Journal of Computers

Contents

Special Issue: E-Service and Applications

Guest Editors: Jason C. Hung and Hsing-I Wang

Foreword of Special Issue on "E-Service and Applications"

Jason C. Hung and Hsing-I Wang

Department of Information Management, Overseas Chinese University, Taichung, Taiwan

E-mail: jhung@ocu.edu.tw, hsing@ocu.edu.tw

This special issue aims to bring together researchers and practitioners to engage in an in-depth discussion on various research and deployment issues regarding e-commerce and e-government. Finally, it creates the new value to information industries. Research in these fields is emergent and crucial, not only in academic and research community, but also in industrial field. In this decade, technological application researches are very popular. This special issue focuses on the state-of-the-art software technology related to these issues, including Resource management for cloud systems, Services integration and management in cloud systems, Optimal deployment configuration, Web services, Vision on novel advances of automated intelligent business agents and e-Commerce, Modeling and simulation of business processes, e-Supply Chains, e-Logistics, e-Procurement for both businesses and government , e-Participation, e-Governance, and e-Voting. Ten papers are included in this special issue.

In the first paper entitled "Measure the Performance of Reducing Digital Divide – the BSC and AHP Approach", Dr. Hsing-I Wang and Chien-Chih Yu discussed to adopt the analytical hierarchy process to prioritize the performance indicators which have been determined in the digital divide balance scorecard. The architecture proposed by the research would be applied to examine the merits of the strategies. This research applies the AHP method to the DDBSC framework. In this paper, the weight of each criterion in the AHP tree is calculated and examples of demonstrating how the indicators are applied to the real cases to determine the performance of the strategies for reducing digital divides are given.

In order to increase the accessibility of the healthcare support e-service, the second paper entitled "A Study on Healthcare Support e-Service Design for Senior Citizens" by Prof. Yuan-Chu Hwang tried to utilize multimedia streaming technology and video clips to present stroke–precaution knowledge for elderly senior citizens. This paper presents an e-service design on stroke-precaution for senior citizens, followed by its implementation and evaluation. With the healthcare support e-service, life quality for elderly senior citizens and their families could be significantly improved.

Belalem Ghalem, Bouamama Samah, and Larbi Sekhri proposed to extend and enrich the simulator CloudSim by auction algorithms inherited from GridSim simulator. In the third paper entitled "An effective economic management of resources in Cloud Computing", authors have the following contributions.

- Proposing a method or approach that will satisfy both the customer and supplier resource.
- Find a solution for customers who want a high quality of treatment but who do not have enough budgets, by the possibility of credit.
- Propose an implementation of auction club using a multi-agent system.
- Allow multi-criteria negotiation and intelligent system for decision support auction.

"A Study of Travel Agencies' Human Resources in Relation to Internet Marketing" written by Huang-Wei Su, Li-Tze Lee, Chiang Ku Fan, and Jason C. Hung proposed this study is to discover whether the travel agencies salespeople feel threatened by the development of Internet marketing. Also, this study seeks to reveal the specific personal thoughts of travel agency salespeople if they do feel threatened. This study provided suggestions for training programs or criteria of recruitment for the human resource directors in travel agencies.

In "KNN-DTW Based Missing Value Imputation for Microarray Time Series Data", Prof. Hui-Huang Hsu, Andy C. Yang, Ming-Da Lu present a novel method to impute missing values in microarray time-series data combining k-nearest neighbor (KNN) and dynamic time warping (DTW). They also analyze and implement several variants of DTW to further improve the efficiency and accuracy of their method. Experimental results showed that their method is more accurate compared with existing missing value imputation methods on real microarray time series datasets.

Prof. Benjamin Penyang Liao pointed out t this study analyses the human relationship starting from the constructing material for the connection line. In the sixth paper entitled "Insight of Value-oriented Human Relationship", a human relationship influence spectrum with seven phases is discussed. This spectrum could provide a good insight of human relationships. Finally, the influence phase transition, influence scope, and characteristics of human relationship influence spectrum are discussed in this article.

"Development of Semantic-CBR Framework for Virtual Enterprises in Project Management" written by Prof. Chouyin Hsu. Semantic-CBR framework is proposed for improving the communication and

integrating the project experiences for virtual enterprises. OWL and RDF are applied for the implementation of the semantic structure of the framework. Moreover, case-based reasoning is utilized for selecting the related project experiences according to the user requirement. Semantic-CBR framework is proposed for improving the communication and integrating the project experiences for virtual enterprises. OWL and RDF are applied for the implementation of the semantic structure of the framework. Moreover, case-based reasoning is utilized for selecting the related project experiences according to the user requirement.

In "A Diffusion of Innovations Approach to Investigate the RFID Adoption in Taiwan Logistics Industry", Prof. Yu-Bing Wang, Kuan-Yu Lin, Lily Chang , and Jason C. Hung purpose this study investigating how RFID technology was implemented and adopted in Taiwan's logistics industry. This study also aimed to determine the concern factors of adopting the RFID system into current company management systems. RFID provider companies should try their best to promote this technology's attributes of innovations: the relative advantages of RFID system to their potential customers, such as the costs reduction, profit capability and improvement of operation performance, design RFID system more friendly and more compatibility for quick adoption; reduce RFID system's difficulty and make it easy to learn and use. RFID providers can provide a trial-able project toward their potential customers.

In the ninth paper entitled "Incremental Mining of Across-streams Sequential Patterns in Multiple Data Streams", Shih-Yang Yang, Ching-Ming Chao, Po-Zung Chen and Chu-Hao Sun propose the IAspam algorithm that not only can handle a set of items at a time but also can incrementally mine across-streams sequential patterns. In the process, stream data are converted into bitmap representation for mining. Experimental results show that the IAspam algorithm is effective in execution time when processing large amounts of stream data.

In the tenth paper, "The Effect of Online Training on Employee's Performance" proposed by Prof. Shu Ching Yang and Chin Hung Lin is aiming at determining the effect of online distance-learning instruction on employees' learning achievement those who took a training program of online distance-learning instruction and those who did not. Since many factors will affect employees' learning achievement, this study also explored the effect of self-efficacy, gender, computer experience, and socioeconomic status, on the learning achievement of employees in a Taiwanese manufacturing company. Additionally, this study investigated students' satisfaction with online distance-learning instruction. The major findings were that learning achievement was similar for online distance learning and traditional face-to-face instruction.

Jason C. Hung is an Associate Professor of Dept. of Information Management, Overseas Chinese University, Taiwan. His research interests include Multimedia Computing and Networking, Distance Learning, E-Commerce, and Agent Technology. From 1999 to date, he was a part time faculty of the Computer Science and Information Engineering Department at Tamkang University. Dr. Hung received his BS and MS degrees in Computer Science and Information Engineering from Tamkang University, in 1996 and 1998, respectively. He also received his Ph.D. in Computer Science and Information Engineering from Tamkang University in 2001. Dr. Hung participated in many international academic activities, including the organization of many international conferences. He is the founder and Workshop chair of International Workshop on Mobile Systems, E-commerce, and Agent Technology. He is also the Associate Editor of the International Journal of Distance Education Technologies, published by Idea Group Publishing, USA. The contact email is jhung@ocu.edu.tw.

Hsing-I Wang is an Associate Professor in the Department of Information Management, Overseas Chinese University, Taichung, Taiwan. She holds a PhD degree in management information systems from the National ChengChi University, Taipei, Taiwan. Her previous research areas include e-government, object oriented system design, applications of artificial neural network, knowledge management, and business managerial issues.

Measure the Performance of Reducing Digital Divide – the BSC and AHP Approach

Hsing-I Wang
Dept. of Information Management
Overseas Chinese University, Taichung, Taiwan,
Email: hsing@ocu.edu.tw

Chien-Chih Yu
Dept. of Management Information Systems
National ChengChi University, Taipei, Taiwan
Email: ccyu@nccu.edu.tw

Abstract—Digital divides are the byproducts of the development of information technologies and digitalization. What concern most of the governments and the international organizations are the disturbance of the advancement of national competitiveness as well as the improvement of human lives caused by the presence of digital divides.
Countries world wide have proposed numerous of strategies to reduce digital divides. However, the absence of the follow-up of the performance of the strategies forms another issue in reducing digital divides. Previous researches built digital divide balance scorecard to realize the objectives as well as to improve the performance of the strategies. This research adopts the analytical hierarchy process to prioritize the performance indicators which have been determined in the digital divide balance scorecard. The architecture proposed by the research would be applied to examine the merits of the strategies.

Index Terms—digital divide, performance measurement, analytical hierarchy process, balance scorecard

I. INTRODUCTION

Digital divides (DD) are the byproducts of the fast development of information technologies and digitalization. The existence of DD not only implies the vanishment of opportunities, it also discloses the existence of poverty, the lack of fundamental literacy, as well as some serious social problems [1]. Countries world wide have put a lot of efforts and proposed numerous of strategies to reduce digital divides. Have the phenomenon disappeared? Take Taiwan for an instance, "Ref. [2]" reported that digital divides could be found between metropolitan and rural area, and between different levels of educations in Taiwan. Evidences indicate that, after all the attempts, reducing

digital divides remain the most important issue and concern in information epoch.

We believe that reducing digital divides is indeed a hardship task, the point is that if the governments have taken appropriate actions and effective strategies. And more important, is whether the performances have been reasonably evaluated. "Ref. [3]" tried to adopt the balance scorecard approach to measure the performances of reducing digital divide. In their researches, the gaps between performances and strategies were located [4], a strategy map was designed [5], and the framework of digital divide balance scorecard (DD-BSC) was developed [6]. However, the indicators have not been validated, neither have the implication of the measurements been demonstrated. This research proposes the analytical hierarchy process (AHP) method to fill up the absences.

This paper is organized as follows: section 2 briefly introduces the DD-BSC framework proposed by [6] and reviews the strategies of reducing digital divides that have been initiated, proposed as well as implemented by governments world wide. Section 3 introduces the AHP method. Section 4 states the research method and the results of the research. Section 5 discusses the findings of this research and the final section gives conclusions of the paper.

II. THE REVIEW OF STRATEGIES OF REDUCING DIGITAL DIVIDES AND DD-BSC FRAMEWORK

Previous digital divides researches can be classified into four categories. The first category observes the phenomenon of digital divide of a single country or compares the degree of digital divide between countries. The second category focuses mainly on the penetration of information technologies. The researches of the first two categories conclude that the opportunities of accessing or using of information technologies explain the major portion of the existence of digital divides. The third category applies different theories to explain the

existence and the causes of digital divides. "Ref. [7]" proposed that digital divides were impossible to be entirely eliminated if they were explained with Diffusion of Innovations, Increasing Knowledge Gap, and Adaptive Structuration Theory. Finally, the fourth category explores digital divides with various research models. Research findings of [8], for example, proved that economic, social, ethno-linguistic and infrastructure were the major constructs that cause digital divides.

"Ref. [3]" integrated previous researches and proposed that the phenomenon of digital divides should be examined from four dimensions including ICT, equal opportunities, information society and national competitiveness respectively. "Ref. [6]" further presented the DD-BSC framework (as shown in figure 1) so that the strategies for reducing digital divides were further classified into four balanced scorecard perspectives including beneficiaries, governmental functions and processes, nation-wide learning and growth as well as public finance.
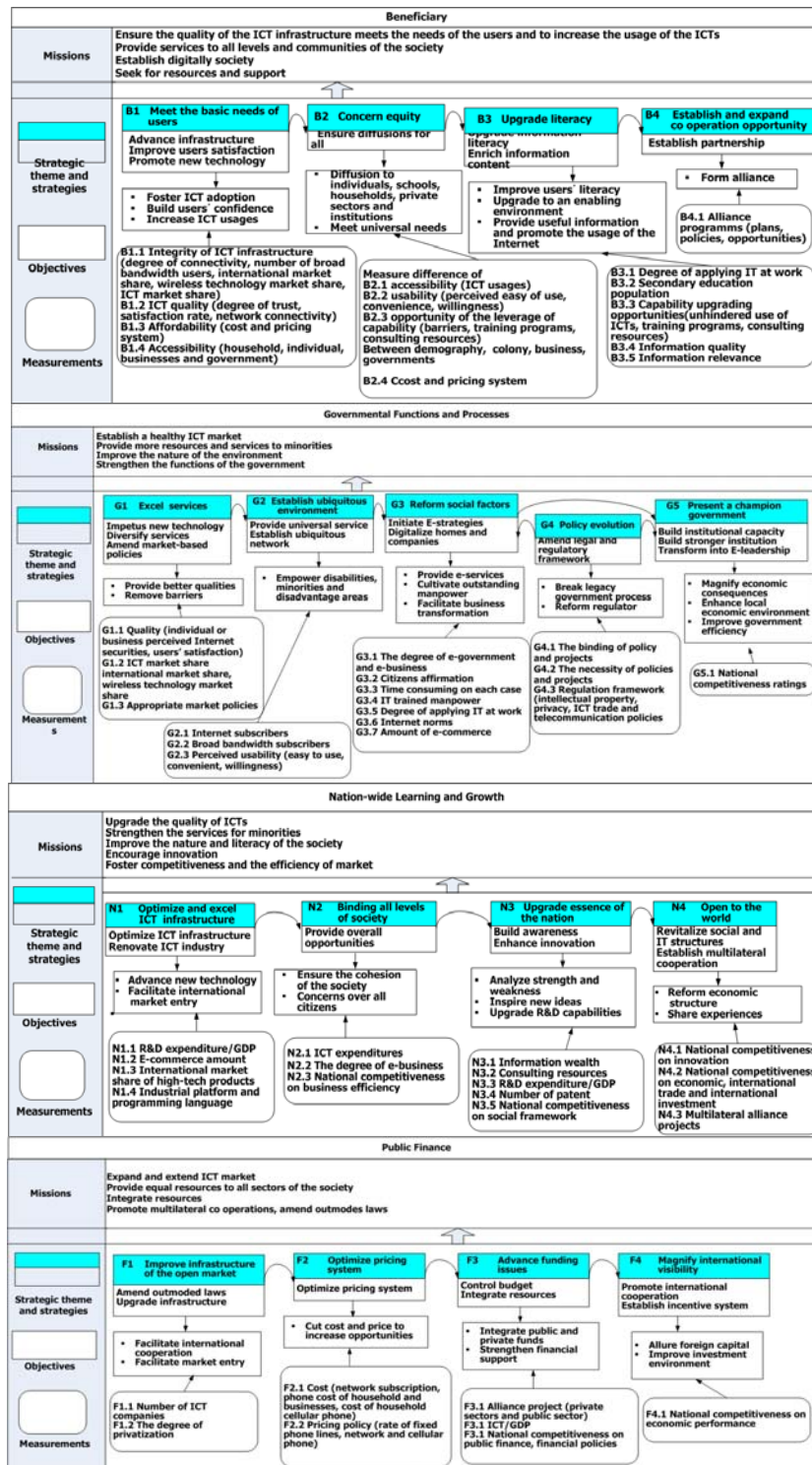

Fig. 1: DD-BSC framework [6]

In the DD-BSC framework, each perspective has its missions. The strategies that are necessary to fulfill the missions are defined along with the strategic objectives. Notice that the measurements are not only specified in the framework, the relationships between the measurements and the strategies are also established. The approach that [6] adopted would help the governments to correctly and efficiently locate the less productive strategies, and furthermore, to take proper actions to recover the drawbacks in time.

III. THE AHP

In the literatures, AHP has been applied to numerous decision making problems. The AHP, developed by [9], is based on a pair-wise comparison of the criteria. The criteria could be prioritized and ranked based on the preferences given by the involved parties. By applying AHP, the interacted effects could be eliminated efficiently. The steps of AHP analysis are firstly, building the hierarchical architecture of the criteria, constructing the pair wised matrix, computing the eigenvalues and eigenvector, and finally validating the consistence of the AHP tree. Figure 2 shows the basic architecture of an AHP tree.
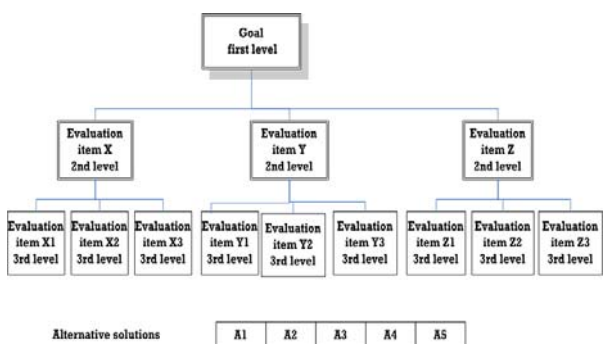
Fig. 2: The AHP structure

The purposes of consistence check are to determine the rationality of the involved parties and to examine the consistence of the entire structure of AHP tree. The consistence is determined by consistency ratio (C.R.), which is the quotient of consistence index (C.I.) and random index (R.I).

$$CR. = \frac{CI}{RI}$$

"Ref. [10]" suggested that RI is closely related to the number of levels of the AHP tree and the number of criteria at each level is less than 7 [11]. The mappings of RI and the number of levels in an AHP tree are given in table 1. "Ref. [11]" also suggested that C.R. should be less than 0.1 to guarantee the consistence of an AHP tree.

The AHP tree is then transformed into a pair wised matrix and the respondent will prioritize each pair with respect to the relative importance of the two criteria. The options for the responses are presented on the AHP

Table 1: RI value for different number of levels

No. of levels	1、2	3	4	5	6	7
RI.	0.00	0.58	0.90	1.12	1.24	1.32
No. of levels		8	9	10	11	12
RI.		1.41	1.45	1.49	1.51	1.48

Resource：[10]

scale from 1 to 9. The odd numbers represent equal important, weak important, essential important, very strong important and absolute important, and the even numbers represent the mid values between two odd numbers [12, 13].

IV. RESEARCH METHOD AND RESULTS

This research adopts the four DD dimensions and the BSC perspective proposed by [6]. The criteria and indicators used in the DD-BSC framework are applied to construct the AHP tree of this research. Figure 3 shows the AHP tree of this research.

A questionnaire was designed in which the criteria of each level were compared in pair. Fourteen experts who are familiar with the subject of digital divide were invited. They include three scholars from universities; two from the digital divide group of the Research Development and Evaluation Commission of Taiwan; four specialists from industries, and four governmental officials. The respondents were required to complete the questionnaires according to their experiences or expertise on the issue of digital divides. Thirteen questionnaires were collected; the response rate was 92%. Expert Choice 2000 was applied to analyze the data. In the validation of the consistence of the answers, one questionnaire showed high CR value 0.54, which meant the comparisons between indicators were inconsistent and the answers were ignored. The remaining 12 responses were analyzed. Table 2 shows the results of the main construct of the AHP tree. Table 2 indicates that the CR value is less than 0.1 and meets the requirement of consistence of AHP.

