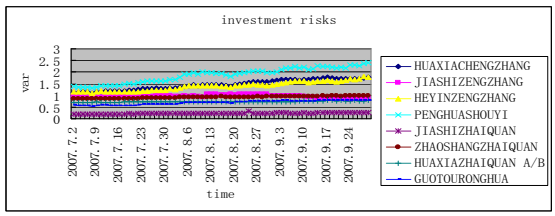


Figure3 Investment Risks

This chart has directly illustrated that the investment risk of bond fund is generally lower than the equity fund. That is because the investment of bond fund always aims at various bonds with fewer risks but more stability. However, stock fund aims at stocks with more risks and bigger fluctuation. [5]

3.5 The Analysis on the Relevance between Investment Risk and the Net Value per Fund

From the table above, six out eight of the target funds' VAR percentage in net per share exceed 40%, which means the investment risk existing is a essential consideration. Among the data, VAR percentage of PENGHUASHOUYI exceeds 1, which results from the events, such as profit sharing, splitting, requesting and redeeming bring out the changes of net per share. This also reminds us the percentage of investment risk in the net value per fund is on a quite high level. So to invest in fund, one could make wiser decision by comprehensively analyzing the investment risks in tables like the above one.

Table 3 Percentage of VAR in the Net Value per Fund

	Percentage of VAR in the Net Value per Fund							
	HUAXIAZENGZHANG	JIASHIZENGZHANG	HEYINCHENGZHANG	PENGHUASHOUYI	JIASHIZHAIQUAN	ZHAOSHANGZHAIQUAN	HUAXIAZHAIQUAN A/B	GUOTOURONGHUA
2007-7-2	0.7066	0.2608	0.6976	0.9054	0.1732	0.7286	0.6656	0.4566
2007-7-3	0.6941	0.2508	0.6762	0.8648	0.1717	0.7150	0.6689	0.4402
2007-7-4	0.7199	0.2494	0.6939	0.8911	0.1737	0.7275	0.6682	0.4528
2007-7-5	0.7368	0.2597	0.7209	0.9258	0.1737	0.7205	0.6673	0.4541
2007-7-6	0.6904	0.2497	0.6702	0.8557	0.1705	0.7171	0.6684	0.4280
2007-7-9	0.6925	0.2439	0.6795	0.8713	0.1730	0.7746	0.6701	0.4424
2007-7-10	0.7151	0.2503	0.7073	0.9109	0.1759	0.7777	0.6690	04585
2007-7-11	0.7112	0.2472	0.6889	0.8985	0.1759	0.7797	0.6696	0.4590
2007-7-12	0.7161	0.2507	0.6851	0.8981	0.1760	0.7815	0.6672	0.4376
2007-7-13	0.7214	0.2552	0.6875	0.9046	0.1776	0.7815	0.6672	0.4393

Continued

	Percentage of VAR in the Net Value per Fund							
	HUAXIAZENGZHANG	JIASHIZENGZHANG	HEYINCHENGZHANG	PENGHUASHOUYI	JIASHIZHAIQUAN	ZHAOSHANGZHAIQUAN	HUAXIAZHAIQUAN A/B	GUOTOURONGHUA
2007-7-16	0.7331	0.2558	0.6977	0.9216	0.1772	0.7815	0.6672	0.4434
2007-7-17	0.7189	0.2505	0.7008	0.9487	0.1784	0.7786	0.6685	0.4430
2007-7-18	0.7215	0.2423	0.6956	0.9518	0.1806	0.7796	0.6707	0.4497
2007-7-19	0.7219	0.2403	0.6866	0.9366	0.1780	0.7799	0.6688	0.4474
…	…	…	…	…	…	…	…	…
2007-9-7	0.7681	0.2143	0.7601	1.0365	0.1928	0.8650	0.6860	0.5225
2007-9-10	0.7759	0.2074	0.7603	1.0130	0.1942	0.8681	0.6913	0.5233
2007-9-11	0.7889	0.2053	0.7492	1.0055	0.1925	0.8646	0.6957	0.5170
2007-9-12	0.7970	0.2021	0.7584	1.0385	0.1944	0.8686	0.6967	0.5209
2007-9-13	0.7890	0.1961	0.7743	1.0175	0.1992	0.8657	0.6983	0.5375
2007-9-14	0.7959	0.1908	0.7688	1.0088	0.1998	0.8881	0.6990	0.5334
2007-9-17	0.7975	0.1891	0.7661	0.9967	0.2011	0.8793	0.7047	0.5344
2007-9-18	0.7914	0.1915	0.7621	2.0552	0.2000	0.8852	0.7056	0.5265
2007-9-19	0.7934	0.1912	0.7583	2.0617	0.2022	0.8904	0.7034	0.5255
2007-9-20	0.7845	0.1898	0.7674	2.0478	0.2051	0.8767	0.7076	0.5300
2007-9-21	0.7861	0.1872	0.7716	2.1480	0.2055	0.8871	0.6974	0.5237
2007-9-24	0.7913	0.1808	0.7969	2.1651	0.2050	0.8854	0.6989	0.5212
2007-9-25	0.7940	0.1890	0.8033	2.1573	0.2066	0.8937	0.7043	0.5281
2007-9-26	0.7964	0.1960	0.8189	2.2177	0.2026	0.8951	0.7015	0.5252
2007-9-27	0.7947	0.1896	0.8465	2.2177	0.2029	0.8827	0.7015	0.5243
2007-9-28	0.7974	0.1904	0.8328	2.2267	0.2036	0.8945	0.7047	0.5322
Average	0.7469	0.2260	0.7264	1.1130	0.1886	0.8306	0.6768	0.4951
	HUAXIAZENGZHANG	JIASHIZENGZHANG	HEYINCHENGZHANG	PENGHUASHOUYI	JIASHIZHAIQUAN	ZHAOSHANGZHAIQUAN	HUAXIAZHAIQUAN A/B	GUOTOURONGHUA
CORREL	0.9890	0.0629	0.9608	0.1657	0.8859	-0.2857	0.9133	0.9850

By analyzing the relevance between investment risk of listed open-ended fund and the net value per fund in this formula :

$$CORREL = \frac{E([X - E(X)][Y - E(Y)])}{\sqrt{D(X)}\sqrt{D(Y)}} \qquad (3\text{-}4)$$

we can get the relevance values of the eight sampling funds, which shows us more than half of them exceed 0.5. That means the two elements present correlation coefficient, as well as the higher return comes from higher risks. Furthermore, in order to appeal more investors, the fund operator could adjust the position structure, for example, by increasing lower risk stocks and reducing higher risk stocks to lower the investment risk for their funds under the same fund performance.