Table 2: Results of the main level

criteria	weight	rank
ICT diffusion	0.335	1
Equal opportunities	0.283	2
Information society	0.208	3
National competitiveness	0.173	4
CI=0.01, n=4, RI=0.9, CR=0.01		
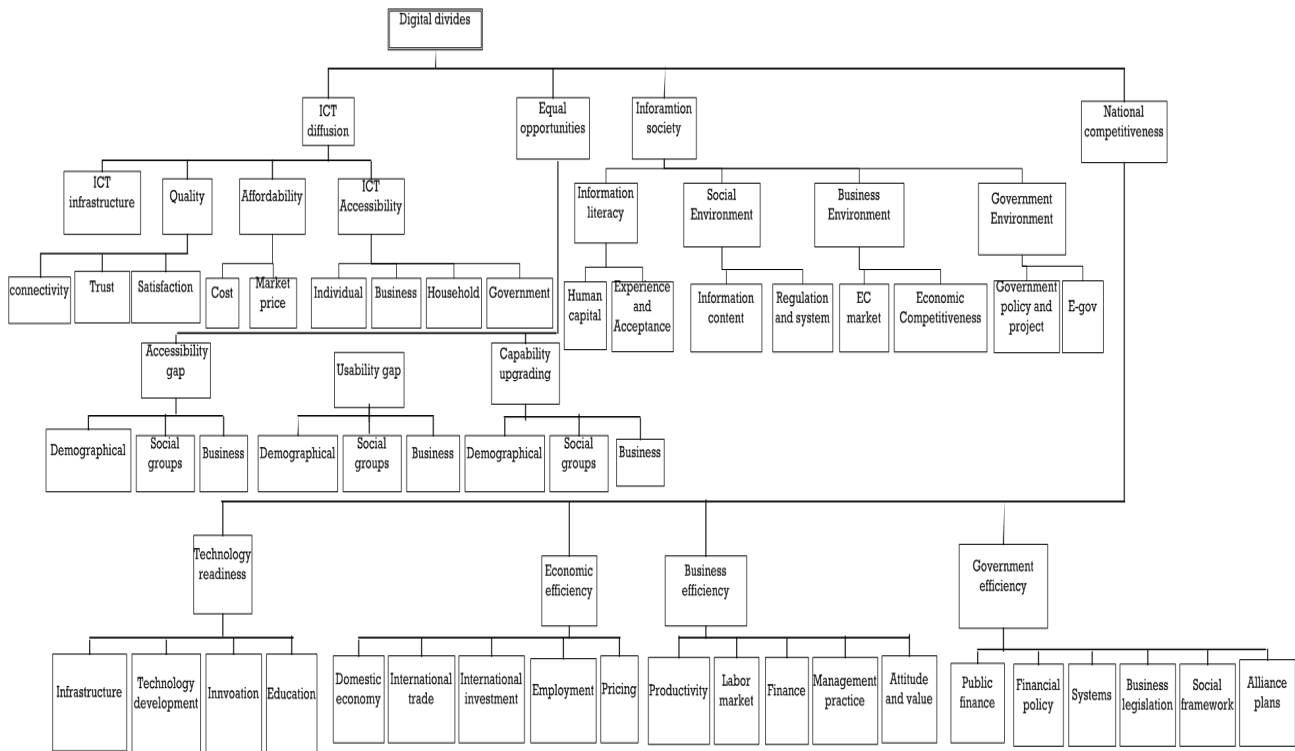
Fig. 3: The AHP tree of this research

To validate the consistence of the entire AHP architecture of this research, this research further calculates the consistent index hierarchy (CIH), random index hierarch (RIH) and consistent ratio hierarchy (CRH) as follows.

CIH = 0.01 * 1 + 0.01 * 0.335 + 0.00* 0.283+ 0.02 * 0.208+ 0.00 * 0.173
= 0.01751

RIH = 0.9 * 1 + 0.9 * 0.335 + 0.58 * 0.283+ 0.9 * 0.208+ 0.09 * 0.173 = 1.70854

CRH = CIH / RIH = 0.010

Again, the CRH value is acceptable and assures the consistence of the design of this research.

Tables 3 give the weights of the criteria in each level of the AHP tree.

V. THE USE OF THE AHP RESULTS

To demonstrate how to apply the results of this research to the real cases, this research applies the weights of the ICT individual accessibility indicators to Taiwan. Table 4 collects the real data of 2008 and 2009 of Taiwan [14, 15, 16]. By multiplying each data with the weights, the score of indicator "individual accessibility" of 2008 is 81.83 while in 2009, the score is 82.88. The score of ICT accessibility is 27.98 in 2008 and 28.34 in 2009. For a country, if the scores are lower than the expected values, the government would have an easier way to locate the less efficient strategies and make appropriate actions. We further notice that table 4 shows that "individual accessibility" is the most important indicator (with the highest weight) on the

"ICT accessibility" level, which means the government needs to take the gap between individuals more seriously in order to gain higher scores.

The second example examines the performance of national competitiveness in Taiwan. According to the national competitiveness report of 2007 [17], the ranks of each category of Taiwan are listed in table 5. Based on the results of this research, the score of national competitiveness of Taiwan in 2007 would be 60.55 (71.1*0.159+58.4*0.251+69.3*0.287+48.5*0.303). The contribution of national competitiveness to reduce digital divide is 10.48 (60.55*0.173). When compare with the score of 2006 (table 6), we notice that the changes are not surprisingly. The government would then have to review the achievements of the strategies more cautiously.

VI. CONCLUSIONS

This research applies the AHP method to the DD-BSC framework that [6] proposed. In this paper, the weight of each criterion in the AHP tree is calculated and examples of demonstrating how the indicators are applied to the real cases to determine the performance of the strategies for reducing digital divides are given. The contributions of this research are summarized as follows:
1. The importance of each indicator of the performance of strategies for reducing digital divides is determined.
2. The AHP method is adopted to calculate the weights of the indicators. Since the AHP considers the relative importance of the criteria evaluated by

specialist in the domain, the bias as well as the interactive effects of criteria would be minimized.

3. In the real situation, the scores gained from the AHP architecture of this research would be compared with the expecting value that governments have set. The comparisons would help the governments firstly, to locate the gap between the outcomes and objectives, secondly, to review the effectiveness of strategies, and finally, to make appropriate actions to mend the gaps.

Nevertheless, the questionnaires were distributed in Taiwan; the importance of the indicators may be different if the AHP structure is evaluated in other countries. Fortunately, the design of this research is flexible and can be applied to all countries even if the weights are different or the importance of an indicator is changed.

Moreover, during the research, we notice that a large number of indicators have been neglected and were not included in the investigations of digital divides or related issues in Taiwan. As a result, the performance of strategies for reducing digital divides would be difficult to be observed. The same situation may exist in many other countries. This research, however, provides the reference for governments to follow.

Further works of this research would look for international cooperation to construct a more general measurement for countries to evaluate the situation of digital divides and to examine the performance of digital divide strategies.

Table 3: Results of AHP analysis

Level 0	Level 1	CR	Level 2	CR
ICT diffusion (0.335) CI=0.00 n=5 RI=1.12 CR=0.00	ICT infrastructure(0.184)	0.008		
	Quality(0.121)	0.008	Connectivity(0.441)	0.011
			Trusty(0.314)	0
			Users satisfaction (0.275)	0
	Affordability(0.446)	0	Cost (0.588)	0.022
			Market pricing (0.412)	0
	ICT accessibility (0.249)	0.008	Individual accessibility (0.338)	0.011
			Business accessibility (0.189)	0.02
			Household accessibility (0.265)	0.03
			Government accessibility (0.208)	0
Equal Opportunities (0.283) CI=0.00 n=3 RI=0.58 CR=0.00	Accessibility gap (0.183)	0.02	Demographical (0.188)	0.01
			Social groups (0.188)	0.01
			Business (0.188)	0
	Usability gap (0.304)	0.03	Demographical (0.188)	0
			Social groups (0.188)	0
			Business (0.188)	0
	Capability upgrading (0.513)	0	Demographical (0.188)	0
			Social groups (0.188)	0
			Business (0.188)	0
Information society (0.208) CI=0.01 n=4 RI=0.9 CR=0.01	Information literacy (0.395)	0	Human capital (0.463)	0.05
			Experience and Acceptance (0.537)	0.03
	Social Environment (0.200)	0	Information content (0.574)	0
			Regulation and system (0.426)	0.02
	Business Environment (0.186)	0	EC market (0.463)	0.05
			Economic Competitiveness (0.537)	0.05
	Government Environment (0.219)	0	Government policy and project (0.708)	0.05
			E-government (0.292)	0.09

Level 0	Level 1	CR	Level 2	CR
National competitiveness (0.173) CI=0.00 n=5 RI=1.120 CR=0.00	Technology readiness (0.159)	0	Infrastructure(0.241)	NA
			Technology development(0.132)	
			Innovation (0.206)	
			Education (0.421)	
	Economic efficiency (0.251)	0.01	Domestic economy (0.266)	
			International trade (0.146)	
			International investment (0.153)	
			Employment (0.257)	
			Pricing (0.178)	
	Business efficiency (0.287)	0.01	Productivity(0.22)	NA
			Labor market (0.165)	
			Finance (0.141)	
			Management practice (0.197)	
			Attitude and value (0.278)	
	Government efficiency (0.303)	0.001	Public finance (0.098)	NA
			Financial policy (0.177)	
			Systems (0.267)	
			Business legislation (0.137)	
			Social framework (0.228)	
			Alliance plans (0.094)	

*the second level of national competitiveness are the indicators, no CR values are applied.

Table 4: The application of weights in ICT accessibility (the case of Taiwan)

Level 1	Level 2	Indicators	Weight (A)	2008data (B)	A*B	2009data (C)	A*C
ICT accessibility (0.257)	Individual accessibility (0.342)	Number of households with fixed telephone lines	0.113	0.98	0.1107	0.98	0.1107
		Population of mobile phones	0.149	1.05	0.1565	1.11	0.1654
		Population of network	0.448	0.68	0.3046	0.69	0.3091
		Owns personal computers	0.290	0.85	0.2465	0.84	0.2436

*: data source [15, 16]

Table 5: The application of weights in National Competitiveness (the case of Taiwan)

Level 1	Indicators	weight	order	2007rank	1-index*	Weighted value	Score
Technology readiness (0.159)	Infrastructure	0.251	2	22	0.6	0.151	71.1
	Technology development	0.133	4	6	0.89	0.118	
	Innovation	0.191	3	10	0.82	0.156	
	Education	0.425	1	18	0.67	0.286	
Economic	Domestic economy	0.266	1	23	0.58	0.155	58.4

Level 1	Indicators	weight	order	2007rank	1-index*	Weighted value	Score
efficiency (0.251)	International trade	0.147	5	16	0.71	0.104	
	International investment	0.148	4	42	0.24	0.035	
	Employment	0.259	2	22	0.6	0.155	
	Price	0.18	3	14	0.75	0.134	
Business efficiency (0.287)	Labor	0.216	2	11	0.8	0.173	
	Labor market	0.158	4	11	0.8	0.126	
	Finance	0.136	5	14	0.75	0.101	69.3
	Management	0.201	3	24	0.56	0.113	
	Attitude value	0.289	1	21	0.62	0.179	
Government efficiency (0.303)	Public finance	0.094	6	21	0.62	0.058	
	Financial policy	0.167	3	3	0.95	0.158	
	Regulation system	0.257	1	29	0.47	0.121	48.5
	Business legislation	0.136	4	28	0.49	0.067	
	Societal framework	0.246	2	37	0.33	0.081	
	Alliance	0.099	5	NA	NA	NA	

* : index = 2007 rank / 55 (total countries investigated)

Table 6: Comparisons of National Competitiveness of Taiwan and other countries 2006-2007

category year country	Overall rank		Technology readiness		Economic efficiency		Business efficiency		Government efficiency		overall		Contribution to DD	
	2006	2007	2006	2007	2006	2007	2006	2007	2006	2007	2006	2007	2006	2007
Taiwan	18	13	74.4	71.10	49.6	58.4	77.8	69.3	44.1	48.5	59.96	60.55	10.37	10.48
Denmark	5	6	82.7	88.5	51.8	61.7	87.2	86.4	74.7	74	73.81	76.8	12.77	13.29
Finland	17	15	90.4	77.9	43.3	45.1	76.1	64.7	74.9	62.0	70.39	61.06	12.18	10.56
Japan	24	22	74.2	71.6	64.7	60.3	54.3	49.7	44.2	36.8	56.99	51.93	9.86	8.98
Korea	29	31	53.7	65.9	44.7	40.6	31.0	31.0	26.8	34.2	36.79	39.94	6.36	6.91
Norway	13	11	81.2	80.5	58.2	61.4	69.4	71.7	67.5	62.8	67.89	67.83	11.74	11.73
Singapore	2	2	84.9	86.0	86.7	80.7	80.6	83.9	77.2	79.1	81.76	82.00	14.14	14.19
Sweden	9	9	84.0	81.7	54.0	59.6	75.1	77.7	54.7	63.8	65.02	69.56	11.25	12.03
United States	1	1	91.0	88.5	88.1	85.8	80.0	75.7	65.5	58.4	79.4	75.00	13.74	12.98
China	15	17	43.7	50.6	87.2	90.9	53.5	56.1	55.3	57.9	60.94	64.5	10.54	11.16
India	27	29	27.8	23.9	80.1	69.8	67.7	67.9	37.8	35.6	55.59	51.6	9.62	8.93

Data source: [17, 18]

REFERENCES

[1]   Bridges.org, R.: 'Spanning_the_Digital_Divide - Understanding and Tackling the Issues,' 2000. 1http://www.bridges.org/spanning/pdf/spanning_the_digital_divide.pdf.

[2]   Yu, C. C and Wang, H.I.: 'Digital Divide in Taiwan: Evidences, Comparisons, and Strategies', Electronic Government, an International Journal, 2004, 1(2), 179-197.

[3]   Yu, C. C. and Wang, H.I.: 'Measuring the Performance of Digital Divide Strategies: The Balanced Scorecard Approach', EGOV 2005, R.

Traunmuller (Eds.), Lecture Notes in Computer Science, 2005, 3591(0), 151-162.

[4] Yu, C. C. and Wang, H.I.: 'Towards Refining Digital Divide Strategies via Strategy Gap Analysis – the Method and Case Study', Journal of International Management Studies, 2008, 3(2), 28-36.

[5] Yu, C. C. and Wang, H.I.: 'Developing Strategy Maps for the Formulation of Digital Divides Strategies', Proceedings of the 7th International Conference on Electronic Business, Taipei, Taiwan, 2007.

[6] Yu, C. C. and Wang, H.I.: 'Strategy Mapping in the Process of Formulating Digital Divide Strategies,' Electronic Government, An International Journal, 2009, 6(2), 143–161.

[7] Manson, S. M. and Hacker, K. L.: 'Applying Communication Theory to Digital Divide Research', IT & Society, 2003, 1(5): 40-55.

[8] Bagchi, K.: 'Factors Contributing to Global Digital Divide: Some Empirical Results', Journal of Global Information Technology Management, 2005, 8(3): 47-65.

[9] Saaty, T. L.: 'A Analytic Hierarchy Process', New York, McGraw Hill, 1980.

[10] Saaty, T. L.: 'Decision Making for Leaders - The Analytic Hierarchy Process for Decisions in A Complex World', Resource Allocation, Pittsburgh, RWS Publications, 1988.

[11] Saaty, T. L.: 'Multi Criteria Decision Making: The Analytic Hierarchy Process - Planning, Priority Setting', Resource Allocation, Pittsburgh, RWS Publications, 1990.

[12] Deng, J.Y. and Tseng, K H.: 'AHP Theory and Its Applications - 1', Journal of Chinese Statistics, 1989, 27(6): 13707-13724.

[13] Deng, J.Y. and Tseng, K H.: 'AHP Theory and Its Applications - 2', Journal of Chinese Statistics, 1989, 27(7):13767-13786.

[14] Directorate – General of Budget, Accounting, and Statistics, Executive Yuan, Taiwan. http://www.dgbas.gov.tw.

[15] RDEC, '2008 Digital Divide Report', Research, Development and Evaluation Commission, Executive Yuan, Taiwan. http://www.rdec.gov.tw/public/Attachment/92313 455871.pdf.

[16] RDEC, '2009 Digital Divide Report', Research, Development and Evaluation Commission, Executive Yuan, Taiwan. http://www.rdec.gov.tw/public/Attachment/91215 951771.pdf.

[17] IMD: 'The World Competitiveness Yearbook,' 2007 International Institute for Management Development, Lausanne.

[18] IMD: 'The World Competitiveness Yearbook,' 2006 International Institute for Management Development, Lausanne.

A Study on Healthcare Support e-Service Design for Senior Citizens

Yuan-Chu Hwang

Department of Information Management, National United University, Taiwan

Email: ychwang@nuu.edu.tw

Abstract—The urgent needs of healthcare assistant for elderly senior citizens increase rapidly. This paper presents an e-service design on stroke-precaution for senior citizens, followed by its implementation and evaluation. The healthcare support e-service design focuses on the aging problems including the degeneracy of memory and vision as well as the recession of attention. These physiological problems become the barriers for senior citizens to learn healthcare knowledge effectively. In order to help senior citizens to obtain appropriate knowledge for stroke-precaution, this study provides a healthcare support e-service and its interface design that specifically addresses the physiological issues for elderly.

In order to increase the accessibility of the healthcare support e-service, we utilize multimedia streaming technology and video clips to present stroke–precaution knowledge for elderly senior citizens. Different from traditional e-learning system provide massive information for user, this study provide a personalized e-service design which only provide related healthcare information according to user's health condition. The healthcare support e-service is designed for senior citizens that provide appropriate user interface addressing their situations. Analysis of learnability, perceived system performance, memoraility and satisfactions are based on different ages and gender. Analytical result indicates our service design could help elderly to leap over the physiological problems they encountered. With the healthcare support e-service, life quality for elderly senior citizens and their families could be significantly improved.

Index Terms—e-Service Design, Healthcare Support e-Service (HSS), Elderly Senior Citizens, Stroke-Precaution

I. INTRODUCTION

Stroke is one of deadly illness in Taiwan. There are around 50 thousand stroke patients per year. The elderly population has increased to 10.4% of the Taiwan society. In order to improve the life quality for elderly, the urgent needs of healthy knowledge for elderly to prevent stroke should be take into consideration.

The lack of professional caring service for elderly is one of the most urgent problems. Usually, the elderly is not well take cared in their family. There are 17% of elderly could not get any help in time when they encounter emergent events. The institutional caring service could provide many professional services for elderly. However, the institutional caring service is not available to all elderly owing to the cost is not affordable for every family. There should be alternative solutions to help elderly obtain necessary information to improve their life quality.

Adopting e-service and ICT (information and communication technologies) in healthy caring service is one of the solutions to broaden the knowledge channel for elderly. Due to the physiological recession from aging, traditional system design may not provide suitable assistant for elderly. There should be a service design that is specifically addressing on the problems that elderly encountered. In this study, a healthcare support e-service (denoted as HSS) design on stroke-precaution for elderly is proposed. The stroke-precaution HSS provides personalized healthcare information for elderly to learn and enrich their knowledge. The user interface design could help elderly to overcome the aging problems. The personalized recommendation service design could provide useful healthcare information for senior citizens. The provision of numerous multilingual video clips of stroke-precaution knowledge could support those illiterate senior citizens conquer the reading problem.

Remain sections of this paper are organized as follows. Section 2 presents the related guidelines for service design to elderly. Illustration of the stroke-precaution Healthcare Support e-Service (HSS) is presented in section 3. The evaluation methods and the analytical results are introduced in section 4. Finally, the conclusions and future research directions are presented in section 5.

II. SERVICE DESIGN GUIDELINES FOR ELDERLY

The service design guidelines for elderly can be unfolded into three parts. Firstly, the service needs for elderly senior citizens should be discovered. Secondly, some appropriate design guidelines for elderly senior citizens should be considered for providing the stroke-precaution HSS. Thirdly, the evaluation method of e-service design for elderly senior citizens will be discussed. In this section, these issues will be elaborated.

A. The Service Needs For Elderly

According to Morrell and Echt (1996), the comprehension ability of word, working memory and space recognition ability will decrease due to the aging issue. [12] Communicating with senior citizens with more

simple language, and provide assistant actively, these are few tips for helping elderly senior citizens to relieve their loadings. In order to provide service for elderly senior citizens, it is important that we consider that their physical conditions are declining.

There are two major issues for considering the service needs of elderly. The social issue and the knowledge issue. The social issue is the lack of labor power to take care the elderly. The knowledge issue is the lack of appropriate channel to obtain healthcare information and knowledge for elderly to take care themselves.

The family structure has changed due to the birthrate decrease. The whole society must face the challenges that the population of elderly has increased which raise the healthcare issue for elderly. In Taiwan, 65.2% of elderly suffer from a chronic ailment and require medical service periodically. Ministry of the interior report indicates that ideal living method for elderly is "to live with their children". Traditional culture in Taiwan also agrees that family should support and take care of their elderly. However, support and caring the elderly has becomes a burden for young families especially in the economy recession era. The precaution healthcare service for elderly could lower the loading for take care the elderly and improve the life quality for every family significantly.

The lack of suitable channel for senior citizens to obtain accurate healthcare knowledge is a vital problem nowadays. The healthcare support service for stroke-precaution requires professional medical knowledge that is not available for all elderly and their family members. Since the healthcare for elderly senior citizens requires family support, it is necessary for all family members to access proper stroke-precaution knowledge to provide support for their elderly.

Without the knowledge to identify the hazardous factors for stroke-precaution, it is very dangerous for elderly when they encountering the stroke crisis. Correct channel for elderly to gain healthcare information could benefit their knowledge as well as the general level of their health. However, the medical or healthcare knowledge may not available for elderly due to the incapability to access the suitable healthcare information sources. Even the internet and e-health service could provide much healthcare information for elderly; they are usually confused and might not able to identify which knowledge is suitable for them. This issue is obvious especially in the case that the healthcare information provides massive information for elderly that they cannot understand their true needs. Since most stroke patients are taking cared by their family, the stroke related healthcare knowledge should also be available for elderly as we as their family members to provide necessary support for the stroke patients.

B. E-Service Design Guidelines For Elderly

ICT-based e-service could become a convenient healthcare information channel for elderly, but the physiological recession issue may trouble elderly for using those ICT-based e-services. Previous studies indicate that usage of information technology is inverse proportion to user's age. [13, 14, 15] The visionary

recession will confuse elderly to accommodate to the rapid brightness change [5]. Elderly senior citizens are hardly to keep long-time concentration on specific subject [11]. They will become distracted owing to niggling details, which will showdown learning and reduce the capability of information gathering [2, 6]. Previously study indicate that the short-term memory for senior citizens are not recession due to aging [9, 4], but senior citizens are difficult to process and recall memory especially on space and location subjects [3, 1]. Those issues will influence the utilization of ICT-based healthcare support e-service for elderly senior citizens, and worthy to pay attention on the possible solutions.

A Human Computer Interface (HCI) design should consider the perception of user and the communications between users [8]. Efficient service design is to diminish the communication gap to fulfill user's needs. Since 60% information system errors come from operational fault, proper instruction designs for user could eliminate the operational problems they encountered. These operational instructions for e-service designs are more important for senior citizens who may have learning disorder of the ability to read. Alternative information channel to provide necessary support for e-service operation could utilize all kinds of media elements including test, graphic, audio, animation and video [7]. The disorder to distinguishing colors may raise the difficulty for senior citizens to use traditional e-service interface and should be specific redesigned to fulfill their urgent needs.

Following are some interface design guidelines suggested from previous research, including layout design, readability design, and operation design. The design guidelines of layout design for senior citizens focus on diminishing the physiological recession to simplify the layout to help elderly concentrate on the provided information. For example, succinct layout design, proper length of the interface layout [10], centralized the layout of information provision. To improve the operational instruction for e-service design, it requires providing clear and consistent user interface for senior citizens to understand the meaning of each operation element. Large buttons are easy to comprehend the meaning of each element and is highly recommended.

For readability design, it is design to focus on how to decrease the difficulty for senior citizens to read the content of provided information from e-service platform. For example, Larger the font size and contrast the fonts to the background would help elder to identify the content. High-density content organization will interfere with reading fluency. Moreover, since senior citizens may not understand the usage of hyperlink, eliminate unnecessary hyperlink and replace text-link with graphic one will help elderly understand the interface [11]. Those suggestions will reduce the loading of memory for senior citizens and strengthen the wills-to- learn new things.

Besides the aging issue will influence the service design for senior citizens, the gender issue may also determine the use of ICT-based e-service. As mentioned in Kerschner and Hart (1984), male user have higher acceptance rate for new technologies than female user

[14]. In this study, we will propose a healthcare support e-service design that addressing the service need for senior citizens and we would like to evaluate whether different ages and gender will influence the usage of ICT-based healthcare support e-service.

Gender digital divide for senior citizens do exist, the definition of gender digital divide is to represent the gap of their ability and opportunity to use information technology between different genders. Gender digital divide may also influence the learning effect of senior citizens. According to Huyer and Sikoska (2003), female users do have less advantage in using ICTs and internet [18].

In Taiwan, the gender digital divide issue is reported by Research, Development and Evaluation Commission of Executive Yuan in 2008, the report items of gender digital divide for senior citizens can be unfolded into three parts, including the ICTs usage rate, Searching for healthcare information, and Using ICT course for learning.

For the ICTs usage rate, the giant gap exists especially for the senior female citizens. With the age increased, the gap also increased. The ICTs usage rate for female senior citizens is lower than male for the group of age 61 to 64 years old and the group of age above 65 years old. The female lagging indicator for the group of 61 to 64 years old is 15.9% lag behind, and 9.7% fail behind for the group above 65 years old. The internet usage rate for age 61 to 64 female users fail behind the male user up to 20.9%.

The gender gap exist for using ICTs for searching healthcare information, there are average 81.4% female users will search healthcare information through the internet, while the male user is left behind around 10%. This means female citizens care about the healthcare more than male users. Female citizens have urgent needs for the healthcare information than male users.

The gender digital divide exist in using ICTs course for learning. Female citizens have stronger wills and higher proportion that using ICT course for learning. Previous studies [16, 17] also provide similar outcome that female learn things more than male user, especially for the elderly senior citizens.

III. HEAHTHCARE SUPPORT E-SERVCIE DESIGN ON STROKE-PRECAUTION

The major issues of the healthcare support e-service design for senior citizens are the service function design and its user interface design. The service function design of the proposed stroke-precaution HSS can be unfolded into four key functions, including health status estimation, physical examination record, personalized multimedia stroke-precaution course materials, and achievement test. We will introduce the service functions in the following subsections. However, this paper emphasis on the user interface design for elderly since it is the foundations for elderly to utilize other service functions.

A. Stroke-precaution HSS Functions

We deployed the HSS on touch screen computers; user could interact with the HSS easily. The four function components of HSS on stroke-precaution are tightly connect with each other. A HSS system screen sketch is shown in Figure 1.

Physical examination record. The physical examination records are the fundamental healthcare information for the service system. User will get their physical examination records from hospital or other examination reports. User's physical examination records provide necessary information for the Health status estimation function.

Health status estimation. This function is a computational function that gathering all the physical examination records or other information sources from elderly to provide the health status report. Various healths related factors and criteria are provided from doctor and specialist for establishing the computational mechanism. This function is connected to a recommendation service that generates the personalized healthcare course materials for their urgent needs.

Personalized healthcare course materials. In this study, ten categories of stroke-precaution healthcare knowledge that contains 82 multimedia video clips in three major languages in Taiwan. These healthcare course materials are stored in the multimedia database and connected with the streaming server. Elderly senior citizens could retrieve these course materials from internet or other

Figure 1. The user interface sketch of stroke-precaution HSS.
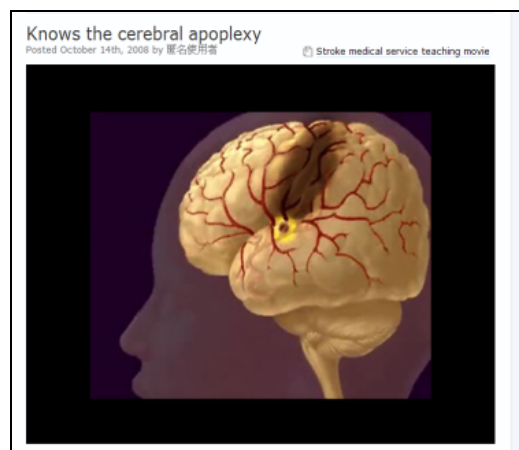

Figure 2. A sketch of course material in Stroke-precaution HSS.

service platform located in hospital. Benefit from the ICT, elderly could get personalized healthcare course materials based on their physical examination. A sketch of course material on stroke-precaution is shown as Figure 2.

Achievement test. Since the course material provided a channel for elderly to obtain personalized healthcare knowledge, it requires a mechanism to confirm the learning outcome. Achievement test provide the evaluation function for senior citizens to understand whether they can comprehends those stroke-precaution healthcare knowledge.

B. The Interface design for elderly

Owing to physiological recessions, elderly senior citizens may feel trouble to use e-services under traditional user interface design. Traditional user-interface layout design usually contains rich information for providing extensively choice for users. However, the senior citizens are not able to concentrate on specific target for a long time. They also have some trouble to get familiars with those fancy layout designs, instead, those dazzlingly designs may becomes the barrier for senior citizens to understand how to use the proper e-service functions. In short, the "simple" layout design would be more suitable for elderly senior citizens.

In this study, we redesign a service user interface for elderly senior citizens to operate the Stroke-precaution HSS. Some ICT features are utilized for providing intuitional operation instructions, such as the touch-screen, the voice-based instructions, etc. The significant features of our proposed design can be unfolded into two parts:

(1)Improving visionary recession problems. In order to present a comfortable service layout for elderly, our proposed interface are designed with 24 point large font and control the information within 110 words per page. Scrolling are not necessary in our interface designs, this will establish convenient reading and operating environment for elderly.

(2)Concentrating on the healthcare information. For the purpose to avoid senior citizens confused by irrelevant buttons, our design combines large and clear images with text descriptions of each button in the service system. Elderly senior citizens will get voice-based assistant in several major languages used in Taiwan.

C. Factors Considered in Stroke-precaution Healthcare Service

Personalized recommendation service on Stroke-precaution HSS is design considering the following factors. Those healthcare related factors can be classified as "Controllable factors" and "Uncontrollable factors".

Controllable factors used in our Stroke-precaution HSS including their age, family historical records, races, gender, and previous medical records on apoplexy. Uncontrollable factors can be advanced classified into three subset. Uncontrollable disease related factors including hyperlipidemia, hypertension, cardiopathy, fatness, diabetes, and mental stress. Uncontrollable environment factor such as temperature difference will also influence the stroke-precaution. Habit factors such as

exercise insufficiency, smoking, drinking, and their food habit will be also considered in the service system.

IV. EVALUATION METHOD AND DATA ANALSIS

A. Evaluation Process for Healthcare Service Design

The evaluation process is design to compare whether the Stroke-precaution HSS could help elderly obtain useful healthcare information. Our proposed Stroke-precaution HSS is compared with a traditional stroke healthcare information service platform, which is design on traditional interface layout. The research design is shown in Table 1. The evaluation matrix for healthcare support e-service design is presented in Figure 3. The evaluation process includes four steps.

Table 1. Research design of HSS Designs

	Experiment Treatments	Post-test
Experimental Group	Stroke-precaution healthcare support e-service for senior citizens	O_1
Control Group	Traditional e-service designs on healthcare information system	O_2

O_1, O_2: Questionnaire contains 12 items in four dimensions

Step 1. The introduction process to describe the service system to the elderly subjects. Elderly will be randomly assigned to either group.

Step 2. Elderly subjects are suggested to practice on using the assigned service system; researchers will observe whether elderly have trouble in using the service system.

Step 3. A task will be assigned for elderly to obtain healthcare knowledge related to his/her personal benefit.

Step 4. An evaluation questionnaire will be completed by subjects to estimate their perceptions.

The evaluation matrix is based on usability design of the e-service and could be unfolded into four dimensions. There are 12 questions included in the questionnaire. The validity analysis of the questionnaire is introduced in next subsection. Learnability represent whether the service interface could facilitate the usability for elderly. System performance is to determine whether the overall performance could fulfill user's expectation. Memoraility is design to measure whether the e-service interface is easy to use for elderly and could be memorized after a
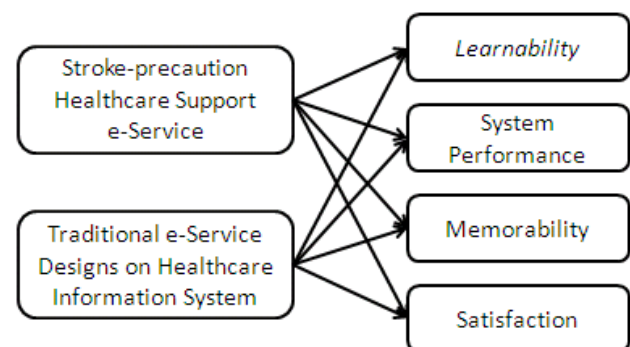

Figure 3. Evaluation matrix for healthcare service designs.

period. The satisfaction is for measuring whether the users satisfy the provided service through the healthcare service.

B. Subjects Discriptions

To understand the perceptions of the healthcare service design from elderly, the subject was selected according to definition from Ministry of the interior in Taiwan. Only aged over 55 years old was considered as the subject of this study. We ask for the subject randomly from the hospital and elderly communities, subjects were randomly assigned into different experiment groups. There are totally 57 subjects considered in this study, due to the missing data of one subject. The effect subject were 56 subjects. The subject descriptions is shown in Table 2 and 3. The experimental group was given the Stroke-precaution HSS, while the control group was given the traditional e-service design on healthcare service.

Table 2: The subject descriptions on gender

	Number of Subjects		Number of Subjects	%
Experimental Group	30	Male	21	37.50%
		Female	9	16.07%
Control Group	26	Male	13	23.21%
		Female	13	23.21%

Table 3. The subject descriptions on ages

	Subgroups	#	%
Experimental Group	Age 50~58	4	7.14%
	Age 59~66	7	12.50%
	Age 67~75	10	17.86%
	Age over 76	9	16.07%
Control Group	Age 50~58	10	17.86%
	Age 59~66	7	12.50%
	Age 67~75	5	8.93%
	Age over 76	4	7.14%

In this study, we would like to know whether the difference exist between different service designs, especially on the age and gender of senior citizens. The data analyses on four different factors are shown as following tables. Table 5 to 8 represents the data analysis between different ages of senior citizen subjects. Table 9 to 12 represents the difference between difference gender of senior citizen subjects.

Table 4. Summary of validity analysis

Dimension	Item number	questions	α
Learnability	1,2	2	.888
System performance	3,6,11,12	4	.917
Memoraility	4,5	2	.904
Satisfaction	7,8,9,10	4	.920
Overall		12	.970

To understand the creditability and stability of our measuring instrument, the validity analysis is shown as Table 4. The Cronbach's of the questionnaire is 0.970. As the four sub-dimensions, Cronbach's α of the four sub-dimensions are higher than 0.7. Overall speaking, the performance of the measuring instrument is acceptable.

C. Data Analysis Results on Different Ages

The data analysis result in Table 5 indicates that the learnability in our proposed Stroke-precaution HSS is not influenced by the ages of users. Various ages' elderly could facilitate the usability for elderly. Although there are not significant difference between experimental group and control group, but the overall evaluation on learnability in our Stroke-precaution HSS is higher than traditional service design.

Table 5. The difference of ages on Learnability

	Subgroups	Avg.	Std.	F
Experimental Group	Age 50~58	8.50	.577	.649
	Age 59~66	8.43	1.512	
	Age 67~75	9.00	.816	
	Age over 76	8.11	1.965	
Control Group	Age 50~58	5.70	2.058	1.642
	Age 59~66	4.71	1.799	
	Age 67~75	7.20	1.789	
	Age over 76	5.50	1.915	

Data analysis shown in Table 6 indicates that the perceived system performance in control group has significant difference (*P<.05) over different ages. This means that traditional service design could not fulfill all users' need especial those senior elderly. However, in our Stroke-precaution HSS, the age issue did not influence the overall performance. Based on the result, elderly senior citizens think the performance in our Stroke-precaution HSS is overall acceptable in all ages users.

Table 6. The difference of ages on perceived system performance

	Subgroups	Avg.	Std.	F
Experimental Group	Age 50~58	17.25	1.500	1.675
	Age 59~66	16.43	1.512	
	Age 67~75	17.80	1.989	
	Age over 76	16.00	2.062	
Control Group	Age 50~58	10.30	3.622	3.932*
	Age 59~66	10.00	3.512	
	Age 67~75	15.60	2.191	
	Age over 76	13.25	2.754	

*P<.05 ; ** P<.01 ; *** P<.001

Table 7. The difference of ages on memoraility

	Subgroups	Avg.	Std.	F
Experimental Group	Age 50~58	8.50	1.000	1.554
	Age 59~66	8.00	.816	
	Age 67~75	8.90	1.370	
	Age over 76	7.89	1.054	
Control Group	Age 50~58	4.60	1.776	5.451**
	Age 59~66	4.71	1.496	
	Age 67~75	7.80	1.095	
	Age over 76	6.50	1.915	

*P<.05 ; ** P<.01 ; *** P<.001

The measuring of whether the age of elderly in different healthcare service will influence the memoriality is shown in Table 7. Data analysis indicate the memoriality in control group has significant difference (*P<.01) over different ages. This means that traditional service design could not make elderly senior citizens in different ages to remember the usage of the interface design. Overall speaking, in our Stroke-

precaution HSS, the age issue did not influence memoriality of elderly. Based on the result, elderly think the memoriality in our Stroke-precaution HSS is acceptable in all ages. We can see the simple service design could make elderly concentrate on the target of the provision service, and elderly senior citizens could remember the operation and feel easy to use that service.

Table 8. The difference of ages on satisfaction

	Subgroups	Avg.	Std.	F
Experimental Group	Age 50~58	18.00	1.826	.845
	Age 59~66	17.00	1.633	
	Age 67~75	18.20	1.751	
	Age over 76	17.33	1.658	
Control Group	Age 50~58	11.30	2.869	3.867*
	Age 59~66	11.14	2.968	
	Age 67~75	16.00	2.828	
	Age over 76	13.75	2.630	

*$P<.05$; ** $P<.01$; *** $P<.001$

Data analysis shown in Table 8 indicates that the user satisfaction in control group has significant difference (*$P<.05$) over different ages. This means that traditional service design could not satisfy the needs from elderly. However, in our Stroke-precaution HSS the age issue did not have significant influence between different ages. Based on the experiment result, elderly feel satisfied in our Stroke-precaution HSS for all different ages.

In general, the service design for elderly is quite different from traditional design paradigm. Due to physiological recessions, elderly are may encounter many problems when using traditional service design. Our proposed HSS is redesigned based on previous study for providing elderly useful healthcare knowledge on stroke precaution. The data analysis indicates that service design could bring elderly a better environment for them to obtain healthcare information.

D. Data Analysis Results on Different Gender

In the previous data analysis result, we found that our proposed Healthcare support e-service design could fulfill the actual needs of senior citizens. Senior citizens have better impressions on the service design that focus on their aging problems. In this subsection, we would like to analysis whether the outcome will be influenced by different gender of senior citizens. Following subsections will introduce the difference of gender on e-service learnability, system performance, memoriability and satisfaction. The summarized data will presented in Table 9 to 12.