4 Conclusion

This paper introduced the VAR model on the basis

of foreign research and some adjustments of writer's own. By considering the investment weight of stocks and funds, we deduce the investment risk of the stocks as well as of the bonds, and furthermore we can get the investment risk of the listed open-ended funds in China.

We come to the conclusion that: First, we can prove the traditional theory that the investment risk of bond funds is generally lower than the equity funds by comparing the VAR of the two kinds of fund. Second, the percentage of investment risk in the net value per fund is rather high. As a result, investors ought to reconsider the possible risk when investing. While they are investing, they could make full use of the model constructed and adjusted in this paper to calculate the investment risk, then consider the risk bearing capability of their own and make correct decision at last. Third, from the correlation coefficient between investment risk and the net value per fund, we will find the positive correlation between them. Fund operators could keep alert on the investment risk by calculating the fund risk and make adjustment in their position. By this way, they can lower the investment risk of their fund, appeal more investors and expand their business.

References

[1] CHEN Bo-wen, the model of measurement of financial risks——VAR model and the development, [J], Group Economy Research(in Chinese), 2007- 05S, 201-202

[2] Pneza, P, Bansal, VK., the measurement of market risks on VAR, CHA Xiang translate[M], BeiJing: Machinery Industrial publisher(in Chinese), 2001, 24-26

[3] Jorion, R, VAR:the value of risks, ZHANG Hai-yu translate [M], BeiJing: CITIC publisher(in Chinese), 2000, 33-36

[4] SONG Feng-ming, TAN Hui (2004) the measurement of liquidity risks in VAR, [J], Number Economy And Technology Economy Research(in Chinese), 2004-6

[5] WANG Chun-feng. the management of risks in financial market [M], Tian Jin: Tian Jin University publisher(in Chinese), 2001,4-8

[6] WU Xi-zhi, Statistics: from data to conclusion[M], BeiJing : China Statistics Publisher(in Chinese), 2004,13-15

Research on Multipliable Template Pattern Recognition of License Plate

Ying Yang Xiuli Zhang Lin Sheng Peng Zhang

School of Mechanical Engineering and Automation, Northeast University, Shenyang, Liaoning, 110004 ,China

Email: yangyingsy@163.com

Abstract

There is unavoidably a limit to either the pattern recognition of multipliable templates modeling based on the Taylor formula(MTMT) or the statistical pattern recognition based on traditional templates matching(TTM) when using any of them singly to recognize the plate image, considering the changes of illumination, the surrounding environment of the license plate and the license plate itself. An algorithm combining both of them is therefore proposed to recognize the plate image after it is segmented. First, In the day, the number of the white pixels that are added up in a certain range is marked as B. If the number of the whit pixels of the plate image to be recognized is calculated is higher then B, MTMT(the parameters of the noise and different measurement and the error of the eigenspace that could be minimized by the robust regression) is used to recognize the plate image；Or TTM is used to recognize the plate image. It has been proved that the result of the recognition that has good adaptability to the license plate picture with noise interfering and no uniform illumination is better.

Keywords：Pattern Recognition; License Plate; Image; Template; Pixels

1 Introduction

The statistical pattern recognition method is widely used in the pattern recognition. It is a useful way when the plate image is shot in uniform illumination and the plate is clean. But in most cases, single recognition couldn't get good result, because of the dirt on the plate and no uniform illumination. Diversiform methods are adopted aiming at the case of no uniform illumination in this paper.

The pattern recognition of multipliable templates modeling based on the Taylor formula is adopted in this paper in the case of no uniform illumination. PCA is adopted to extract and classify the feature of characters and wipe out overmuch redundancy information. In order to enhance the recognition rate, the image to be recognized is unwrapped in the eigenspace by the Taylor formula. Then it is adjusted by the noise parameters and different rotary parameters. Besides, the error of the eigenspace is minimized by the robust regression, so that it will enhance Robust. Then, the character which has the minimum of the weight Euclidean distance is treated as the recognition result. It enhances the recognition rate by these methods above.

2 Prerecession

The image is put into the recognition system after the image acquisition, plate location and character segmentation (Rectification and interpolation have been done before).

Before the recognition, firstly, PCA is adopted to reduce the dimension, which usually adopts SVD: according to matrix X provided that it has orthogonal matrix U and V, then X could be written as: $X=U \wedge V^{T}$. So that it is called one SVD of matrix X. The image is reduced to the d dimension.

The multipliable pattern basis is approximatively denoted as f, and the pre-d dimension of U is treated as

the recognition basis,

$$\mathbf{f} \approx \sum_{l=1}^{\mathrm{d}} m_l U_l \qquad (1)$$

Let $m_l = \mathrm{f} \bullet U_l^T$, due to U is orthogonal matrix, so m_l is the projection matrix of f in the space of d vectors, and U_l is called recognition basis. Finally, row vector m^T=[$m_1, m_2, \ldots, m_d$] is used to denote f.

3 Multipliable templates modeling recognition

3.1 Training

The digital, alphabetic and Chinese characters training bases are reconstructed and trained according to the structure of the plate. The swatches are made up of the rotated and noisy ones.

3.3 Multipliable templates modeling recognition

The algorithm which combines the pattern recognition of multipliable templates modeling based on the Taylor formula, and the statistical pattern recognition based on the traditional templates matching is to accomplish the recognition of the plates in this paper. The whole method could be divided into two classes:

1. In the case of uniform illumination (for example, the front lights of the car are off), the statistical pattern recognition based on traditional templates matching is used to recognize.

2.When the front lights are on, the pattern recognition of multipliable templates modeling based on the Taylor formula is used to recognize the license plate image with no uniform illumination.

In the day, the number of the white pixels that are added up in a certain range is marked as B. If the white pixels number of the characters about the plate to be recognized isn't larger than B, it belongs to class 1, or it belongs to class 2.