Table 9. The difference of gender on Learnability

Groups		Average	Standard Deviation	t
Experimental Group	Male	8.67	1.528	0.802
	Female	8.22	.972	
Control Group	Male	5.85	1.951	0.386
	Female	5.54	2.106	

Comparing the data, we found the learnability of male users and female users are similar in each group. Both male and female users have better learnability in our proposed HSS design. The data analysis result in Table 9 indicates that the learnability in our proposed Stroke-

precaution HSS is not influenced by the gender of senior citizens significantly. However, even there are not significant difference between experimental group and control group, but we found the experimental group has better learnability than the control group. This means the learnability in our proposed Stroke-precaution HSS design could help senior citizens understand the operations than traditional service design.

Table 10. The difference of gender on perceived system performance

Groups		Average	Standard Deviation	t
Experimental Group	Male	21	17.10	0.993
	Female	9	16.33	
Control Group	Male	11.69	3.660	0.000
	Female	11.69	4.070	

Data analysis shown in Table 10 indicates that the perceived system performance between experimental group and control group is not significant for senior citizens. Comparing the data we found that male users in experimental groups have much higher perceived system performance than others. The perceived system performance of Stroke-precaution HSS design is a higher than traditional service design for male user. For female users, traditional design has little advantages than the HSS design. In general, the design in our Stroke-precaution Healthcare Service did not influence the perceived system performance significantly for those senior citizen subjects in different gender.

Table 11. The difference of gender on memoriability

Groups		Average	Standard Deviation	t
Experimental Group	Male	8.43	1.326	0.902
	Female	8.11	.601	
Control Group	Male	5.62	2.142	0.192
	Female	5.46	1.941	

The measuring of whether the gender of elderly senior citizens in different HSS service design will influence the memoriality is shown in Table 11. Data analysis indicates that there is no significant difference on the memoriality in each group. This means the gender of elderly senior citizens did not influence them to remember the usage of the HSS interface. However, comparing the data of each group, we found our proposed HSS design has higher memoriability than traditional design. Overall speaking the gender issue did not influence memoriality of senior citizens in our Stroke-precaution HSS design. The memoriality in Stroke-precaution HSS design is acceptable in both gender. We can see the "simple" service design could make elderly concentrate on the provided healthcare support e-service, and senior citizens could easy to remember the operation and feel easy to use that service.

Table 12. The difference of gender on satisfaction

Groups		Average	Standard Deviation	t
Experimental Group	Male	17.71	1.848	0.450
	Female	17.44	1.333	
Control Group	Male	12.62	3.015	0.116
	Female	12.46	3.711	

Comparing the data, we found the male users and female users are similar satisfaction in each group. However, both male and female users have higher satisfaction in our proposed HSS design. Although the data analysis in Table 12 indicate the gender of senior citizens did not have significant difference in our proposed HSS design and traditional design. Even there are not significant difference between experimental group and control group, but we found both male and female users in experimental group have better satisfaction than the control group. This means the satisfaction in our proposed Stroke-precaution HSS design could make senior citizens feel better when operate healthcare services. Based on the experiment result, senior citizens feel satisfied in our Stroke-precaution Healthcare Service in all different gender.

V. CONCLUSIONS

Current healthcare service platform is rarely designed for elderly. Through continuous learning or participant in the social activities, a good healthcare support e-service would help elderly senior citizens keep a youthful body and healthy mental status. In order to provide a stroke-precaution healthcare support service for elderly senior citizens, this study combines ICT-based learning mechanism with personalized healthcare course materials and a friendly service interface addressing the recession issues of senior citizens. Doctors and specialists from hospital provided the course materials for our stroke-precaution HSS. The intuitive service interface design could help senior citizens access various healthcare knowledge for improving their health status. Analytical result indicates our service design could help elderly senior citizens to leap over the physiological problems they encountered. With stroke-precaution HSS, elderly senior citizens and their families could get better life quality.

REFERENCES

[1] K. E. Cherry, D. C. Park, and H. Donaldson, "Adult age differences in spatial memory: Effects of structural content and practice", *Experimental Aging Research*, vol. 19, pp.333–350, 1993.

[2] S. L. Connelly, and L. Hasher, "Aging and inhibition of spatial location", *Journal of Exp*erimental Psychology: *Human Perception and Performance*, vol. 19, pp.1238–1250, 1993.

[3] M. W. Denny, J. Dairiki, and S. Distefano, "Biological consequences of topography on wave-swept rocky shores. I. Enhancement of external fertilization". *Biol. Bull.* vol. 183, pp.220–232. 1992.

[4] A. R. Dobbs, and B. G. Rule, "Adult age differences in working memory". *Psychology and Aging*, vol.4, pp.500–503 , 1992.

[5] D. W. Kline, and C. T. Scialfa, "Sensory and Perceptual Function: Basic Research and Human Factors Implication". In A. D. Fisk, W. A. Rogers （Eds.）, *Handbook of Human Factors and the Older Adult*. San Diego, CA: Academic Press, 1996

[6] L. Kotary, and W. J. Hoyer,"Age and the ability to inhibit distracter information in visual selective attention". *Experimental Aging Research*, vol.21, pp. 159-171, 1995.

[7] C. Mok, Designing business: Multiple media, multiple disciplines. San Jose, CA: Adobe Press, 1996.

[8] J. Preece, , A Guide to Usability Human Factors in Computing, New York: Wiley Computer Publishing, 1998.

[9] T. A. Salthouse, "The nature of the influence of speed on adult age differences incognition". Developmental Psychology, vol.30, pp.240-259, 1994.

[10] SPRY. A Guide for Web Site Creators, 1999. http://www.spry.org/publications/publications.html

[11] M. Vercruyssen, "Movement control and the speed of behaviour". In A. D. Fisk & W. A. Rogers （Eds.）, *Handbook of human factors and the older adult*. San Diego, CA: Academic Press, pp. 55–86, 1996.

[12] R. W. Morrell, and K. V. Echt, "Instructional design for older computer users: The influence of cognitive factors". In W. A. Rogers, A. D. Fisk & N. Walker (Eds.), *Aging and skilled performance: Advances in theory and applications* (pp. 241-265). Mahwah, NJ: Erlbaum. 1996

[13] R. Adler, and M. Furlong, "Older Americans and the Information Superhighway". *Report of a National Survey.* San Francisco: SeniorNet. 1994

[14] P. Kerschner and K. Hart, "The aged user and technology". In. *Communications Technology and the Elderly* (Dunkle R. ed.), Springer, New York, pp. 135–144. 1984

[15] E. Zandri, and N. Charness, "Training older and younger adults to use software". *Educational Gerontology*, vol.15, pp.615-631, 1989

[16] C. A. Vedura, and E. H. Worthy, *Education for Adults: a Synthesis of Significant Data.* National Council on the Aging, Inc. Publications, Washington, DC. 1982

[17] K. P. Cross, *Adults as learners.* San Francisco: Jossey-Bass,1981.

[18] S. Huyer, and T. Sikoska. "Overcoming the gender digital divide: understanding ICTs and their potential for the empowerment of women." *Instraw research paper series*; no. 1. 2003. Retrieved 14th February, 2007, from http://www.uninstraw.org/en/docs/gender_and_ict/synthesis_paper.pdf.

Yuan-Chu Hwang was born and grew up in Taiwan, R.O.C. Yuan-Chu Hwang received his Ph.D. from MIS Department of National Chengchi University in 2007.

Currently, he is an Assistant Professor of Information Management Department in National United University in Taiwan. His research and teaching interests include e-service innovation, ubiquitous commerce and privacy/trust issues for social mobile applications.

Dr. Hwang was the member of IEEE. He is also the international committee member of AICIT, and NCM, IICIT, IDC, IMS, NISS international conferences.

An Effective Economic Management of Resources in Cloud Computing

Ghalem Belalem
Dept. of Computer Science, Faculty of Sciences, University of Oran (Es Senia), Oran, Algeria
Email: ghalem1dz@yahoo.fr

Samah Bouamama
Dept. of Computer Science, Faculty of Sciences, University of Oran (Es Senia), Oran, Algeria
Email: samahu@hotmail.com

Larbi Sekhri
Dept. of Computer Science, Faculty of Sciences, University of Oran (Es Senia), Oran, Algeria
Email: Larbi.sekhri@univ-oran.dz

Abstract— In Cloud computing, the availability and performance of services are two important aspects to be lifted, because users require a certain level of quality service in terms of timeliness of their duties in a lower cost. Several studies have overcome this problem by the proposed algorithms borrowed from economic models of real world economy to ensure that quality of service, our job is to extend and enrich the simulator CloudSim by auction algorithms inherited from GridSim simulator, but its algorithms do not support the virtualization which is an important part of Cloud Computing, why we introduced several parameters and functions adapted to the environment of cloud computing as well as users to meet their requirements.

Index Terms— Cloud computing, CloudSim, Auction model.

I. INTRODUCTION

Cloud Computing environment is hinting at a future in which we won't compute on local computers, but on centralized facilities operated by third-party compute and storage utilities [9]. This environment, with a great flexibility and ease of use, the availability, of data and of services, becomes one of the biggest problems to treat and to improve [12].

With cloud computing, companies can scale up to massive capacities in an instant without having to invest in new infrastructure, train new personnel, or license new software. Cloud computing is of particular benefit to small and medium-sized businesses who wish to completely outsource their Data center infrastructure, or large companies who wish to get peak load capacity without incurring the higher cost of building larger Datacenters internally.

The CloudSim simulator was proposed to simulate and evaluate the various approaches suggested with the platforms of Cloud Computing [3] [6]. The execution of the simulator makes it possible to make announce a whole of the breakdowns in the treatment of the requests subjected by the users.

GridSim is an event management tool for simulation of heterogeneous Grid resources. It supports the modeling of network entities, users, machines, and the network, including network traffic.

It is able to model and simulate the behavior of Grid applications (execution, scheduling, allocation and monitoring) in a distributed environment consisting of multiple organizations of the grid, but is unable to withstand the demands infrastructure and application levels under the cloud computing paradigm.

In particular, there is very little or no support in the existing grid toolkits for simulation modeling of virtualization to the application to allow the management of resources and applications. Therefore, modeling cloud infrastructure simulation toolkits should provide support to economic entities such as brokers and exchanges cloud for the negotiation of services between customers and suppliers. Among the currently available simulators discussed in this document GridSim that offers support for managing economic resources [2] thing that CloudSim does not have. The work presented in this research is to satisfy the customer in terms of providing a high quality treatment in a minimum cost and in the Clouds Computing offering any improved of algorithms auction of GridSim specific of CloudSim to simulate an economic environment in the cloud computing, with the goal of providing virtualized resources to treat Cloudlet from users.

We offer a hybridization of the two main aspects of simulators GridSim and CloudSim, which are the implementation of economic models of bidding GridSim and virtualization of computing resources CloudSim.

The rest of the paper is structured as follows: in Section 2, we define cloud computing environments as an important scientific trend. Section 3 is dedicated to the simulator GridSim; define the simulator and its uses in the Grid. We present in Section 4, the simulator CloudSim, The following section identifies changes to CloudSim to ensure proper management of resources in

the simulator CloudSim. The different experiments are presented in Section 6. We end our paper with a summary and some extension work that we will consider doing so.

II. CLOUD COMPUTING

Cloud computing can be defined as "A type of parallel and distributed system consisting of a collection of interconnected computers and they are virtualized and dynamically generated and submitted as one or more computing resources based on service level agreement established through negotiation between the provider and consumers" [4]. Some examples of new infrastructure are Microsoft Azure Cloud Computing [8], Amazon EC2, Google and Aneka [15].

Computing power in cloud computing environments is provided by a collection of data centers, which are typically installed with hundreds of thousands of servers [5]. The layered architecture of a typical cloud-based data centers is shown in Figure 1. In the lower layers there are huge hardware resources (storage and application servers) which feed data centers. The servers are managed transparently by the level of virtualization [14], services and toolkits that enable the sharing of their capacity through virtual instances of servers. These virtual instances are isolated from each other, this helps to achieve fault tolerance and isolation of the security environment.
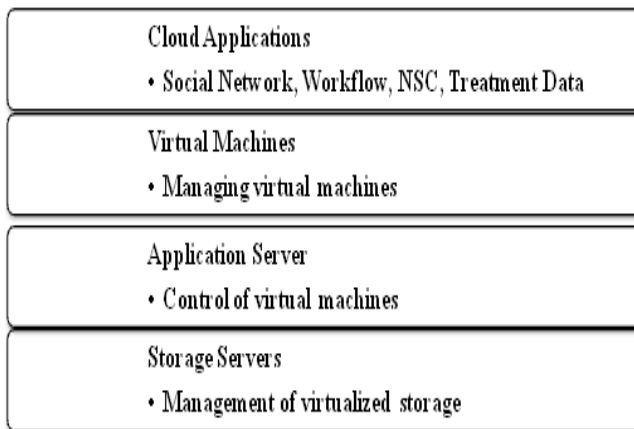

Figure 1. Typical DataCenter.

Cloud applications such as social networking, game portals, applications, economic and scientific workflows running in the highest layer of the architecture. The patterns of actual use of many real applications vary with time, mostly unpredictable ways. These applications have different requirements of Quality of Service (QoS) depending on the criticality of the time and modes of user interaction (online / offline).

III. GRIDSIM

GridSim [5] toolkit was developed by Buyya et al to address the problem of near impossibility of performance evaluation of real large scaled distributed environments (typically Grid systems but also P2P networks) in a repeatable and controlled manner. The GridSim toolkit is a Java based simulation toolkit that supports modelling and simulation of heterogeneous Grid resources and users spread across multiple organizations with their own policies. It supports multiple application models and provides primitives for creation of application tasks, mapping of tasks to resources and managing such tasks and resources.

IV. CLOUDSIM

CloudSim [3] is a framework developed by the GRIDS laboratory of University of Melbourne which enables seamless modelling, simulation and experimenting on designing Cloud computing infrastructures. CloudSim is a self-contained platform which can be used to model data centers, service brokers, scheduling and allocation policies of a large scaled Cloud platform. It provides a virtualization engine with extensive features for modelling the creation and life cycle management of virtual engines in a data center and provides flexibility to switch between space-shared and time-shared allocation of processing cores to virtualized services.
CloudSim framework is built on top of GridSim framework also developed by the GRIDS laboratory.

Figure 2 shows the conception of the CloudSim toolkit [6]. At the lowest layer, we find the SimJava [11] that implements the core functionalities required for higher-level simulation such as queuing and processing of events, creation of system components (services, host, data center, broker, virtual machines), communication between components, and management of the simulation clock. In the next layer follows the GridSim toolkit that support high level components for modelling multiple Grid components Such as networks, resources, and information services.

The CloudSim is implemented as layer by extending the core functionality of the GridSim layer. CloudSim provides support for modelling and simulation of virtualized Datacenters environments such as management interfaces for VMs, memory, storage, and bandwidth. CloudSim layer manages the creation and execution of core entities (VMs, hosts, Datacenters, application) during the simulation period. This layer handle the provisioning of hosts to VMs based on user requests, managing application execution, and dynamic monitoring. The final layer in the simulation stack is the User Code that exposes configuration functionality for hosts (number of machines, their specification and so on), applications (number of tasks and their requirements), VMs, number of users and their application types, and broker scheduling policies. A Cloud application developer can write an application configurations and Cloud scenarios at this layer to perform a cloud computing scenario simulations.
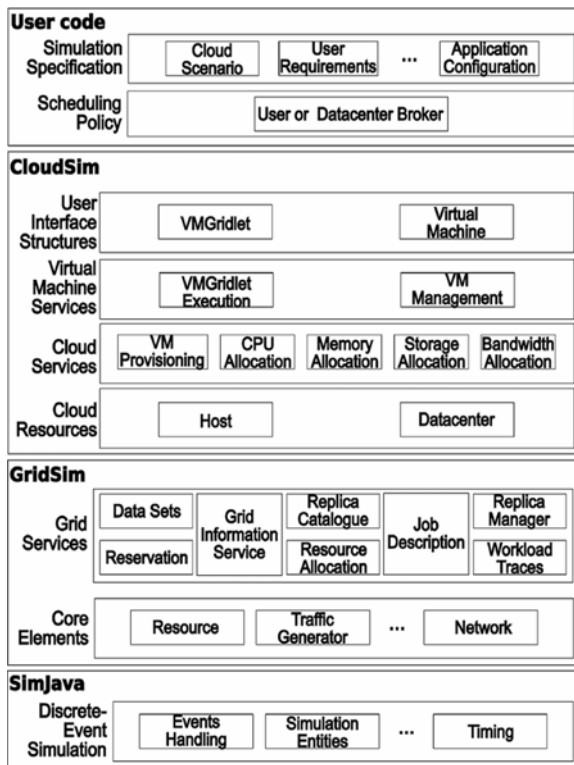
Figure 2. CloudSim architecture [9].

Figure 3. General view auction mechanism.

V. IMPROVED THE CLOUDSIM SIMULATOR

To improve the quality of treatment service of users cloudlets, we made a comparative study between the simulator CloudSim who inherits own bidding algorithms of GridSim simulator, as well as those we have implemented and integrated into the simulator GridSim specific of CloudSim, after decompiled CloudSim to make the changes mentioned above on the simulator GridSim and then recompiled with the toolkit GridSim amended. GridSim implements the four types of auction are: First Price Auction sealed bid (FPSB), English auction, Dutch auction and Continuous Double Auction (CDA), these types of betting revolves around one principle which is the increase or the starting price set by the supplier or the reduction of a very high price to reach the reserve price set by supplier to win the auction with a price that satisfies both the customer and supplier. The diagram below illustrates the general principle of auctions in the Clouds Computing.

Initially, the user submits Cloudlets to his broker. In the Cloud, a broker is responsible for submitting and monitoring Cloudlets on the user's behalf. The broker creates an auction and sets additional parameters of the auction such as Cloudlet length, the quantity of auction rounds, the reserve price and the policy to be used (e.g. English or Dutch auction policy). As the broker also plays the role of auctioneer, it posts the auction to itself; otherwise, the auction would be post to an external auctioneer. The auctioneer informs the bidders that an auction is about to start. Then, the auctioneer creates a call for proposals (CFP), sets its initial price, and broadcasts the CFP to all the bidders. Providers formulate
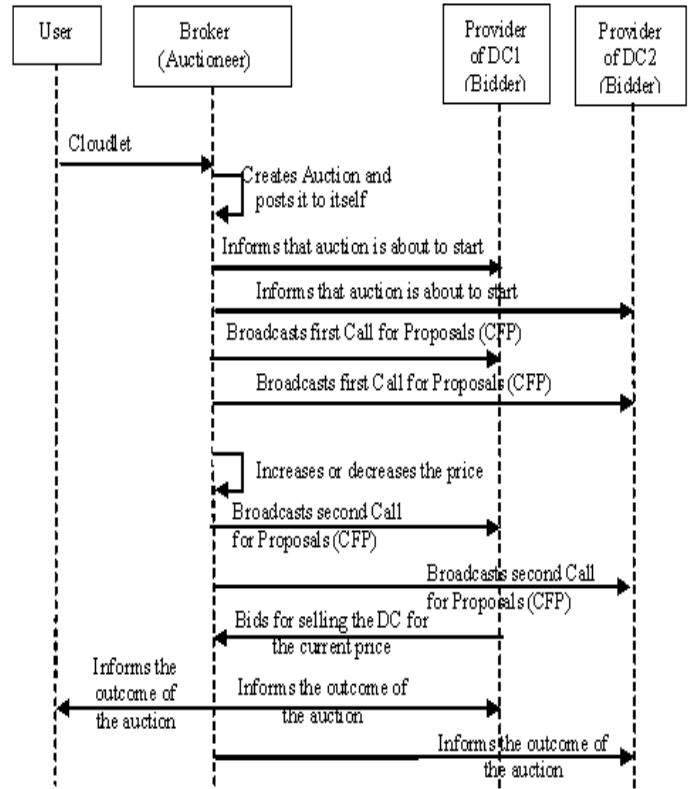
bids for selling a service to the user to execute the Cloudlet. The first time that bidders evaluate the CFP, they decide not to bid because the price offered is below what they are willing to charge for the Data center. This makes the auctioneer to increase or decrease the price and send a new CFP with this new price. Meanwhile, the auctioneer keeps updating the information about the auction. In the second round, a bidder decides to bid. The auctioneer clears the auction according to the policies specified before. Once the auction clears, it informs the outcome to the user and the bidders.

A. First Price Sealed Bid Auction

Bidders are not aware of each other's offers [1]. In addition, it is a single round auction. When bidders receive a call for proposals, they can verify the minimum price and either decide to bid or not to bid for the good. The auctioneer waits a given time for the bids and then allocates the good to the bidder who has valued the good the most. Part of this class is presented in Algorithm 1.

B. English Auction (EA)

This economic model [7] is an ascending auction in which the auctioneer tries to find the price of a good by proposing an initial price below the supposed market value and slowly raising the price until no bidder is interested in paying the current price for the good. Then the best past bid is chosen. Part of this algorithm is presented in Algorithm 2.

Algorithm 1: FirstPriceSealedBidAuction

```
Var double Min_Price, Current_Price,
Reserve_Price, Final_Price;
/*ReservePrice is the maximum price at which a
seller is willing to sell a Data center*/
/*CurrentPrice is the price of asker*/
MessageCallForBids Msg ;
MessageBid best
int  AuctionID, currentRound;

/* This method is called when a round is
started*/
void onStart(int round)
        if (round == 1) then
        Min_Price:=Reserve_Price;
   Current_Price:=Reserve_Price;
   end if
/* Creates a call for proposal that is broadcast to
all bidders*/
    Msg = new
MessageCallForBids(AuctionID,AuctionProtocol,Mi
n_Price,currentRound);
    broadcastMessage(Msg);
end;

/* This method is invoked when a round
finishes*/
void onClose(int round)
   best = getFirstBid();
        if (best != null) then
        double price = best.getPrice();
          if (price >=Reserve_Price)
          then
            FinalPrice:=price;
         Winner_Id:=best.getBidder()
      else
      Final_Price:=Current_Price;
      end if
    else       Final_Price:=Current_Price;
   end if
end;
```

Algorithm 2: English Auction

```
Var double Min_Price, Current_Price,
Reserve_Price, Final_Price;
/*ReservePrice is the maximum price at which
a seller is willing to sell a Data center*/
/*CurrentPrice is the price of asker*/
MessageCallForBids Msg ;
MessageBid  bestBid
int   NumberOfRounds ;
boolean  shouldIncrease

/* This method is called when a round is
started */
void onStart(int round)
    if(round == 1)then
        initialBidders =getBidders();
        Current_Price :=Min_Price;
    end if
        broadcastMessage(msg);
end;
/* This method is invoked when a round
finishes */
void onClose(int round)
boolean stop = false;
    if (!shouldIncrease|| round
==NumberOfRounds)
        stop = true
    else  shouldIncrease = false;
        double increase=(Max_Price–
Min_Price)/(NumberOfRounds - 1);
        Current_Price= Current_Price +
increase;
    end if
   if (stop) then
        if (bestBid != null) then
        double price = bestBid.getPrice();
                if (price >= Reserve_Price)
        then
                Final_Price:=price;
                Winner:=bestBid.getBidder()
            else Final_Price:=Current_Price;
            end if
        else Final_Price:=Current_Price;
        end if
   end if
end;
```

C. Dutch Auction (DA)

This economic principle [7] is a descending auction and differs from the English auction in the sense that the auctioneer starts by issuing a call for proposals with a price much higher than the expected market value. The auctioneer then gradually decreases the price until some bidder shows interest in taking the good for the price announced. Part of this algorithm is presented in Algorithm 3.

D. Continuous Double Auction (CDA)

The Continuous double auction [10] works with a system of bids and asks. The price is found by matching asks and bids. The auctioneer accepts asks and bids and tries to match them. The auctioneer informs the price to the bidder and the seller when a match is made. Part of this algorithm is presented in Algorithm 4.

The principle of continuous double auction is shown in the diagram in Fig. 4.

Among the improvements in our work is that we have introduced a specific parameter to the user who is the budget in order to manage its budget by the number of Cloudlet and their importance in terms of size.

The user will bid with the broker representing the Data Center on an auction mentioned above, when the user tries to treat all its Cloudlet in minimum time and with less cost. CloudSim GridSim and assume that the size of the Cloudlet before treatment is fixed this means

Algorithm 3:   Dutch Auction

Var double Min_Price, Current_Price, Reserve_Price, Final_Price, Current_Price;
 /*ReservePrice is the minimum price at which a seller is willing to sell a Data center*/
 /*CurrentPrice is the price of asker*/
 MessageCallForBids Msg ;
 int   NumberOfRounds ;
 boolean shouldIncrease
 	
 /* This method is called when a round is started */
 public void onStart(**int** round)
     if(round == 1) **then**
       Current_Price:=MaxPrice;
   end if
       broadcastMessage(msg);
 end;
 	
 /* This method is invoked when a round finishes */
 public void onClose(**int** round) {
       if (round >=NumberOfRounds) **then**
       if (bestBid == **null**) **then**
       setFinalPrice(getCurrentPrice())
           else double decrease=
           Max_Price/(NumberOfRounds - 1);
       Current_Price:=Current_Price - decrease;   **end if;**
   end if;
 end;

Figure 4. Continuous Double Auction Diagram.

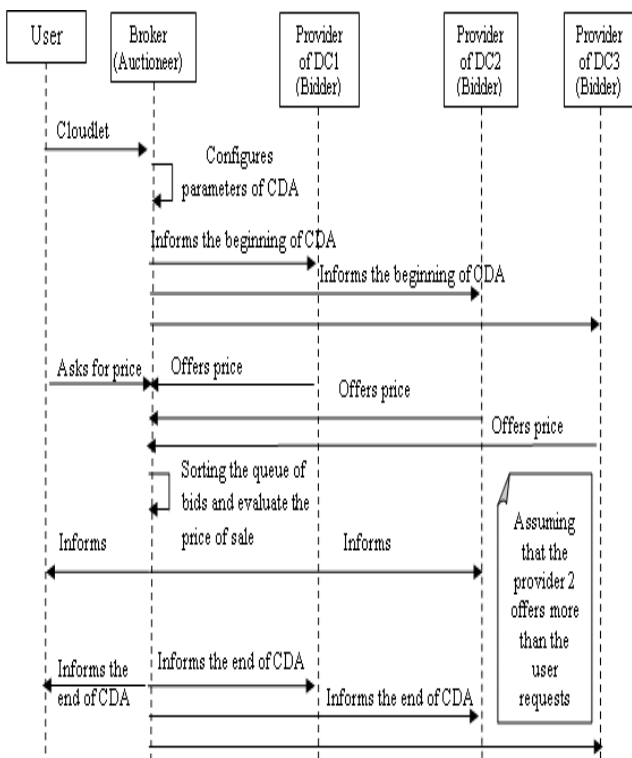| Algorithm 4: Continuous Double Auction |
|---|
| Var LinkedList asks,bids; /*the list of offers and requests*/
 Comparator compAsks,compBids;
 /* Called when a bid is received.*/

 public void onReceiveBid(MessageBid bid)
 Collections.*sort*(asks,compAsks);
       if (asks.size() > 0) **then**
     MessageAsk ask = (MessageBid)asks.getFirst();
       double priceAsk = ask.getPrice();
       double priceBid = bid.getPrice();
           if(priceBid >= priceAsk){
           double finalPrice = (priceAsk + priceBid) / 2;
         match(ask, bid, finalPrice);
         asks.remove(ask)
         else   bids.add(bid);
         end if
     else   bids.add(bid) ;
     end if;
 end;

 /* Called when a ask is sent by a provider. */
 public void onReceiveAsk(MessageAsk ask)
 Collections.*sort*(bids,compBids);
       if(bids.size() > 0) then
     MessageBid bid = (MessageBid)bids.getFirst();
       double priceAsk = ask.getPrice();
       double priceBid = bid.getPrice();
           if(priceBid >= priceAsk)then
             double finalPrice = (priceAsk + priceBid) / 2;
           match(ask,bid,finalPrice);
           bids.remove(bid)
           else   asks.add(ask);
           end if
       elseasks.add(ask);
     end if;
       end; |

that all have the same budget Cloudlet view that it depends on their size using the following formula:

$$UnitaryPriceKB = Budget/TotalSizeOfAllCloudlets \quad (1)$$

From the Formula (1), we can deduce the estimated cost to cloudlet given by the formula below:

$$CostOfCloudlet = UnitaryPriceKB*SizeOfCloudlet \quad (2)$$

In specific auction algorithms of GridSim, the price offered by the Data Center is random and that generated by the user depends on this price, does not guarantee the true estimate of data center cost and the treatment of a cloudlet is why we have proposed a function that generates the offer price. This function estimates the cost of executing the Cloudlet and the cost of their transfers

before and after the execution assuming that its size before and after treatment is the same (see Formula 3).

$$GenBid = (CostPerCPU*TimeCPU)+2*(SizeOfCloudlet* CostPerBandWidth) \qquad (3)$$

Each algorithm has 3 major auction prices for auction process, these prices are: max price, min price and reserve price.

The reserve price is the same in all algorithms of auction and it is equal to the Formula 2, the maximum price in the Dutch and English auction is equal to Formula 4.

$$Maximum\ price = Budget/NB\_Cloudlet \qquad (4)$$

For the minimum price of Dutch auction is zero, while that of the English auction is equal to a price that exceeds the offer price of formula (3) to respect the procedure of the English auction described above.

VI. SIMULATIONS

We conducted four sets of simulation to prove that the auction algorithms provide good quality service in terms of cost and duration of minimum bid. The aim of the first series is to see the impact of size variation of the cloudlet in its cost, therefore we have fixed the number of datacenter to 5, the number of host to 20 each has 2 processors which costs $3 (Price of treatment) with 2MB of RAM which costs $0.5 (price list), 128 MB / sec which costs $0.01 (cost of the bandwidth), 500GB of storage which costs $0.1 (price of storage), the number of virtual machine is 2, the number of user is 1 with a budget of $300 and 20 Cloudlet varied their size between 100 and 1000 KB in increments of 100, the number of rounds for the Dutch and English auction is 10 duration 120 ms, Figure 5 shows the resulting graph.
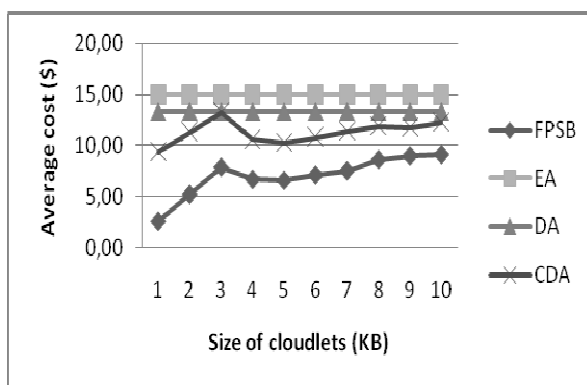

Figure 5. Average cost of Cloudlets according to their size

We can note from the Figure 5 that FPSB lower the cost of Cloudlet more than other auction models despite the increased size of Cloudlet because it is the user who set the starting price this is dependent on its budget, unlike the CDA, where the price proposed is carried by the user and the datacenter divided by two, we also note that despite the size variation of Cloudlet, the average cost DA and EA is stable because in this example the

reserve price coincides with the ask price, with a net increase of EA versus DA.

The aim of the second set of simulation is to see the impact of varying number of data centers on the average cost of cloudlet, we set the size to 300KB Cloudlet 50 each, and varied the number of data center from 1 to 10 by no one, the figure below shows (see Figure 6) that the average cost of four model bid.
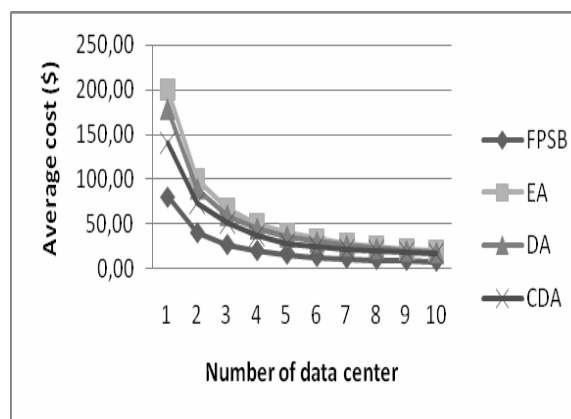

Figure 6. Average cost of Cloudlets according the number of datacenter

The third set of simulation has to stumble to see the impact of number de cloudlets on the average time of the auction, we found the following result (see Figure 7), all bidding models are slow in terms of increasing the number of Cloudlets, but the auction CDA is much faster than other auction for it is not a continuous enhancements such as DA or EA, FPSB in second place because the price generated by the user is compared with the reserve price and the supplier awarded directly, but EA much time compared to DA because it attempts to maximize the greatest possible gain of the supplier without it has a price limit.
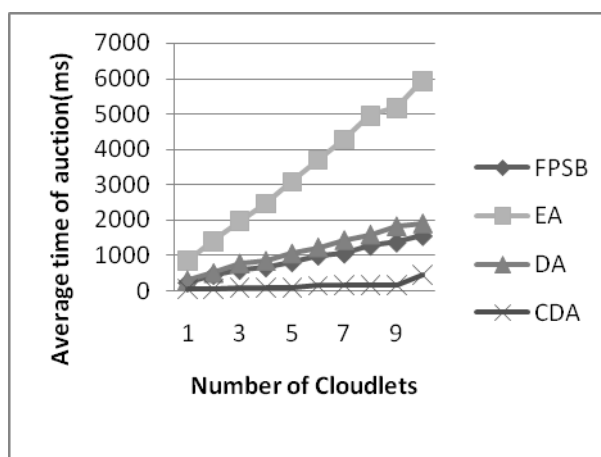

Figure 7: Impact of variation in number of cloudlet on the average time of the auction

Fourth set of simulation has to stumble to see the impact of varying the number of users on the average time of auction, Figure 8 shows that increasing the number of users to a negative impact on the average time

of the auction, but CDA is the fastest of all the models auction saw its operating principle, we also note the EA and DA will coincide in the point (6 ; 150000) that is to say, 6 users both bids are the same time, but beyond 6 EA takes less time compared to the DA this is due to that the competitor to raise prices as quickly as possible to break the price of other competing offers.
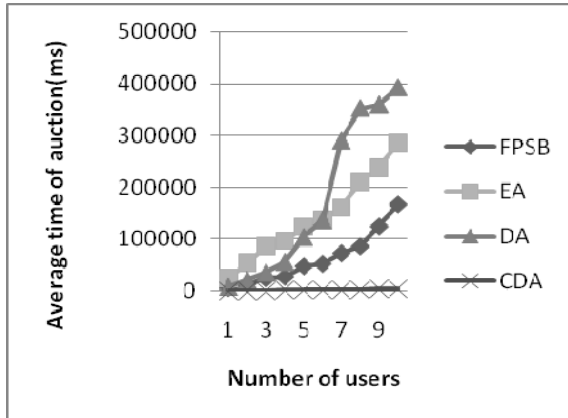

Figure 8. Average time of auction the number of users

VII. CONCLUSION

The recent efforts to design and develop Cloud technologies focus on defining novel methods, policies and mechanisms for efficiently managing Cloud infrastructures.

The main objective of our work is to satisfy customers Clouds, while providing economic functions that reduce the cost of processing cloudlets and improved auction algorithms implemented in GridSim to reduce the time auction and to assure a rapid and effective acquisition of computing resources. The experimental results obtained are very satisfactory and meet our expectations. For a continuation of our work, several perspectives can be envisaged:

- Proposing a method or approach that will satisfy both the customer and supplier resource.
- Extend this work by a process of discovery and allocation of appropriate resources in the cloud computing [13].
- Find a solution for customers who want a high quality of treatment but who do not have enough budgets, by the possibility of credit.
- Propose an implementation of auction club using a multi-agent system.

REFERENCES

[1] M. D. D. Assuncao, and R Buyya, "An Evaluation of Communication Demand of Auction Protocols in Grid Environments", *In Proceedings of the 3rd International Workshop on Grids Economics & Business (GECON 2006)*, pp. 24-33, Singapore, May 2006.

[2] G. Belalem, B. Yagoubi and S. Bouamama, "An Approach Based on Market Economy for Consistency Management in Data Grids with OptorSim Simulator", *International Journal of Information Technology and Web Engineering (IJITWE)*, vol. 3, no. 3, pp.1-16, 2008.

[3] R. Buyya, R. Ranjan, and R. N. Calheiros, "Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities", Keynote Paper. In *Proceedings of the 7th High Performance Computing and Simulation* (HPCS 2009) Conference, Leipzig, Germany 2009.

[4] R. Buyya, C. S. Yeo, and S. Venugopal, "Marketoriented cloud computing: Vision, hype, and reality for delivering IT services as computing utilities". *In Proceedings of the 10th IEEE International Conference on High Performance Computing and Communications*, 2008.

[5] R. Buyya, and M. Murshed, "GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing", *The Journal of Concurrency and Computation: Practice and Experience (CCPE)*, vol. 14, no. 13-15, pp. 1175-1220, Wiley Press, Nov.-Dec., 2002.

[6] R. N. Calheiros, R. Ranjan, C. A. F. De Rose, and R. Buyya, "CloudSim: A Novel Framework for Modeling and Simulation of Cloud Computing Infrastructures and Services", *Technical Report*, GRIDS-TR-2009-1, Grid Computing and Distributed Systems Laboratory, The University of Melbourne, Australia, 2009.

[7] R. Jr. Cassady, "Auctions and Auctioneering". *University of California Press*, Berkley and Los Angeles, California, 1967.

[8] D. Chappell, "Introducing the Azure services platform". *White paper*, Oct. 2008.

[9] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud Computing and Grid Computing 360-degree compared", *in Grid Computing Environments Workshop*, GCE'08, Austin, TX, pp.1–10, 12-16 Nov 2008.

[10] D. Friedman, and J. Rust, "The double auction market: institutions, theories, and evidence". Addison-Wesley, 1993.

[11] F. Howell, and R. Mcnab, "SimJava: A discrete event simulation library for java". In *Proceedings of the first International Conference on Web-Based Modeling and Simulation,* 1998.

[12] T. Rings, G. Caryer, J. R. Gallop, J. Grabowski, T. Kovacikova, S. Schulz, and I. Stokes-Rees, "Grid and Cloud Computing: Opportunities for Integration with the Next Generation Network", *Journal of Grid Computing*, vol. 7, no. 3, pp.375-393, 2009.

[13] A. Sharma and S. Bawa, "Comparative Analysis of Resource Discovery Approaches in Grid Computing". *Journal of Computer (JCP),* vol. 3, no. 5, pp. 60-64, 2008.

[14] J. E. Smith, and R.Nair, "Virtual Machines: Versatile platforms for systems and processes". *Morgan Kauffmann*, 2005.

[15] C. Xingchen; K. Nadiminti, J;Chao; S. Venugopal, and R. Buyya, "Aneka: Next-generation enterprise grid platform for e-science and e-business applications". In *Proceedings of the 3rd IEEE International Conference on e-Science and Grid Computing*, pp. 151-159, Bangalore, 10-13 Dec. 2007.