The specific method of class 2 is as follows:

Firstly, the reconstructed value is calculated

Let 57 Chinese characters, 25 alphabetic characters and 10 digital characters be the swatches of the plate recognition. Then every character is rotated in the range of 0°to 5°, or is added noise. At last, it has 4 paradigms. Next, let every character unite into 19×30.Taking Chinese character-base as an example, the dimension of matrix X($\mathrm{x}_{\text{汉}}$) is $19 \times 30 \times 57 \times 4$。

The image to be recognized is projected to the eigenspace of recognition basis. Firstly, let the Chinese characters be rough classified. Then, these characters are minutely classified. The method of four borders area coding [1], which means that the Chinese characters are divided into several subclasses according to the frame feature of Chinese characters, is adopted for rough classification. The pattern recognition of multipliable templates modeling based on the Taylor formula is applied for subcategories. The multipliable templates modeling recognition means that let the plate image to be recognized match the paradigms. But the Chinese characters are maybe different from the paradigms according to the different imaging conditions. So that it is modeled in a series of variable function Γ.

Let $\mathbf{D}(x)$be the plate image to be recognized as $\mathbf{x}_{\mathrm{i}} \rightarrow \mathbf{R}^{n}$, [**D**] be the matrix of the plate image segmented, $[\mathbf{D}]_j$ $(1 \leq \mathrm{j} \leq 57)$ be the j class of the vector [**D**],. According to the analysis of PCA, [**D**] is projected to the eigenspace of the recognition basis, and then the coefficient vector m is obtained. And then it could be expressed as the linear compounding.

In order to enhance the robust, the error of the eigenspace could be minimized by the robust regression,

The function of Geman-Mclure $\rho_{\mathrm{GM}}(\mathrm{x}, \sigma) = \mathrm{x}^2/(\mathrm{x}^2 + \sigma^2)$ is the norm of robust error. Let σ be the measure parameters which control the effect caused by the overstepped points

$$\min \sum_{\mathrm{j}=1}^{\mathrm{d}} \rho(\mathrm{e}_{\mathrm{j}}, \sigma) \qquad (2)$$

$$\mathrm{e}_{\mathrm{j}}(\mathrm{m}) = [\mathrm{D}]_{\mathrm{j}} - \sum_{l=1}^{d} m_l U_{l,j} \qquad (3)$$

The function of Geman-Mclure $\rho_{\mathrm{GM}}(\mathrm{x}, \sigma) = \mathrm{x}^2/(\mathrm{x}^2 + \sigma^2)$ is the norm of robust error. Let σ be the measure parameters which control the effect caused by the overstepped points.

A variable function of $\Gamma(x)$ is brought in this paper according to the matching, Then $\mathbf{D}(x)$ is transformed as $\mathbf{D}(\Gamma(x))$. Next, $\mathbf{D}(\Gamma(x))$ is unwrapped by the Taylor formula. And the high differential coefficients are ignored. So it could be written as:

$$D(\Gamma(X)) \approx D(x) + D_x(x)\Gamma(x) \qquad (4)$$

Where $D_x(x)$ is the first differential coefficient.

Secondly, calculating the error of weights:

In order to recognize the Chinese characters of the plate better, the influence cased by the noise and different image measure to the plate image segmented should be considered. So that, the noise and different measurement of D(x) could be simplified by $S \times D(lx)$. Let l be the noisy parameter, S be the measurement parameter. Then they are brought into the error of the eigenspare which is minimized by the regression:

$$e_j(m,l,S) = \{S \bullet [(D_x(x)\Gamma(l,x) + D(x)]\}_j - \sum_{l=1}^{d} m_l U_{l,j} \qquad (5)$$

Where $\Gamma(l,x) = lx$. Then Eq. (2) could be minimized by multi-scale measurement; m could be calculated with the measurement σ minishing gradually, till the Eq. (2) is converged. So that Eq. (5) could be simplified as:

$$[S \bullet (l+1)D(x)]_j = \sum_{l=1}^{d} m_l U_{l,j} \qquad (6)$$

Let $S' = S \bullet (l+1)$. Then Eq. (6) could be simplified as $[S' \bullet D(x)]_j = \sum_{l=1}^{d} m_l U_{l,j}$

Eq. (2) could be converged by the large measurement of σ, which is treated as the initial value of the small measurement toward the large transform.

If the Eq. 6 could be divided by S', the parameter m_l could be changed as m_l / S'. So the parameter vector could be got when there is a proportion parameter, that is $m_l' = m_l / S'$. After getting the parameter vector m_l', the Euclidean distance between m_l' and the vector parameter of paradigm **q** could be written as:

$$p^2 = \sum_{l=1}^{d} (m_l' - q_l)^2 \qquad (7)$$

That is the error of weights. The minimum of the weight errors could be treated as the recognition result. q_l could be calculated by Eq. (1). Taking "Liao" as an example, the procession is shown in Fig. 1, 2.

The image to be dealt with is the characters segmented in the plate recognition. Fig.3. is shot in the night when the car's lights are on. The nonuniform illumination adds the difficulty to extract the features. We could get Fig. (b) If the statistical pattern recognition of traditional templates matching is only adopted in the feature extraction; nevertheless, using the method:

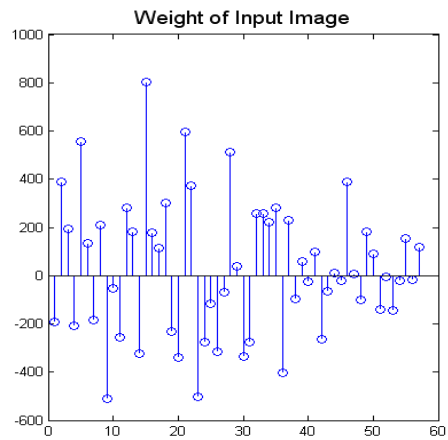


Figure1 The weight of input image

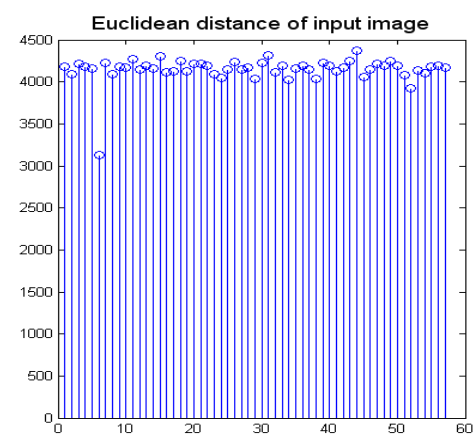


Figure2 Euclidean distance of input image

Firstly, PCA is used to extract the principle analysis. Secondly, it is recognized by the method of four sides coding for rough recognition. Thirdly, it is recognized minutely by the pattern recognition of multipliable templates modeling based on the Taylor formula. Fourthly, the recognition result is the minimum of the weights errors. At last, the result is shown in Fig. (c). The Chinese character is recognized falsely in the