Ghalem Belalem graduated from University of Oran, Algeria, where he received PhD degree in computer science in 2007. His current research interests are distributed systems, grid and cloud computing, placement of replicas and consistency management in large scale systems and mobile environment.

Samah Bouamama a master's candidate in the Department of Computer Science, Faculty of Sciences, University Of Oran, Algeria. Her research interests are Grid and Cloud Computing, replication strategies and consistency management.

Larbi Sekhri received his PhD in Computer Science from Oran University in 2006 (Algeria). His current research area include formal modeling in wireless ad-hoc and sensor networks, systems modeling using Petri nets, automata theory, diagnosability and monitoring of automated production systems and reasoning in artificial intelligence.

A Study of Travel Agencies' Human Resources in Relation to Internet Marketing

Huang-Wei Su

Tung-Fang Design University/Department of Tourism and Leisure Management,
Kaohsiung, Taiwan
Email: hw@mail.tf.edu.tw

Li-Tze Lee

Overseas Chinese University/Department of Accounting and Information
Taichung, Taiwan
Email: leelitze@ocu.edu.tw

Chiang Ku Fan

Shih Chien University/Department of Risk Management and Insurance,
Taipei, Taiwan
Email: ckfan@ms41.hinet.net

Jason C. Hung

Overseas Chinese University/Department of Information Management
Taichung, Taiwan
Email: jhung@ocu.edu.tw

Abstract—When technological advances emerge in electronic commerce, travel agencies rapidly take account of their application in Internet marketing. Internet marketing presents many advantages to its users, such as the ability to break the barriers of time and space, convenience and speed of trade between travel agencies and consumers, ability to provide customers with sufficient information of travel services and products, and low cost. In order to take advantage of the new Internet market, travel agencies have already set up websites to attract potential customers. However, the traditional marketing of travel agencies generally depends on the sales force, because travel agency representatives act not only as salespeople, but also as guides and escorts. Therefore, personal selling offers the benefits of evaluating travel courses in advance, providing information on various choices, and accompanying customers throughout the trip to help them feel at ease. Since Internet marketing and personal selling both have their own advantages, should travel agency salespeople feel threatened by the emergence of Internet marketing? Will the marketing functions impact travel agency salespeople because consumers purchase services on the Internet themselves? The purpose of this study is to discover whether the travel agencies salespeople feel threatened by the development of Internet marketing. Also, this study seeks to reveal the specific personal thoughts of travel agency salespeople if they do feel threatened. At its conclusion, this study will provide suggestions for training programs or criteria of recruitment for the human resource directors in travel agencies. Also, those interested in travel business may use the results of this study as guidelines as they prepare themselves for the impact of Internet marketing.

Index Terms—travel marketing; electronic commerce; internet marketing; travel agency salespeople

I. INTRODUCTION

According to [19], technology has progressed so rapidly that Internet marketing has became more and more important in the "marketing mix." In the past ten years especially, the economy and society have changed rigorously, and not only the incomes of consumers but also the dollar amounts of purchases have risen. Consumers' habits of purchasing travel products are more different than ever. In order to occupy this market, most travel agencies have invested plenty of resources and manpower in Internet marketing in order to provide the new business opportunities and increased convenience the Internet can provide, through which customers can purchase their travel products. As a result, how do traditional travel agency salespeople feel about this change? Do travel agency salespeople feel threatened in their work, since consumers can now purchase travel products directly over the Internet? Or do these professionals view the Internet as a prosperous new mode of business development? The purpose of this research is to examine the proportions of the travel agency salespeople who do feel threatened and to discover the factors and characteristics of those particular travel agency salespeople.

II. LITERATURE REVIEW

In this section, subjects such as electronic commerce and Internet marketing, the marketing place of travel agencies, and a comparison between Internet marketing and travel agency marketing will be discussed.

2.1 Electronic Commerce and Internet Marketing

2.1.1. Electronic commerce: reference [18] defines electronic commerce as "the buying and selling of goods and services online; conduct of commerce via telecommunications-based tools." Electronic commerce is a modern commercial behavior through Internet service. Several categories involved in e-commerce are as follows: business-to-consumer (B2C), business-to-business (B2B), business-to-government (B2G), consumer-to-consumer (C2C), business-to-employee (B2E) and consumer-to-business (C2B) [5]. From the viewpoints of both definition and categorization, the functions of e-commerce include message transformation and communication between businesses and consumers.

The advantages of electronic commerce include the elimination of barriers related to time and distance, low cost and high effect, easy and fast contact with potential global clients, a revolution of consumer services, and simplified operation procedures [2] [5].

Reference [10] also believes that e-commerce can be utilized through marketing and advertising online, building a company's image, selling products and services, trading online, providing service after a sale, and researching consumers. Considering these benefits, it is not hard to understand why business people are in favor of e-commerce.

2.1.2. Internet marketing: according to [9], marketing is a social and management process by which a person or a group can create values and trade products to satisfy people's needs and desires, and this process by the way of Internet is referred to as "Internet marketing." In 1995, both Time and Newsweek announced that the age of the Internet was officially coming. In Britain, online marketing became so popular that [12] revealed that 84% of business were already using or planning to apply e-mail marketing to promote activities.

Enterprises identified the opportunity to utilize the Internet as a medium to provide and transfer information in order to achieve the goals of promotion and sales. Reference [14] and [16] pointed out that executing Internet marketing with promotion and sales could increase sales volume considerably. Reference [11] stated that a business is suitable for Internet marketing when it 1) provides a service, 2) sells fashion products, 3) offers a product for which customers are seeking a low price, and 4) provides a low product delivery cost. In such situations, businesses can easily build a sustainable relationship with clients and achieve increased sales goals.

2.2 Marketing of the Travel Industry

2.2.1. Marketing distribution: marketing distribution is one of the most important parts of marketing mix; a company's choice of the proper marketing channels to push promotion activities will not only enlarge the market, but also lift the amount of sales.

Reference [9] believed that the marketing distribution channel functions by transferring products from providers to consumers, shortening the time and distance between providers and consumers. In travel agencies, [17] used the term distribution mix to describe these concepts. "It is the combination of the direct and indirect distribution channels that a hospitality and travel organization uses to make customers aware of, to reserve, and to deliver its services." Indirect distribution differs from direct distribution in that it assigns part of its responsibility of promoting, reserving, and providing services to one or more other travel organizations. Because the travel business is a business of service, a key aspect is face-to-face contact with customers; in both direct and indirect distribution, personal selling is a critical phase during the trade process. According to [9], personal selling creates a longer-lasting relationship between consumers and providers. In Taiwan, customers and travel agency salespeople become even more familiar because salespeople also sometimes assume the roles of tour leaders and guides, providing all services from the beginning to the end.

2.2.2. Marketing promotion: in the field of marketing management, promotion uses marketing techniques to communicate information with consumers in order to build great images of products [8] [9]. Reference [17] indicated that "promotional mix" combines several factors to develop a communication strategy for the hospitality and travel organization, including personal selling, advertising, sales promotion, merchandizing, and public relations and publicity approaches for a specific time period. These factors can be categorized into personal selling and non-personal selling. Personal selling includes direct contact with people, while non-personal selling includes advertising, sales promotion, and public relations. Well-executed personal selling provides the most effective way to achieve sales. "It is harder to say no to a personal presentation than to an impersonally communicated message" [17]. Therefore, personal selling is one of the most important and most readily applied promotion strategies in travel agencies.

2.3 Comparison between Internet Marketing and Travel Agency salespeople

2.3.1. Advantages of Internet marketing:

Paperless on-line ticketing system provides not only cost reduction but also time saving benefits [3]. Reference [7], [13] and [20] concluded that travel agencies can gain the following advantages by using Internet marketing: 1) sales can proceed in all-day, year-round marketing broadcasting, 2) updates to travel information can take place any time without a printing fee, 3) costs are reduced through the integration of paper and electronic media with multimedia materials, 4) travel agencies can supply hyperlinks with travel information through which consumers can learn and seek more services, 5) companies enjoy better selling effects when

consumers initiate their own browsing of travel information on the Internet, 6) customers can buy the travel products they want directly, and 7) Internet marketing techniques can focus on a specific target population.

According to [15], the customers' buying process follows the following stages: need awareness, information search, evaluation of alternatives, purchase, and post-purchase evaluation. In these stages, a critical issue is the convenience of obtaining information, including travel destination and travel itinerary. Therefore, Internet marketing allows travel agencies to provide much more complete and convenient travel information without the limitation of time and distance. Because of these features that remove such cumbersome barrier for travel agency salespeople, Internet marketing has become the new development area of marketing management for travel agencies.

2.3.2. Disadvantages of Internet marketing: travel products are seldom necessities; therefore, it is often difficult to push sales. The travel industry, as a service business, combines high-quality service with well-trained customer relationship skills. Personal selling plays an important role in the purchase process. Reference [1] indicated that most travel agency consumers understand the importance of tour arrangements, so their buying behaviors require high involvement. Therefore, the selling process is considered part of the assistance given by travel agency salespeople, and such support is not easily replaced by Internet marketing. Reference [6] mentioned that the high technology of Internet marketing will inevitably lead to self-service in trade and will reduce the contact between travel agency salespeople and clients. In the travel industry, marketing performance considers customer relationships valuable; consequently, Internet marketing is not suitable for travel product selling. In addition, customers are cautious about providing their credit card information on the Internet because the never-ending hackers and thus travel agency may consider give other paying methods as alternatives [3].

As previously discussed, Internet marketing handles two traditional marketing values – promotion and distribution – which used to belong solely to travel agency salespeople. However, the Internet can not technically replace the relationship between the consumer and the tour conductor. Nevertheless, in light of the rise in Internet usage for travel needs, how do travel agency salespeople view such a shift? Do they feel threatened? What professional aspects have the Internet taken over that make travel agency salespeople feel so threatened? This study seeks to answer such questions.

III. METHOD

3.1 Research Steps

As discussed in the literature review, several services typically performed by travel agency salespeople will be replaced by Internet marketing, causing travel agency salespeople to feel threatened. In order to realize the

proportions and the characteristics of travel agency salespeople who feel such a way, the research will follow steps as outlined in Figure 1.
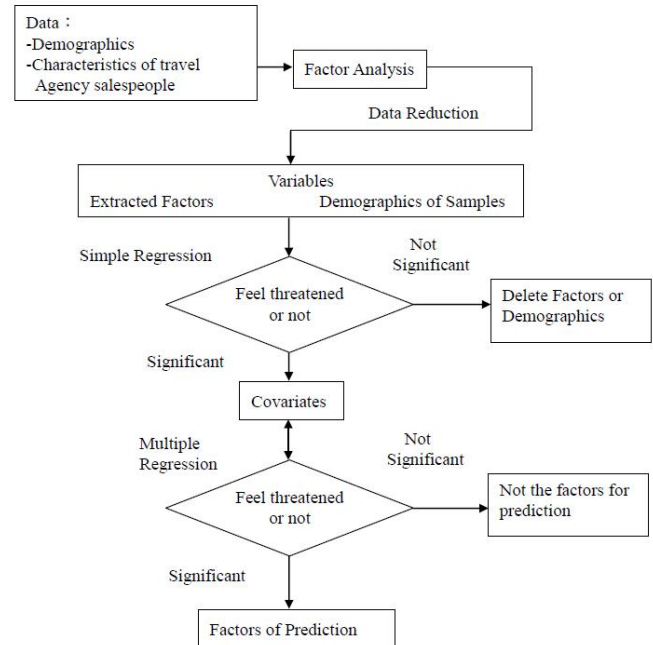

Figure 1. Research Steps

3.2 Hypothesis

According to the research questions, three hypotheses have been formed:

H1: There are significant differences between those who feel threatened and those who do not feel threatened among travel agency salespeople.

H2: There are significant differences in threatened feelings among different demographics of travel agency salespeople.

H3: There is a significant relationship in threatened feelings among salespeople's reaction to different factors of Internet marketing.

3.3 Questionnaire

The questionnaire is composed of two parts. First the questionnaire addresses demographics, including gender, age, professional position, marriage, number of kids, part-time job, and education. Second, the questionnaire addresses the characteristics of travel agency salespeople, using 20 items. The answers are constructed with the Likert scale, with the possible answers being *strongly agree*, *agree*, *neither agree nor disagree*, *disagree*, and *strongly disagree*. According to the pre-test, the reliability is 0.8081 and the validity is 0.7923.

3.4 Sampling

Stratified sampling with proportional allocation is conducted in the sampling. According to the Travel Quality Assurance Association, travel agencies were selected from north, central, and south areas of Taiwan by

proportion. Then, travel agency salespeople were selected randomly from the travel agencies. 400 questionnaires were distributed by mail, and 241 responses were received. The response rate is 60.2%.

3.5 Statistic method

SPSS 16.0 is applied to manage the data analysis, and the statistical method is as follows:

3.5.1. Factor analysis:

(1) Principle component factor analysis: Factors were extracted through the principle component factor analysis and Varimax. When the Eigenvalue is larger than 1 and the factor loading is larger than 0.5, the factor is selected.

(2) Item to total correlation: In order to ensure the core contents of a construct, if the item to total correlation is lower than 0.6, it is deleted.

(3) Internal consistency analysis: For the internal consistency of each construct, Cronbach's α is required to be higher than 0.7 [4].

3.5.2. Logistic regression: the dependent variable is the feeling of being threatened, and the independent variables are the demographics and the characteristics of the travel agency salespeople. Using the Logistic Regression model in simple regression analysis first to find the significant related factors individually, this study then applies those significant factors as covariates to create a prediction model by multiple regression analysis.

IV. DATA ANALYSIS

4.1 Demographics

Among the 241 samples, 68.5% feel threatened by Internet marketing, and 31.5% do not feel so. The results are detailed in Table 1.

4.2 Factor Analysis

Items are deleted when the factor loadings are lower than 0.6, and the 20 items are screened into 13 items. After the factor analysis, the characteristics of travel agency salespeople are extracted into 3 factors; Cronbach's α are all higher than 0.7, as outlined in Table 2.

TABLE I.
DEMONGRAPHICS

	Number	Percentage		Number	Percentage
Gender			Number of kids		
Male	106	44%	0	128	53.1%
Female	135	56%	1	68	28.2%
			2	21	8.7%
Age			3	18	7.5%
20-25	39	16.2%	4	3	1.2%
26-30	69	28.6%	5+	3	1.2%
31-35	55	22.8%			
36-40	24	10.0%	Working experience		
41-45	27	11.2%	1-2 years	59	24.5%
46-50	24	10.0%	3-4 years	77	32.0%
Other	3	1.2%	5~6 year	48	19.9%
			7~8 year	27	11.2%
Education			8 year and more	30	12.4%
Elementary	3	1.2%			
Junior high	15	6.2%	More than 1 job		
Senior high	96	39.8%	Yes	66	27.4%
Institutes	94	39.0%	No	172	71.4%
College	30	12.4%	No response	3	1.2%
Graduate	3	1.2%			
			Travel agent type		
Marital status			In & out bound	80	33.2%
Separated	13	5.4%	In bound	138	57.3%
Live together	14	5.8%	Domestic	23	9.5%
Widower or widow	3	1.2%			
Married	103	42.7%	Threatened feeling		
Single	108	44.8%	Yes	165	68.5%
			No	76	31.5%
			Total	241	100%

TABLE II.

TRAVEL AGENGCY SALESPEOPLE CHARACTERISTICS ANALYSIS

Factor item	Factor loading	Eigenvalue	Cumulative % of variance	Cronbach's α
Internet marketing prospect				
E-commerce means new consumers	.818			
Traditional travel business replaced by the Internet	.780			
Sales achievements higher with e-commerce	.727			
Because of Internet marketing, sales become more difficult	.701	3.612	27.510	.8377
I have had experiences of losing clients to Internet marketing	.673			
I will use e-mail to communicate with customers	.650			
I think Internet marketing will prevail in the travel business	.610			
Work frustration				
I have visited potential customers who declined the sales	.896			
The client doubted my information when they asked for it	.870	3.041	48.905	.8778
For salespeople, clients are diminishing	.861			
Work devotion				
I will visit previous clients, even they do not need my services	.777			
I will prepare necessary documents when visiting potential customers	.658	1.503	62.739	.7509
I devote much time to my sales job	.646			

4.3 Logistic Regressing with Demographics and Characteristics

The simple logistic regression model shows that the gender, education, and travel agency type significantly predicted threatened feelings. See Table 3.

TABLE III.
DEMONGRAPHICS AND THREATENED FEELINGS

Independent variable	β	S.E.	P value	Exp(β)
Gender	0.677	0.290	0.020*	1.9671
Education	-0.624	0.284	0.028*	0.536
Travel agency type	-0.965	0.443	0.029*	0.381

*P<0.05

The proportion of female salespeople who feel threatened is higher than the proportion of male salespeople who do. With an institute or college education or higher, the proportion of salesmen who feel threatened is lower than the number of salesmen with lower education. Within the travel agency type, the proportion of domestic travel agency salespeople who feel threatened is lower than the proportion of inbound and outbound salesmen.

The other simple logistic regression model shows that characteristics of salesmen (internet marketing prospect, work devotion, and work frustration) significantly predicted threatened feeling. See Table 4.

TABLE IV.
CHARACTERISTICS AND THREATENED FEELINGS

Independent variable	β	S.E.	P. value	Exp(β)
Internet marketing prospect	0.580	0.1782	0.001*	1.786
Work devotion	-1.481	0.212	0.037*	1.5558
Work frustration	0.443	0.241	0.000*	0.2273

*P<0.05

Table 4 shows that Internet marketing prospect and work frustration are positively significantly related to threatened feeling. Work devotion is negatively significantly related with threatened feeling.

4.4 Multiple Logistic Regression with Factor

Using both factors of demographics and characteristics as covariates as independent variables, the multiple logistic regression model shows that the Internet marketing prospect, work devotion, and work frustration significantly predicted threatened feeling. See Table 5.

TABLE V.
MULTIPLE LOGISTICS REGRESSION

Covariates	β	S.E	P value	Exp (β)
Internet marketing prospect	0.6089	0.2072	0.0033*	0.5439
Work devotion	-0.4250	0.2763	0.1240	0.6538
Work frustration	1.4808	0.2446	0.0000*	4.3964
Constants	-2.3181	1.6468	0.1592	

*P<0.05

The multiple logistic regression model is as follows:

$$\ln\frac{P}{1-P} = f(x) = -2.3181 + 0.6089 * a - 0.4250 * b + 1.4808 * c$$

a: Internet marketing prospect
b: Work devotion
c: Work frustration

According to this model, if a travel agency salesperson has a higher Internet marketing prospect, is not devoted to his work, and feels frustrated at work, he will have a higher threatened feeling concerning Internet marketing.

V. CONCLUSION

Almost 70% of the travel agency salespeople had threatened feelings with regard to Internet marketing. Furthermore, this feeling had significant relation to individual factors such as gender, education, travel agency type, Internet marketing prospect, work devotion, and work frustration. To predict threatened feelings, a model combining Internet marketing prospect, work devotion, and work frustration was developed by multiple logistic regression. This means that a travel agency salesperson with higher Internet marketing prospect, lower work devotion, and higher work frustration will have higher threatened feelings with regard to Internet

marketing.

Therefore, if a salesperson has confidence in the Internet marketing prospect, ways in which to take the advantage of Internet marketing will be a solution to those threatened feelings. High work devotion improved the communication with clients with whom the salesmen created close relationships, and they think those relationships are not replaceable. Those with work frustration face the inevitable task of improving their selling skills to overcome the drop in clientele due to the rise in Internet marketing.