recognition result provided that we only adopt the statistical pattern recognition of traditional templates matching; The result of Fig.(c) is right so the recognition result of Fig. (c) is better than that of Fig. (b),. Moreover, the quality of these original plate images isn't very good due to the light; some plates, even, couldn't be recognized by the eyes of the persons. But the plate image of high brightness that are difficult to recognize even by human eyes, sometimes, could be

Recognize minutely by the pattern recognition of multipliable templates modeling based on the Taylor formula.

(a) Original image

(b) The single statistical pattern recognition of traditional templates matching

(c) The method of multipliable templates recognition

Figure3 The comparison of methods

Taking the Chinese character "Liao" as an example, the program of the swatch to be recognized m_l' can be described as follows:

$$m_l' = (\text{sample} - \text{me} + \frac{\text{um} * \text{st} - \text{ustd}}{\text{ustd}} + 1) * \text{ustd}$$

/st ,where "sample" is the reconstruction matrix of the one to be recognized, "me" is the average of the reconstruction matrix about the one to be recognized, "st" is the standard excursion of the recognition matrix to be recognized, "ustd" is the standard average obtained by the experiment. "um" is the average obtained by the experiment.

4 Experiment

Besides, swatches-base is updated by adding the image recognized in error into the system. So, the system will have good ability to adapt to different condition. The recognition rates are shown in Table1.

In Figure4, the thin line denotes the result by using the recognition method that has been introduced in this paper; the thick line denotes the result by using the statistical pattern recognition of transitional templates matching singly. According to the result, the recognition rate of the classification of digital, Chinese, alphabet characters based on the pattern recognition of multipliable templates modeling based on the Taylor formula, is a little higher than the one using the ordinary method, then the error rate is a little lower. The recognition rate about the classification of digital characters arrives at 100% after 19 times training. Though other classifications exist the problem of the inaccurate recognition rate and the precise, the recognition rate could arrive at 100% by the way of adjusting the parameters of multipliable templates modeling again and again, or adding the image recognized in error and diversiform standard swatches-base into training gather.

Table1 The recognition results

Classes	Tested number	Errors number	Normal errors	Right rate
Digital characters	1526	0		100%
Alphabet characters	498	15	ADELD	97%
Chinese characters	325	23	湘鄂闽浙黑冀	93%

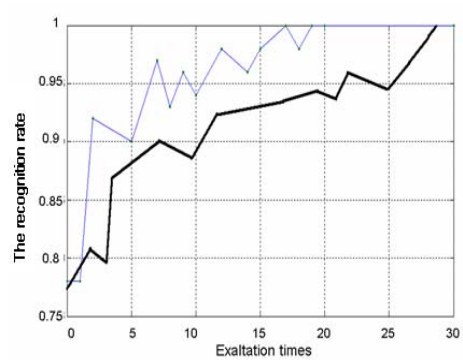


Figure4 The comparison of two ways

5 Conclusion

An algorithm combining both of the pattern recognition of multipliable templates modeling based on the Taylor formula (TTM) and the statistical pattern recognition based on traditional templates matching (TMTM) to recognize the plate image colligates the strongpoint of the two methods of characters recognition. The experiment result is proved that the recognition rate of this algorithm achieve better result than that of TTM or TMTM singly used among the large templates gather and testing gather, especially, in the case of the intricate city roads and the case that the car's front lights are on at night. Although the algorithm proposed in this paper is better than other methods that are singly used, it still exists problems to be solved. For example, the real time of the algorithm needs to be further improved.

References

[1] Bian hu-qi. Pattern recognition(the second edition)[M]. Beijing: Press of University of Qinghua, 2000, 315~329

[2] Wu Yi-fei. Pattern recognition——theory, method and application[M]. Beijing: Press of University of Qinghua, 2003

[3] Liu Ji-Lin, Song Jia-Tao, Ding Li-Ya, Ma Hong-Qing, Li Pei-Hon. Vehicle License Plate Recognition System with High Performance. In: Acta Automatic Sinica, 2003 29(3): 457~465

[4] Hu Chang-bo, Feng Tao, Ma Song-de, Lu Han-qing. The Behavior Recognition based on the principle analysis[J]. Journal of Image and Graphics, 2000.10(5): 818-824

[5] Chen Yang, Chen Song-juan, Guo Ying-hui. Graphics Program and Image Disposal in MALAB 6.X[M]. Xi'an: Press of University of Electronic Science, 2002.10

[6] R.W.Swiniarski，A.Skowron. Rough Set methods in feature selection and recognition，Pattern Recognition Letters，2003.3(24): 833~849

[7] Yang Ying, Sheng Jing and Zhou Wei "The Monitoring Method of Driver's Fatigue Based on Neural Network", IEEE ICMA2007, Harebin, 2007, 3555-3559

[8] Ying Yang, Wei Zhou，Guang-yao Zhao "Driver's Face Image Recognition for Somber Surroundings based on computer vision", DCABES2007, Hubei**,** pp1125-1128

[9] Ying Yang, Dongliang Zhu "The Simulation Research of Control Arithmetic for Automobile ABS Based on MATLAB "International Conference on Mechanical Engineering and Mechanics 2007,Wuxi, 2058-2062

The Linearity Compensation Circuit Design Of Accurate Temperature Measurement

Anan Fang[1] Xiaoli Ye[2] Qingwu Lai An Zhao

Electronic Department , Nanchang University, Nanchang, Jiangxi 330031,China

Email:1 fanganan@ncu.edu.cn; 2 ye6d525@hotmail.com

Abstract

The larger and heavier of a device ,the bigger the thermal inertia after heating through industrial control ,which results in the generated temperature after heating lagging behind the desired one so as to affect its accuracy. This is an important factor that the heating accuracy of a large-sized device cannot be improved, and also a difficult problem that puzzles designers all the time. As to some thermocouple, software has been developed to compensate and modify it artificially. But it is difficult for this kind of software compensation to be suitable for all thermocouples since they vary in parameters. To find a better way of compensation, the author finally found the existing relationships among thermocouple, thermoelectric potential and temperature by studying thermocouple carefully. The real-time, linear and optimal design is realized by using operational amplifiers to compute high-order power series functions, and in this way the above mentioned problem has been solved and the accuracy of measurement has been improved extensively. The design totally meets the wide-ranged and accurate temperature control through the application to the injection molding machine and the result is satisfying.

Keywords： Thermocouple; Linearity Compensation Circuit; Thermoelectric Potential

1 Introduction

For a product formed by heating, its qualities such as whether the surface is luminous and smoothly or not, whether the inner density is good consistently or not have close relationship with the accuracy of the generated temperature. So how to control the temperature becomes pretty important. How to get the temperature after accurate measurement and accurate control is a major factor of increasing the high-class rate of the finished products .While in practice ,the weight of the inner heating parts weighs near up to several dozen kilograms and the volume of them is so huge, which makes the thermal inertia increases during heating. Plus the nonlinearity of thermocouple gets bigger with temperature in the wide-ranged temperature measurement, so the temperature measurement is not accurate as to make the temperature after heating lag behind the desired temperature.

To overcome the temperature drift caused by thermal inertia, we have designed a thermometer of high accuracy to make the controlled temperature is no more than ±0.5℃.On one hand, we design a thermocouple linearity compensation circuit to measure the temperature in actual wide range and ensure the accuracy of the temperature sampling. On the other hand, the high-accuracy thermometer we design can measure the temperature of every set point in one second and deliver it to SCM to deal with the data and computes advanced controlled variable by using the theory of fuzzy control to have control in advance, thus it can realize the purpose of accurate temperature measurement and the purpose of controlling heating.

2 The design of the main part of linear compensation circuit used for temperature measurement

The linear detection circuit used for thermocouple temperature measurement and its experimental data

In order to make the designed system to measure temperature accurately, the question as how to choose the amplifier properly to linearly amplify the measured value becomes quite important. To ensure the higher accuracy of the measured temperature, we have set a few points to measure the temperature. Low-power consumption CMOS integrated circuit CD4051 is chosen to constitute the analog multi-path conversion circuit to perform control during different time. The temperature of different point can be measured by choosing different thermocouple in different time, then the measured temperature of every point is delivered to CPU to be dealt with. X0……X7 can be chosen respectively by adding the control signal delivered by CPU to the control ends of A,B,C of the multiplexer CD4051.The thermoelectric potential of thermocouple outputs from the 3th pin of the multiplexer.(see Figure1)

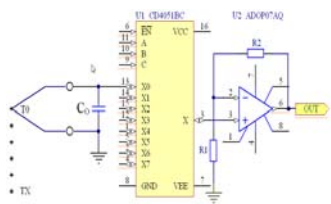


Figure1 the conversion circuit of the thermocouple (only illustrating one-path thermocouple)

In this figure, T0....TX are thermocouples, TX is the number of set thermocouples .Co is a smoothing capacitor, Because the thermoelectric potential of thermocouple is very small, if the leakage current of capacitance is big, there will be a drift voltage. Suppose the leakage current of Co is 0.1μA, the drift voltage is 0.1μA×1000=100μA, which is amplified 1000 times after the rear stage. Thus the detected signal cannot be guaranteed to be linear. A Tantalum capacitor whose leakage current is extremely small has to be chosen. To make the signal-to-noise ratio of the output signal of thermocouple high, we add a stage of preamplifier so as to ensure the output of thermocouple and avoid the interference of other signals. The line break protection circuit of thermocouple is composed of other components.

The linear amplification circuit used for thermocouple temperature measurement and its experimental data designing a perfect linear amplification circuit is a prerequisite for ensuring the measurement's accuracy. Because the output voltage of thermocouple is extremely low, only tens of μV/ ℃ ,the good-linearity ,low-offset ,high-sensitivity ,low-drift ,str ong-ability of common mode rejection operational amplifiers ADOP07 has been chosen to design & constitute the thermocouple amplification circuit ,which adopts three operational amplifiers to structure an inphase & parallel differential proportional circuit of high input impedance

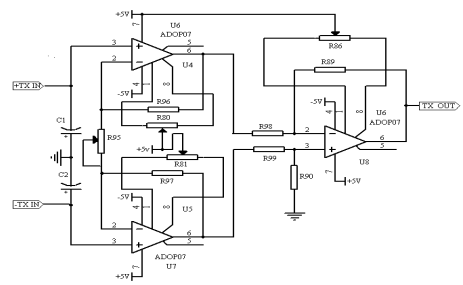


Figure2 the thermocouple amplification circuit

The range of gain adjustment of the circuit is very wide and all common mode voltages are put on two sides of resistor R95, thus the output voltage would not be influenced and they are irrelevant with feedback resistances R96, R97 .The deduced result is the amplified common mode rejection ratio is:

$$CM\frac{CM}{C}\frac{C}{C} \qquad (1)$$
$$RRE$$

Obviously, the ability of common mode rejection of this circuit is mainly determined by the consistency of common mode rejection ratio of the two operational amplifiers. Through actual detection the linearity of the amplification circuit can attain the anticipated goal after elaborate design and adjustment.(see table 1)

Table 1 the actual measurement result of the linear amplifier

IN (mV)	OUT (v)	IN (mV)	OUT (v)	IN (mV)	OUT (v)
0.1	0.58	1.0	6.01	1.5	9.03
0.2	1.18	0.6	3.57	1.1	6.53
0.3	1.81	0.7	4.19	1.2	7.16
0.4	2.38	0.8	4.77	1.3	7.78
0.5	2.96	0.9	5.41	1.4	8.36