REFERENCES

[1]   Assael, H. (1987). *Consumer Behavior and Marketing Action*. 3rd ed., Boston: Kent Publishing Company.

[2]   Avery, S. (2002). A move to e-procurement improves order efficiency, *Purchasing*, vol. 131(3), pp. 47.

[3]   Chao, C. N., Mockler, R. J. & Dologite, D. G. (2009). Online Ticketing in the Airline Industry: The Global Impact, *Review of Business Research,* vol. 9(2), pp. 176-179.

[4]   DeVellis, R. F. (1991). *Scale Development Theory and Applications*, London: SAGE.

[5]   Dumitrache, M. (2010). E-Commerce Applications Ranking, *Informatica Economică,* vol. 14(2), pp. 120-132.

[6]   Gallagher, J. (2002). Balancing between high-tech & High-touch, *Insurance & Technology*, vol. 27(3), pp. 22-27.

[7]   Higgins, B. (2001). Low maintenance Producers can pull in higher commissions, *National Underwriter / Life & Health Financial Services,* vol. 102(50), pp. 56.

[8]   Kinnear, T.C. & Bernhardt, K.L. (1990). *Principles of Marketing*. 3rd ed., Glenview, IL: Scott and Foresman.

[9]   Kotler, P. (2000). *Marketing Management*, 10th ed., Upper Saddle River, N.J: Prentice-Hall.

[10] Kozinets, R.V. (2002). The field behind the screen: Using netnography for marketing research in online communities, *Journal of Marketing Research,* vol. 39(1), pp. 61-72.

[11] Kumar, S.R. (2001). Combining on-line and off-line marketing strategies, *Ivey Business Journal*, vol. 66(2), pp. 14-6.

[12] Mazur, L. (October, 2001). E-mail may well be the future of Web marketing*, Marketing,* pp. 16.

[13] McAllister, M. (2001). Online choices shape healthcare buying, *Health Management Technology*, vol. 22(12), pp. 10.

[14] McMaster, M. (2002). In the field: Penguins put the Web to work, *Sales and Marketing Management*, vol. 154, pp. 24.

[15] Mill, R. C., Morrison, A. M. (1985). *The Tourism System: An Introductory Text*. Englewood Cliffs, N.J.: Prentice-Hall.

[16] Monti, R. (2000). Better Web Designs Boosts Sub Sales. *Folio: The Magazine for Magazine Management,* vol. 29 (2), pp. 54.

[17] Morrisson, A. M. (1996). *Hospitality and Travel Marketing,* 3rd ed*., New York, N.J: Delmar Publishers.*

[18] Movahedi, S. J. (2002) *E-Commerce. Reference & User Services Quarterly*, vol. 41(4), pp. 316-325.

[19] Rizzi, J. (2001). In the Mix, *Target Marketing*, vol. 24, pp. 13.

[20] Thompson, S. (2002). Doritos to Net teens, *Advertising Age*, vol. 73, pp. 8.

KNN-DTW Based Missing Value Imputation for Microarray Time Series Data

Hui-Huang Hsu, Andy C. Yang, Ming-Da Lu

Department of Computer Science & Information Engineering,
Taipei, Taiwan, R.O.C.
Email: huihuanghsu@gmail.com, andyyung0215@gmail.com

Abstract—**Microarray technology provides an opportunity for scientists to analyze thousands of gene expression profiles simultaneously. However, microarray gene expression data often contain multiple missing expression values due to many reasons. Effective methods for missing value imputation in gene expression data are needed since many algorithms for gene analysis require a complete matrix of gene array values. Several algorithms are proposed to handle this problem, but they have various limitations. In this paper, we develop a novel method to impute missing values in microarray time-series data combining k-nearest neighbor (KNN) and dynamic time warping (DTW). We also analyze and implement several variants of DTW to further improve the efficiency and accuracy of our method. Experimental results show that our method is more accurate compared with existing missing value imputation methods on real microarray time series datasets.**

Index Terms—**microarray time series data, missing value imputation, dynamic time warping, k-nearest neighbor.**

I. INTRODUCTION

Recently, microarray technology has become one of the important tools in biological researches. Microarray makes it possible to monitor thousands of gene expressions in a single experiment. Numerous gene expression data can be generated simultaneously via this high throughput biological technology. Microarray time-series data are generated by using cDNA microarray technology. It was previously found in many studies that gene expression values represent the reaction of each gene after the hybridization effect across time [1,2]. Each gene expression value represents different reaction degrees resulted from experiments. These quantitative values are in the format of logarithm. This kind of data provides a possible means for the inference of transcriptional regulatory relationships among the genes on the microarray gene chips. The discovery of specific gene pairs with highly-correlated relations could provide valuable information for biologists to predict important biological reactions [3].

Nevertheless, microarray gene expression data usually contain multiple missing values. A specific portion of gene expression values that do not exist in microarray gene expression raw data are called missing values. For subsequent analysis of microarray gene expression data, these kind of missing values need to be effectively estimated and imputed. The reason why missing values occur may result from experimental operations, experimental inaccuracy, or unobvious reaction at that time slot of certain genes [4]. If there is a particular gene I with one missing value at time slot J, then YIJ is denoted to represent the target missing value.

In this paper, we propose a novel approach combining k-nearest neighbor (KNN) and dynamic time warping (DTW) to impute missing values in DNA gene expression microarray time series raw datasets. Furthermore, we also survey and analyze several techniques for accuracy and efficiency refinements of the DTW algorithm to further improve our method. Compared with existing imputation methods, our method is more effective because we not only keep the advantage of the KNN method but also make improvements on it to reach a better result. Our method can impute missing values efficiently in spite of the existence of outliers and time delays. Moreover, our method takes DTW as the similarity measurement of two genes so that it is still practical while handling gene sequences of different lengths. Experimental results show that the proposed method with specific variants of DTW outperforms others in imputing missing values in microarray time series data.

The remaining of this paper is organized as follows. In Section II, discussions on existing imputation methods for missing values in microarray time-series dataset are given. The proposed method is delineated in detail in Section III. Section IV described the involved datasets and the estimation of imputation accuracy. The experimental result is then presented and discussed in Section V. The concluding remarks are drawn in Section VI along with future work.

II. LITERATURE REVIEW

An early study summarizes and implements three methods: singular value decomposition based method (SVD-impute), weighted k-nearest neighbor (KNN-impute), and row average imputation [5]. The results in the paper show that the KNN imputation appears to be a better solution for missing value estimation than SVD-impute. The paper also mentions that the best number of k is proved to be set between 10 and 20. Both SVD-impute and KNN-impute surpass the commonly used row average method or filling missing values with zeros. Several imputation methods are proposed based on the method introduced in the work. For example, Kim *et al.* develop a new cluster-based imputation method called

sequential k-nearest neighbor (SKNN) method [6]. The method imputes the missing values sequentially from the genes with least missing values, and uses the imputed values for the later imputation. This study is typically an example showing the effectiveness of KNN with some improvements on it. Besides KNN or KNN-based imputation methods, there are still other works proposing several methods of different aspects. Oba *et al*. propose an estimation method for missing values based on Bayesian principal component analysis (BPCA) [7]. The method combines mathematical theorems and needs no modelling parameters which are difficult to determine. The results outperform the KNN and SVD imputations according to the authors' evaluations. Moreover, an imputation method called LLS-impute based on the local least squares formulation is proposed to estimate missing values in the gene expression data [8].

For existing imputation methods, BPCA is shown to outperform others. However, it is not easy to decide the number of principal axes while applying BPCA for missing value imputation [9]. Existing methods for microarray missing value imputation mainly utilize k-nearest neighbor (KNN) or KNN-like approaches to estimate the missing values. When applying KNN to impute missing values in microarray time series-data, we have to choose a number of k similar genes without missing entries at the same time slot (experiment) as the target missing value. Besides, we still need to estimate how similar the two genes of interest are to identify whether the two genes have regulatory relations. However, most of the similarity measurements take the statistical or mathematical correlations among genes into consideration. These principles may be unsuitable to the microarray time-series data because of the existence of outliers. Outliers influence much on the correlation coefficient measurements, especially when there are two or more outliers occurring in the time-series data set. A study suggests that outliers do exist in certain microarray datasets [10]. Also, when identifying similarity of two genes in microarray time-series data, comparing local similarity is usually more important than comparing all time slot points. The reason is that even genes with known regulations may have reaction delay or offsets among time axis in microarray experiment results [11]. As a result, it is necessary to apply a similarity measurement method that has the capability of pointing out local similarity and is also effective even with certain existing outliers in microarray time-series data.

III. METHOD

To overcome the difficulty mentioned above and to achieve better imputation results, we propose a novel missing value imputation method based on KNN and dynamic time warping (DTW). With our method, local similarity and shifted trends of gene expression data between each gene pair can be discovered. Studies argue that it is more important to observe and find out whether there exist sub-sequences with highly similar relations when analyzing whole microarray time series data. However, most of existing studies and approaches take

Euclidean distance that identifies global similarity into account as the measurement to determine whether two genes are similar or not. Also, DTW is practical even though the two gene sequences to be aligned are of different length, where this often happens in real microarray dataset. Besides, offsets among time-axis do not influence much on DTW because DTW generates corresponding mappings of two sequences according to their similar shapes. Outliers will not influence on DTW critically compared with Euclidean distance or Pearson correlation coefficients. With our method, better imputation results can hence be achieved. Moreover, we also try our method with several variants of DTW to further improve its efficiency and accuracy of imputation. This section briefly describes the KNN method and the DTW algorithm, followed by the combined method for missing value imputation.

A. K-Nearest Neighbor Imputation Method

The k-nearest neighbor (KNN) method selects genes with expression values similar to those genes of interest to impute missing values. For example, if we consider gene G that has one missing values at experiment time slot T, KNN would find K other genes that have a value at experiment time slot T, but with expression values most similar to Gene G in experiments time slot points except for T. A weighted average of values at experiment time slot T from the chosen K closest genes is then used as estimation for the missing value in gene G. As for the weighted average, the weighted value of each gene in the K closest similar genes is given by the similarity of its expression to that of gene G. To determine the k closest genes which are similar to the target gene G with missing values to impute, matrices such as Euclidean distance measurement, Pearson correlation coefficient, or other distance-based similarity measurements are applied. Usually, Euclidean distance measurement is the most commonly used metric for this purpose. The steps of KNN imputation are as follows:

Step1. In order to impute the missing value G_{IJ} for gene I at time slot J, the KNN-impute algorithm chooses k genes that are most similar to the gene I and with the values in position k not missing.

Step2. If Euclidean distance measurement is employed for two gene expression vectors $g_x = <e_{x1}, e_{x2}, e_{x3}...e_{xn}>$ and $g_y = <e_{y1}, e_{y2}, e_{y3}...e_{yn}>$, the Euclidean distance between g_x and g_y can be calculated as follows:

$$\text{dis}(g_x, g_y) = \sqrt{\sum_{t=1}^{n} (e_{x,t} - e_{y,t})^2} \qquad (1)$$

Step3. The missing value is estimated as the weighted average of the corresponding entries, in the selected k expression vectors:

$$G_{IJ} = \sum_{i=1}^{k} W_i \times e_{iJ} \qquad (2)$$

Step4. The weighted value

$$W_i = \frac{1}{dis(g^*, g_i) \times \Delta} \tag{3}$$

$$\Delta = \sum_{i=1}^{k} [1/(dis(g^*, g_i))] \tag{4}$$

,where

and g* denotes the set of k genes closest to g_i

When applying the KNN-based method for the imputation of missing values, there are no constant criteria for selecting the best k-value and similarity measurements. Both k-value and similarity measurements have to be determined empirically. It is not constant for determining the k value. Choosing a small k value produces poorer performance after imputation. On the contrary, choosing a large neighbourhood may include instances that are significantly different from those containing missing values. However, on study shows that setting k-value between 10 and 20 brings the best results for KNN imputation [5]. With this result, KNN can be an effective and intuitive imputation method if it works with a proper similarity measurement for genes.

B. Dynamic Time Warping

Dynamic time warping (DTW) is a commonly-used algorithm in voice and pattern recognition [12,13]. It has been shown that DTW performs well to find out the similarity for a pair of time series data [14,15]. In this paper, we combine the KNN method with the DTW algorithm as the similarity measurement to estimate the missing value in microarray time-series data. In general, the dynamic time warping (DTW) method is used to warp and match generic sequences of numbers that can be viewed as curves in a proper coordinate system. The aim of DTW is to obtain a precise matching along the temporal axis, and to maximize the number of point-wise matches between two time series. The alignment of temporal patterns by DTW has traditionally been used in the recognition of speech signals. This method is a widely-used algorithm for string comparison and for the alignment of time series data. If two series with time points are given as input, the DTW algorithm can select the best possible alignment between them by minimizing a local distance between the series points.

DTW is a recursive algorithm that matches each two-point pair from the first element to the last element on input sequences. After the table recording all local optimal paths and corresponding points is completed, a multiple of its last computed value returns the DTW distance between the two sequences. If we are going to align two sequences that are similar with observation, the application of Euclidean distance or Pearson correlation coefficient of these two sequences may have poor performance due to shifts on time axis. With DTW mapping method, local similarity can be found as the best path within the two comparison sequences. As a result, if two genes with similar gene expression values at certain parts in microarray time series data are analyzed by DTW, it is more precise for similarity measurement because DTW can discover their similarity that cannot be identified with other similarity measurements. Equations of DTW algorithm are as follows:

Distance of two time slot pints:
The distance between the elements of the two time series is computed as:

$$dis(i, j) = |x_i - y_i| \tag{5}$$

Base Conditions:
$e(0,0) = 0$;
$e(1,1) = dis(x_1, y_1)*W_D$;
$e(i,0) = \infty$ for $1 \leq i \leq I$;
$e(0,j) = \infty$ for $1 \leq j \leq J$; \hfill (6)

where W_D is the weighted value for the paths in the diagonal direction.

Recursive Relation:

$$e(i, j) = \min \begin{cases} e(i, j-1) + dis(x_i, y_j)*W_V \\ e(i-1, j-1) + dis(x_i, y_j)*W_D \\ e(i-1, j) + dis(x_i, y_j)*W_H \end{cases} \tag{7}$$

where W_V, W_D, and W_H denote the weighted value for the paths in the vertical, diagonal, and horizontal directions respectively.

Output: DTW distance for two sequences X and Y:

$$DTW(X,Y) = \frac{1}{n+m} * e(i, j) \tag{8}$$

where length of X and Y are n and m respectively.

C. Imputation Method Combining KNN and DTW

While handling missing value imputation problems, we combine the KNN method with the DTW algorithm. DTW is utilized as the similarity measurement between gene expression values at time slots. The combined method substitutes equation (9) and equation (10) for equation (3) and equation (4) respectively.

$$W_i = \frac{1}{DTW(g^*, g_i) \times \Delta} \tag{9}$$

where

$$\Delta = \sum_{i=1}^{k} [1/(DTW(g^*, g_i))] \tag{10}$$

Missing values for the target gene are hence imputed with our proposed method.

D. Refinement of the Algorithm

DTW is a widely-used method to align sequences in many fields. It works well especially to discover local similarity of two sequences even with different lengths. To further test our imputation method, we survey and analyze some variants of DTW and try to advance the efficiency and accuracy of our method. Variants of DTW are usually categorized for two purposes: speed up and accuracy. In this sub section, we describe these two sorts of refinements for our proposed method.

1)   Computational Efficiency

The critical disadvantage of DTW is its high computational cost. The time complexity of the traditional DTW algorithm is O(n*m) for two input sequences with length n and m, respectively. Despite the high throughput and computational ability of modern computers, speeding up the calculations for DTW distance is still essential when the size of involved data is very large. As we will show in Section IV, we use the Spellman's dataset to perform missing value imputation with totally 6178 genes in the dataset. However, if we naively use the original DTW algorithm to calculate DTW distance of the total 6178 genes, the calculation time needed is awfully amazing and reduces the practicability of the algorithm. As a result, several methods to speed up the calculation of DTW are proposed. Among all existing methods, we find the most useful one proposed in [16]. The authors propose an approximation of DTW called FastDTW that has linear time and space complexity. In other words, if there are two time series sequences with near the same length n, time complexity for FastDTW is only O(n). The method uses a multilevel approach that recursively projects a solution from a coarse resolution version of original data and refines the projected solution. With FastDTW, the computational cost can hence be reduced. FastDTW works somehow like the "divide and conquer" technique in the algorithm field. It uses a multilevel approach with three operations:

*a)   **Coarsening**: Coarsening means that FastDTW shrinks a time series into a smaller one which represents the same curve as accurately as possible with fewer data points.*

*b)   **Projection**: After FastDTW performs the coarsening step, it will find a minimum- distance warping path at a lower resolution, and use the path to guess another minimum-distance warping path in a higher resolution.*

*c)   **Refinement**: Finally, FastDTW refines each warping path in every resolution projected from a lower resolution with local adjustments.*

FastDTW cut the points needed from original time series to the lowest resolution from 1/1 to 1/2 to 1/4 and so forth, and then projects paths from each lower resolution to a higher one. For example, if there are 18 points in an original time series, FastDTW cuts the points needed from 18 with a two-times reducing rate. This forms every resolution in the coarsening process. However, according to our testing, we find that coarsening with a three-times reducing rate performs better than coarsening with a two-times reducing rate in terms of the dataset involved. This is because the dataset we use only contain 18 or 17 time points and only need two times of coarsening. As a result, we modify the FastDTW algorithm and set the coarsening rate from 1/1 to 1/3 to 1/9 as shown in Fig. 1. With three-times reducing rate, we can retrieve better computational efficiency with almost the same accuracy. Finally, FastDTW gives a specific tolerance region for projecting

the warping path from a lower resolution to increase the probability that paths of FastDTW runs through paths of real DTW. This procedure is performed to slightly improve the accuracy of FastDTW.

2)   Accuracy of imputation

The other attractive issue for the improvement of the DTW algorithm is in accuracy increasing. Although DTW has been used in various fields with success, there is still a drawback called the singularity problem [17]. In some cases, DTW would lead to unintuitive alignments where a single point on one time series is mapped onto a large subsection of the other time series. In other words, a specific point on one input sequence may map into even more than three points on the other sequence. This kind of undesirable behavior is so called the singularity problem. As we mentioned in previous sections, DTW works well due to its capability of discovering local similarity of two sequences so that forms a dynamic mapping which is better than usual global similarity measurement like Euclidean distance or Pearson correlation coefficient. However, when the two sequences to be aligned are highly similar but with only slightly different height of the peaks mapped on the two sequences, DTW will perform a one-to-many mapping. This reduces the effectiveness of the algorithm because this kind of mapping is not expectative theoretically and will easily fail to find obvious and intuitive alignments for sequences. As a result, modifications on DTW to mitigate the singularity problem are essential.
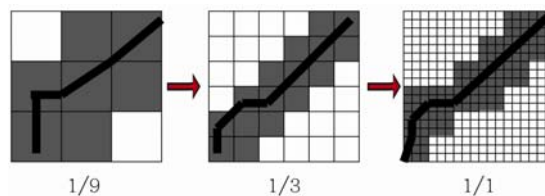

Figure 1.   Modification of FastDTW

Several works to mitigate this problem are developed with certain effectiveness respectively. We survey and analyze existing methods aiming to reduce the singularity problem and choose four of them to implement on our proposed missing value imputation method. Within the four methods, three of them are typical constraining modifications of DTW developed one decade ago, while the other one is a recently-revised method. In the following paragraph, we give a brief description on the four methods.