The linearization circuit of thermocouple and its experimental data

The relationship between the thermoelectric potential of thermocouple and temperature is nonlinear ,the maximum nonlinearity error is －1% when the relationship curve of the thermoelectric potential of K-type thermocouple and temperature is 0 ℃～600℃. (See Fig.3a the characteristic curve of nonlinearization) .So by necessity, there is an optimization for linearization to ensure the accuracy of measurement. According to the way of polynomial linearization, suppose temperature is T, the coefficient of every item is a0，……，aN, then the thermocouple potential E of thermocouple can be expressed by

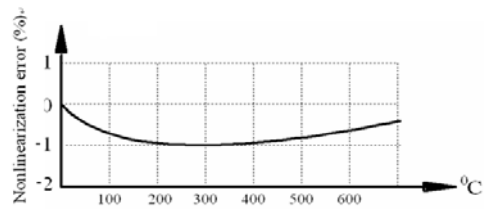
$$E = a_0 + a_1 T + a_2 T^2 + \cdots\cdots a_N T^N \qquad (2)$$

Through many times of analysis about the experimental data, if a high-order power series function can be acquired then there is a circuit of perfect linearity structured. We can get enough accuracy when it is second-order power for the K-type thermocouple. Because generally the approximate expression of the thermoelectric potential of K-type thermocouple is

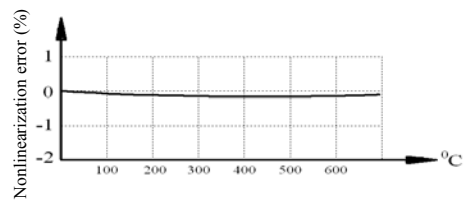
$$V_{out} = -0.776 + 24.9952 V_{IN} - 0.0347332 V_{IN}^2 (mv) \quad (3)$$

When temperature is 200℃,the thermoelectric potential of K-type thermocouple is checked out to be +8.137mV ,then we can get Vout=200.31mV according to the above formula,; when temperature is 400℃, the thermoelectric potential is +16.395mV,also we can get Vout=399.68mV. From above statement,200 ℃ corresponds to 200.31, 400℃ corresponds to 399.68,the differences only are 0.31mv and 0.32,respectively.So we know that the linear relationship between temperature and output voltage is perfect.

The key of a linearization circuit is square operation. We choose an integrated circuit chip AD538 which is the best one suitable for square operation whose accuracy is 0.5% and dynamic range is wide. There is a high-accuracy reference voltage source set in this chip. Three input ports VX（pin15），VY（pin10）and VZ（pin2）can form the functional relationship as Vout=VY(VZ/VX)mv. A circuit of square operation can be structured by AD538, ADOP07 and their external resistors and without other external components any more. The external resistors R91、R92、R131、RW2 of the operational amplifier U10 determine the gains of coefficients of first-order power and second-order power. R132、R133、R134、R135and R138 provide the bias voltage －0.776mV of above formula. Linearization improves greatly by elaborately adjusting the parameters of the components. The error of actual detection reduces from original －1% to 0.1~0.2% for nonlinear error. See Figure3 b the characteristic curve after linearization. The designed linearization circuit see Fig.4



a the characteristic curve before linearization



b the characteristic curve after linearization

Figure3 the characteristic curve before and after linearization

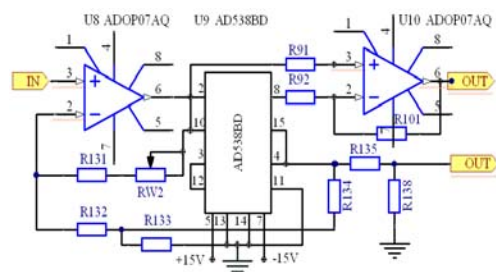


Figure4 the circuit of linearization

The cold terminal compensation circuit of thermocouple

The thermoelectric potential of thermocouple and the temperature of reference node (cold node) & measured node must keep constant. According to the standard, the thermoelectric potential of reference node is the one at 0℃. So when the temperature of reference node is not 0℃, a equivalent thermoelectric potential equal to the temperature of reference node should be added. We choose the temperature transducer AD592 that is suitable for K-type thermocouple to measure the temperature of reference node. If AD592 is provided a certain voltage then there is an output voltage proportional to absolute temperature acquired. Because the sensitivity of AD592 is 1μA/℃,the reference node of K-type thermocouple which is centralized as 25℃ and whose temperature coefficient is 40.44μV/℃ can be compensated. So when the ambient temperature is T , the output voltage of U12 can be altered by adjusting the resistor value of RW1 in the differential proportional amplifier U12 so as to compensate temperature. The circuit is illustrated in Fig.5. Q1 is emitter voltage follow, which is used to increase the driving ability of rear stages.

A/D conversion and SCM interface circuit

After the measured signal has been amplified, the result after the above dealing procedure is inputted to pin IN of the conversion integrated circuit 1C7135. MC1403 supplies a +2V reference comparative voltage to A/D converter to implement analog/ digital signal processing .The result is transferred to SCM W78E58B to perform such functions as control, display and reset etc. so as to complete a serial process of accurate measurement and intelligent control.

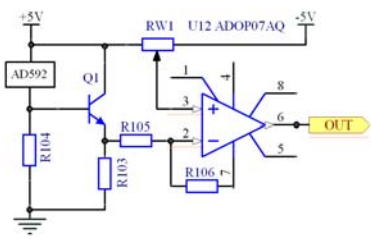


Figure 5 the cold terminal compensation circuit of thermocouple

3 The analysis of the results of experimental data

Through the elaborate design from the beginning of thermocouple linearity detection circuit used for temperature measurement to the linearization circuit of thermocouple and the adjustment to the parameters of the designed circuits, we can see that precision of the detected data and the accuracy of the controlled temperature are guaranteed in the whole wide range of temperature measurement after analyzing the results of the detected data, which reach the purpose of accurate temperature measurement and heating through control.

4 Conclusion

Thermocouple is temperature-dependent transducer made through See beck effect in physics; it is sintered with two different conductors. Its resistor is almost zero and the output thermoelectric potential is extremely small, only microvolt for every degree. All of these have made the measurement become very difficult. Besides the nonlinear factor of thermocouple itself, external electric field and other interferences would also affect the accuracy of measurement .For those reasons, it is difficult to design a temperature measurement system used for such large-sized device as injection molding machine whose accuracy needed is high .The circuit fits in with design demand through our efforts and it has been successfully used for the alteration of injection molding machine.

References

[1] Tomas w. Kerlin, Practical Thermocouple Thermometry , Instrumental Society of America, May 1, 1999

[2] Daniel D. Pollock, Thermocouples: Theory and Properties CRC; 1 edition September 17, 1991

[3] Astm Committee, E20 On Temperature Measure (Hardcover - April 1993) Manual on the Use of Thermocouples in Temperature Measurement/Pcn: 28-012093-40 (Astm Manual Series), ASTM International; Rev Sub edition ,April 1993

[4] Thomas Engel, Thermodynamics, Statistical Thermodynamics, and Kinetics (Hardcover), Prentice Hall; 1 edition ,February 19, 2005

[5] Terrell L. Hill, An Introduction to Statistical Thermody namics (Paperback), Dover Publications; New Ededition ,January 1, 1987

Prostate Ultrasound Image Segmentation Algorithm Based on Wavelet Transform*

Shubin Yang[1] Ying Tian[2]

1 School of Electrical & Information Eng., WIT, Wuhan, Hubei, China, 430073

Email: yshubin@sina.com

2 Department of Radiology, Xiangfan Central Hospital, Xiangfan, Hubei, China, 441021

Email: tianying610@163.com

Abstract

Because of the quality of the prostate ultrasound image, it is very difficult to define boundary of prostate. In order to get an accurate estimate result, a segmentation algorithm based on wavelet transform is designed. Using the dyadic wavelet transform, approximate coefficients and detailed coefficients at different scales can be generated and they determine the energy field. Then the energy field drives the discrete dynamic contour (DDC) model and the final contour is attained. Experiment proved that this algorithm can efficiently segment prostate ultrasound image well.

Keywords: Image segmentation; Wavelet transform; Discrete dynamic contour; Energy field

1 Introduction

Prostate ultrasound image segmentation is very important in the diagnosis and the treatment of prostate disease[1]. Image segmentation is the grouping of a set of pixels that share similar characteristics, such as intensity and texture. In segmenting the prostate, segmentation can be described as a procedure that separates the foreground of the image, which refers to the set of pixels that is mapped from structures inside the prostate, and the background. Image segmentation is a well-known problem in computer vision, and there are many well-established techniques for solving this problem. Three categories segmentation techniques widely used: threshold technique, edge-based technique and model-based technique[2]. The threshold technique is efficient in obtaining a segmentation if different regions of the image have a high contrast in their intensities or other measurable features. However, if two distributions have a significant overlapping region, its detection error will be large. Also, the threshold technique is sensitive to noise or intensity inhomogeneity. Edge-based techniques find the edge by locating regions where pixel values change significantly in a neighbourhood. Two widely used edge-detection techniques are the gradient-based procedure and the zero-crossing procedure. However, both methods are sensitive to undesirable fluctuations caused by noise. The deformable model consists of a closed parametric curve that moves under the influence of the internal force and the external force. Since the smoothness of the contour is constrained by the internal force, the deformable is not susceptible to local variation caused by noise. Also, disadvantage of the model is that it requires an initial contour, and the final contour is often sensitive to how the initial contour is chosen. The proposed segmentation algorithm combines the idea of the segmentation techniques introduced above. First, wavelet transform[3][4] with different wavelet bases and the same transform framework is processed on prostate image. Since the wavelet-based framework

* supported by Hubei Provincial Dept. of Education's Science Project

allows the choice of different wavelet bases that suit different purposes, its range of applications is wider than that of conventional gradient-based approach. After applying wavelet-based edge-detection technique on image, the modified DDC mode[5] is used to attain output contour, in which the external force is assigned according to wavelet-transformed image[6]. Then the final prostate contour is attained by DDC processing again using the output contour as the input initial contour. Experiment proved that this algorithm can not only decrease the effect of noise but also efficiently segment prostate ultrasound image well.

2 Dyadic wavelet transform

Suppose there is a 2-D smoothing function θ(x, y), there are two wavelets, which are the partial derivatives of θ(x, y) in x and y direction respectively:

$$\varphi^1(x,y) = -\frac{\partial \theta(x,y)}{\partial x} \quad \varphi^2(x,y) = -\frac{\partial \theta(x,y)}{\partial x}$$

andif $f \in L^2(R^2), then$

$$\begin{bmatrix} w^1_{2^j} f(x,y) \\ w^2_{2^j} f(x,y) \end{bmatrix} = 2^j \begin{bmatrix} \frac{\partial}{\partial x}\left\{ f * \theta_{2^j} \right\}(x,y) \\ \frac{\partial}{\partial y}\left\{ f * \theta_{2^j} \right\}(x,y) \end{bmatrix} = 2^j \nabla \left\{ f * \theta_{2^j} \right\}(x,y) \tag{1}$$

The modulus of this gradient vector is proportional to the wavelet transform modulus

$$M_{2^j} f(x,y) = \sqrt{\left| w^1_{2^j} f(x,y) \right|^2 + \left| w^2_{2^j} f(x,y) \right|^2} \tag{2}$$

M2j can be used to identify significant variations in different scales. The traditional pyramidal multiresolution representations are not translational invariant. However, Translational invariance is important in pattern recognition. When a pattern is translated, its representations should be translated but not modified. Patterns are more difficult to be identified if its representation depends on its location. In digital image processing, the discrete dyadic wavelet transform[7] is used to overcome this problem.

3 Energy Field Introduction

The energy field used in the algorithm is defined based on the approximated coefficients of the dyadic wavelet transform. In a typical prostate ultrasound image, the interior of the prostrate is relatively dark (i.e., the greyscale value is relatively low), while the exterior is relative bright (i.e., the greyscale value is relatively high). It can be expected that, in the smoothed image generated using dyadic wavelet transform, where the greyscale value of a pixel near the edge is computed by averaging the pixel value of the original image in its neighbourhood, the greyscale value near the edge is between the two extremes. Assuming the four initial points a user enters are reasonably close to the edge, the greyscale value of these four pixels on the smoothed image will be very useful in defining the boundary. Specifically, a set of pixels on the smoothed image that has greyscale values close to a weighted mean of these four values approximately define the boundary of the prostate. Therefore, for the DDC model, one can define an energy field that penalize the difference between this weight mean and the greyscale value of pixels on the smoothed smoothed image a j, represented by the set of approximate coefficients generated using dyadic wavelet transform at the scale 2^j. For example, one can define this energy field, denoted by E im1, as

$$E_{im1}[m,n] = -\left| a_j[m,n] - f\left(\left\{ a_j\left[x_{n,} y_n \right] : n = 1,2,3,4 \right\} \right) \right| \tag{3}$$

Where f (.) takes some kind of weighted mean of its argument and [xn , yn] denotes the coordinates of the user-defined initial point n. The negative of the absolute difference is taken because the DDC model moves to the peak of the energy field. To completely define the energy field E im1, it is required to choose the scale 2^j, at which the smoothed image is obtained, and how the weighted mean is defined. Recall that the scale 2^j should be large enough so that the small-scaled variations, mostly caused by noise, are suppressed, and it should be small enough so that the outline of the prostate is not too coarse. By experiment, it is found that j = 4 is appropriate for prostate boundary detection

problem. On the other hand, f could be defined so that it takes the arithmetic mean of its arguments.

4 Algorithm flow and experiment

4.1 Algorithm flow

Based on theory discussed above, Algorithm flow is designed below:

Step 1 input original ultrasound prostate image and four initial contour points are given;

Step 2 wavelet transform is processed using fast dyadic wavelet transform on the original input image and obtain M2j the modulus of the gradient vector at scale 2^j and scale approximation aj;

Step 3 initial contour is obtained from the four initial points given in step 1 using initialization method of DDC[4];

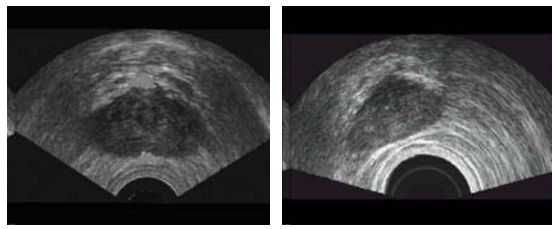
Step 4 according to the initial contour points, compute the E im1 with a j using the formula （3），then use the DDC mode with energy field E im1 to attain the output contour;

Step 5 modify the output contour attained in step 4 using DDC mode with energy field attained from M_2^j and then final contour is gotten.

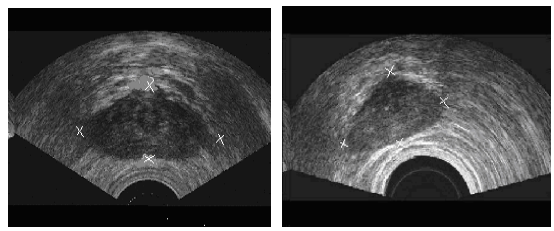
4.2 Experiment result

In order to test the performance of the designed algorithm，two typical prostate ultrasound images are processed with the algorithm. Fig.1 (a), (b) are original typical practical prostate ultrasound images. Fig.1 (a) is a ultrasound image with prostate lying in the central. Figure1 (b) is a ultrasound image with prostate lying in the left. Fig.1 (c), (d) are the images given the four initial points. Figure1(e), (f) are the final segmentation result images.

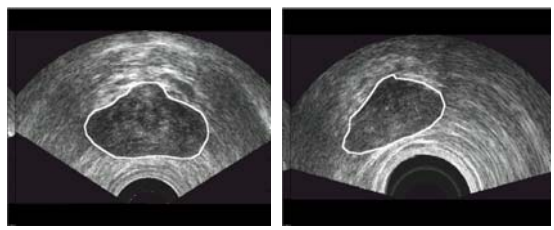
From the experiment results shows the algorithm can segment the difference position original prostate ultrasound images. Radiological doctor also thinks the segmentation result is correct.



(a) original image 1 (b) original image 2



(c) four initial points of (a) (d) four initial points of (b)



(e) segmentation result of (a) (f) segmentation result of (b)

5 Conclusion

The proposed algorithm approaches more intelligently. First of all, it can be observed that the deviation of the initial contour usually occurs near the top of the prostate. Also, it is observed that the intensity contrast between the interior and the exterior is often very high around the top of the prostate. An energy field based on the intensity of the smoothed image can be used to drive the DDC model. It is discovered that by using the spline interpolated contour as the initial contour of a DDC run that is driven by the new energy field, it is possible to generate a contour that is closer to the fact contour. It is expected that if this new contour is used as the initial contour of the gradient-modulus-based DDC run, a more accurate boundary can be defined. The boundary defined by the proposed segmentation algorithm is reasonably. A program that can perform this task is valuable to the medical industry, as the doctor can get important information about the prostate

automatically and quickly.

References

[1] National Cancer Institute, www.cancer.gov/cancertopics/ pdq/ prevention/prostate

[2] D. L. Pham, C. Xu, and J. L. Prince, Current methods in medical image segmentation, Annu. Rev. Biomed. Eng., Vol. 2, pp. 315-337, 2000

[3] QIN Qian-qin, YANG Zong-kai. Pratical wavelet analyse[M]. XI'AN:XI'AN jiaotong university publishing company,1994

[4] S. Mallat and S. Zhong, Characterization of Signals from Multiscale Edges, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 14, pp. 710-732, July 1992

[5] S. Lobregt and M. Viergever, A discrete dynamic contour model, IEEE Trans. on Medical Imaging, vol. 14, no. 1, pp.12-24, Mar. 1995

[6] P. Angel and C. Morris, Analyzing the Mallat wavelet transform to delineate contour and textural features, Computer Vision and Image Understanding, vol. 80, pp, 267-288, 2000

[7] S. Burgiss, R. Whitaker, and M. Abidi, Range image segmentation through Pattern analysis of the multiscale wavelet transform, Digital Signal Processing, vol. 8, pp. 267-276, 1998

Shu-bin Yang: male (1971-) associate professor, postgraduate director, high membership of Chinese Institute of Electronics, major in signal, image processing. Published more than 20 papers.