*a)   **Windowing**: Berndt and Clifford proposed a restricted version of DTW so that the allowable paths in the cost matrix are limited with a warping window : |i-j| ≦ w, where w is a positive value [18]. The windowing constraint makes the corners in the cost matrix of the DTW algorithm pruned so that possible paths of modified DTW are reduced. This constraint may mitigate the seriousness of singularity but it is not able to prevent it. The point in one sequence is limited to match a number of points on the other sequence.*

b) Slope weighting: Kruskall and Liberman proposed a modification of DTW so that the recursive equation in the original DTW algorithm is replaced by $r(i,j) = d(i,j) + min\{r(i-1,j-1) , X*r(i-1, j), X*r(i, j-1)\}$ where X is a positive real number [19]. With this constraint, the warping path is biased toward the diagonal if the weighted value X gets larger. Since each step in recursive procedures of DTW searches for minimum distance summed up so far, the warping path with slope weighting tends to walk through the diagonal direction if larger weighted value X is assigned. This modification of DTW takes the weighted value into consideration that it tries to slightly encourage the warping path goes in diagonal direction to reduce singularity.

c) Step patterns (Slope constraint): Itakura proposed a permissible step for the warping path with $r(i,j) = d(i,j) + min\{r(i-1,j-1) , r(i-1, j-2), r(i-2,j-1)\}$[20]. With this constraint, the warping path is forced to move one diagonal step for each step parallel to an axis. In other words, if the DTW algorithm selects the horizontal or vertical adjacent path in one step, the subsequent path is forced to move one diagonal step.

d) Derivative Dynamic Time Warping: Keogh and Pazzani introduced a modification of DTW [21]. This modification of DTW is called derivative dynamic time warping (DDTW). The author considered only the estimated local derivatives of gene expression values in sequences instead of using the whole gene expression values themselves. DDTW uses a modified estimation to substitute for equation (5) while calculating the distance between two time points from two time series. This estimation equation is as follows:

Distance for two time points in two sequences
$$dis(i, j) = | E(X_i) - E(Y_i)|^2 \text{ while}$$
$$E(X_i) = \{ (X_i - X_{i-1}) + [(X_{i+1} - X_{i-1}) / 2] \} / 2 \text{ , and}$$
$$E(Y_i) = \{ (Y_i - Y_{i-1}) + [(Y_{i+1} - Y_{i-1}) / 2] \} / 2 \quad (11)$$

DDTW takes moving trends of certain subsequences into account in order to judge the similarity of the two sequences. Instead of the distance between two points, DDTW is said to be more sensitive to discover local similarity of two sequences.

Among four above-mentioned methods, windowing and slope weighting are intuitive because they simply form the constraints to force the warping path not to go along the horizontal or vertical direction too much. Step patterns also try to mitigate the singularity problem by leading the warping path to cross the diagonal if the previous step goes along the X-axis or Y-axis. DDTW seems to work well on certain datasets, but it is not suitable for some cases such as sequences with great portions of empty values. The four variants of the DTW algorithm may successfully mitigate the singularity problem under specific situations.

For the four variants of DTW mentioned above, we consider that slope weighting should bring the best results for imputation. This is because slope weighting is more flexible that slightly encourages the warping path goes to the diagonal. For the microarray dataset we use, what

counts lies in local similarity of two genes. Forcing the warping path to go to the diagonal too much may mitigate the singularity problem, but it is also at the risk of losing the alignment that reveals the best mapping of two genes. Slope windowing will be effective if the window size w is small. On the other hand, imputation will be brittle as the window size w is set to be too large. Step patterns also forces the warping path with its constraints. This may results in possible loss of the best mapping. DDTW is not suitable for our microarray time series dataset. The reason is because of the great portion of missing values. Our goal is to retrieve the most proper modification of DTW so that our proposed imputation method can acquire the best results. To fit this need, we have implemented the four above modifications for DTW on our proposed method. We also compare the imputation effectiveness resulted from of these modifications of DTW in order to improve the accuracy of our proposed method. Experimental results stand by our assumptions that performing windowing and slope weighting brings a better result. The detail will be presented and discussed in Section V.

IV. DATASETS AND PERFORMANCE ASSESSMENT

In order to evaluate the effectiveness of the proposed imputation method, we implement the method on real microarray datasets. In this section, we first give a brief description about the dataset involved in our experiments. Subsequently, performance measurement of imputation methods is explained. The NRMS equation is used to assess whether the imputation method are effective or not.

A. Real Microarray Dataset

Spellman *et al*. and Cho *et al*. provided the yeast microarray dataset (http://genome-www.stanford.edu/cellcycle) [2,22]. The data was obtained for genes of Yeast Saccharomyces cerevisiae cells that were collected with four synchronization methods: alpha-factor, cdc15, cdc28, and elutriation. Spellman's dataset is widely used as the real dataset in microarray research [5,7,8]. These four subsets of the dataset contain totally 6178 gene ORF profiles with their expression values across individual amounts of time slots. For example, the alpha subset contains 18 time points with seven minutes as the time interval, while the cdc28 contains 17 time points with ten minutes as the time interval. Here we pick alpha and cdc28 sub-datasets in Spellman's microarray datasets as the testing data because they are representative samples. Alpha sub-dataset contains missing values with nearly uniform distribution, while cdc28 sub-dataset contains a great portion of missing values. These four kinds of sub-datasets record the gene expression reactions during different phases in cell cycle. However, some of the 6178 gene ORF profiles are incomplete with missing values at certain time slots which are the missing values that we are going to impute and estimate. The Spellman's raw dataset is tab-delimited matrix-liked data.

B. Assessment of Imputation Accuracy

After the imputation for missing values, we have to assess the performance of our method and the comparison with existing imputation methods. For assessment of imputation accuracy, genes with missing values in microarray gene expression data are first filtered to generate a complete matrix. As for Spellman's dataset, there are about 4304 genes in the complete matrix. Missing values with different missing rates ranging from 1%, 5%, 10%, 15% and 20% of the data in the complete matrix are deleted at random to create testing datasets. Afterward, we impute missing values in testing datasets with our method to recover the introduced missing values for each data set. The estimated values are compared to those in the complete matrix. The commonest way for the assessment of missing value imputation is to calculate the normalized root mean square (NRMS) error. The most commonly-used equation for NRMS error is as follows:

$$\textbf{NRMS} = \sqrt{mean[(y_{predict} - y_{known})^2]} \Big/ std[y_{known}] \quad (12)$$

where $y_{predict}$ and y_{known} are estimated values and known values in the complete matrix, respectively, and $std[y_{known}]$ is the standard deviation of known values. After generating complete matrices and randomly-removed testing datasets, we will impute missing values with our method. NRMS errors will also be calculated as the assessment to be compared with those of other imputation methods.

V. EXPERIMENTAL RESULTS

A. Mission Value Imputation

We combine the KNN method and the modified DTW algorithm based on FastDTW, along with four variants of DTW to impute missing values in alpha and cdc28 testing datasets. NRMS errors are then calculated as the assessment to determine whether an imputation method is effective or not. An imputation method is said to outperform other methods if and only if its NRMS value is lower than that of others. We first implement our method by using FastDTW-based DTW algorithm. Then we also experiment on the four variants of DTW for accuracy improvement. Moreover, we add the original KNN imputation method, zero imputation, row average imputation, BPCA imputation and LLS imputation in our experiment. We try to compare and explain imputation results from all mentioned methods. Imputations are performed on alpha sub-dataset and cdc28 sub-dataset in Spellman's yeast microarray datasets. Missing rates range from 1%, 5%, 10%, 15%, to 20% individually. The number of K for KNN is set from 10, 15, 20, 50, and 100. DTW with weighting value ranges from 1.2 to 1.8 because we find that the effectiveness is reduced if the weighting value is larger than 1.8. DTW with windowing parameter ranges from 2 to 5 for the same reason. For each experiment, we run 10 times and calculate the average value to reduce the randomness. We then pick out parameters that generate the best result for each method and compare the NRMS values among these methods.

B. Results and Discussion

With the experimental results, we find that the most proper parameter for each variant of DTW differs. For observation convenience, here we merely list the result of each method generated with the parameter that brings the best outcome. We observe and compare the results above and hence make some summaries. First, for all experimental results, we find that methods relative to KNN including KNN, FastDTW, and FastDTW with variants of DTW retrieve the best results when the number of K is set between K =10 and K = 20. This stands for Troyanskaya's research in 2001. As a result, while applying KNN or KNN-like methods to impute missing values in microarray time series data, setting the number of K between 10 and 20 generates the best result empirically. Assigning the value of K less than 10 or more than 20 will not bring a better result.

Besides, we find that the best result occurs when we apply our proposed method with FastDTW-based modification and slope weighting with weighted value between 1.5 and 1.8. For cdc28 sub-dataset, FastDTW with windowing and slope weighting both surpass other imputation methods. This indicates that DTW works well with slightly weighted value that forces the warping path to the diagonal direction. But if we put too much force, it will generate unfavorable results. As for the various missing rate of experimental dataset, we find that our proposed method with FastDTW-based modification works better than the traditional KNN imputation method in most cases. Moreover, if we add proper variants such as slope weighting with weighted value between 1.5 and 1.8, or windowing with window size = 2 for the improvement of DTW, it will bring better results. The proposed DTW imputation method outperforms other methods including BPCA, especially when the missing rate is large such as 15% or 20%.

Moreover, there is an interesting discovery. We find that applying DTW with Step Patterns and DDTW even makes the imputation results worse on the contrary. The reason may possibly because DTW with Step Patterns is not suitable for handing microarray dataset. Because it forces the warping path to across only the limited region it forms so that expectative alignments of two genes may hence lose. The same situation happens on DDTW. DDTW could also generate the warping path that DTW is not originally supposed to generate because DDTW takes vectors of time slot points on two time series of DTW into consideration. However, two regulatory genes may only have reactions on a few time points of the whole sequences. In this case, measuring the distance between two time points is more proper than calculating vectors of them. Furthermore, DDTW fails to work in the case that a portion of empty values exist just like what we find in the microarray dataset. Experimental results support our assumptions mentioned in previous sections that

performing windowing and slope weighting improves the accuracy of the proposed method. In order to reduce the problem of singularity, several techniques are proposed.

As shown in Fig. 2, average imputation and zero imputation seem to be brittle. These two simple imputation methods provide limited help. The imputation method that only utilizes KNN with FastDTW achieves better results than using KNN. This proves that taking DTW distance as the similarity measurement is more suitable than taking Euclidean distance while handling microarray time series data. However, if we modify the algorithm with step patterns or DDTW, the accuracy of imputation decreases. This shows that it is improper to apply step patterns and DDTW for the dataset. The reason is discussed in previous paragraph. BPCA and LLS seem to outperform KNN and other brittle imputation methods. Using FastDTW with windowing results in better results than LLS brings. The most effective method is using FastDTW with slope weighting

that slightly outperforms than BPCA. Sequences of effectiveness of these imputation methods may change a little bit in certain percentage of missed data. This may result from the randomness while deciding which values to be removed in the complete matrix.

Fig. 3 illustrates almost the same situation as Fig. 2. Basically results of all imputation methods are worsened a little. This is because the cdc28 sub-dataset contains more missing values than the alpha sub-dataset. Theoretically, NRMS error increases while missing values are getting more in the dataset. We can also see that using FastDTW with windowing and slope weighting both outperform BPCA. Furthermore, even using FastDTW brings better results than BPCA when the missing rate is larger than 15%. This shows the weakness of BPCA while dealing with microarray time series dataset with a large portion of missing values. To summarize, using our proposed method with the variant of slope weighting can retrieve the best imputation results.
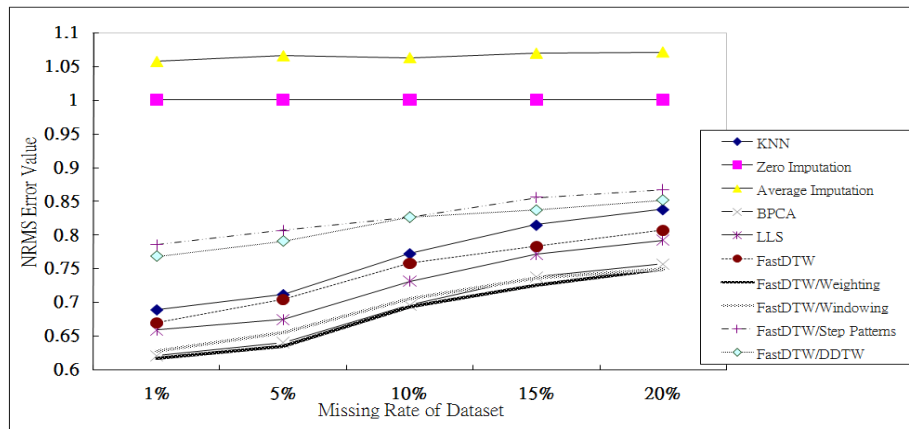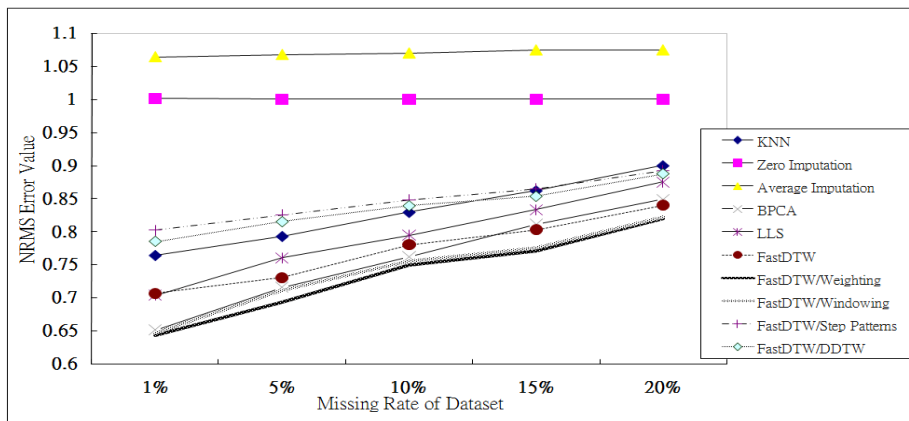

Figure 2. Imputation results of alpha dataset

Figure 3. Imputation results of cdc28 dataset

VI. CONCLUSION

Missing value imputation is very important in microarray gene expression time series data. In this paper, we propose a novel method that combines the traditional KNN method with the DTW algorithm to perform the imputation. We also implement variants of DTW both

for efficiency increasing and accuracy improvement to achieve better imputation results. Experimental results show that our proposed method with the DTW variant of slope weighting outperforms other imputation methods in terms of accuracy assessment. In the future, we aim to take external information for genes such as annotations from gene ontology into consideration to further improve

the imputation method. We believe our approach facilitates research for microarray gene expression data.

REFERENCES

[1] J. DeRisi, R. Iyer, and Brown P, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol.278, pp.680-686, 1997.

[2] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Mol. Biol. Cell*, vol.9, pp.3273-3297, 1998.

[3] D.S.V. Wong, F.K. Wong, and G.R. Wood, "A multi-stage approach to clustering and imputation of gene expression profiles," *Bioinformatics*, vol.23, pp.998-1005, 2007.

[4] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect in the classifier accuracy," *Classification, Clustering and Data Mining Applications*, pp.639-648, 2004.

[5] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol.17, pp.520-525, 2001.

[6] H. Kim, G.H. Golub, and H. Park, "Imputation of missing values in DNA microarray gene expression data," in: *Proc. of the IEEE Computational Syst. Bioinformatics. Conf.*, pp.572-573, 2004.

[7] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol.19, pp.2088-2096, 2003.

[8] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: local least squares imputation," *Bioinformatics*, vol.21, pp.187-198, 2005.

[9] X. Wang, A. Li, Z. Jiang, and H. Feng, "Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme," *BMC Bioinformatics*, vol.7, pp.1-10, 2006.

[10] A.C. Yang, H.H. Hsu, and M.D. Lu, "Outlier filtering for identification of gene regulations in microarray time-series data," in: *Proc. of the 3rd Intl. Conf. on Complex, Intelligent and Software Intensive Syst.*, pp.854-859, 2009.

[11] V.S. Tseng, L.C. Chen, and J.J. Chen, "Gene relation discovery by mining similar subsequences in time-series microarray data," in: *Proc. of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biol.*, pp.106-112, 2007.

[12] C. Furlanello, S. Merler, and G. Jurman, "Combining feature selection and DTW for time-varying functional genomics," *IEEE Trans. on Sig. Processing*, vol.54, pp.2436-2443, 2006.

[13] H.M. Yu, W.H. Tsai, and H.M. Wang, "Query-by-singing system for retrieving karaoke music," *IEEE Trans. on Multimedia*, vol.10, pp.1626-1637, 2008.

[14] C. Myers, L. Rabiner, and A. Roseneberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol.ASSP-28, pp.623-635, 1980.

[15] L. Rabiner, A. Rosenberg, and S. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol.ASSP-26, pp.575-582, 1978.

[16] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol.11, pp.561-580, 2007.

[17] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol.ASSP-26, pp.43-49, 1978.

[18] D. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time Series," in: *Proc. of the Workshop on Knowledge Discovery in Databases*, pp.359-370, 1994.

[19] J.B. Kruskall and M. Liberman, "The symmetric time warping algorithm: from continuous to discrete," in: *Time Warps, String Edits, and Macromolecules: The theory and Practice of String Comparison*, 1983.

[20] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol.ASSP-23, pp.52-72, 1975.

[21] E. Keogh and M. Pazzani, "Derivative dynamic time warping," in: *Proc. of the 1st SIAM Intl. Conf. on Data Mining*, pp.1-11, 2001.

[22] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, and D. Lockhart, "A genome-wide transcriptional analysis of the mitotic cell cycle," *Mol. Cell*, vol.2, pp.65-73, 1998.

Hui-Huang Hsu is an associate professor of Computer Science and Information Engineering at Tamkang University, Taipei, Taiwan. He received his Ph.D. degree in electrical engineering from the University of Florida in 1994. His current research interests are in the areas of machine learning, data mining, bioinformatics, and multimedia processing. Dr. Hsu is a senior member of the IEEE

Chao-Hsun Yang has an alias of **Andy C. Yang.** He is a Ph.D. candidate in the Department of Computer Science and Information Engineering at Tamkang University, Taipei, Taiwan. He received his M.S. degree from the Department of Computer Science and Information Engineering at Tamkang University, Taiwan, R.O.C., in 2005. His research interests are in bioinformatics and computer algorithms.

Ming-Da Lu is a Ph.D. student in the Department of Computer Science and Information Engineering at Tamkang University, Taipei, Taiwan. He received his M.S. degree from the Department of Computer Science and Information Engineering at Tamkang University, Taiwan, R.O.C., in 2008. His research interests are in bioinformatics and Internet Applications.

Insight of Value-oriented Human Relationship

Benjamin Penyang Liao

Department of Information Management, Overseas Chinese University,
No: 100, Chiao Kwang Rd., Taichung 407, Taiwan, R.O.C.
Email: benjaminpyliao@gmail.com

Abstract-- **Human relationship is important for everyone. The relationship connection is generally constructed by the value production. The insight of value system is firstly described with the discussions of value definition, framework of value system, value evaluation with positive and negative measures, and value characteristics. Then a framework of value-oriented human relationship is analyzed with the emphasis of the two value lines discussions. Finally, the insight of human relationship influence spectrum is discussed. The influence spectrum includes seven phases, which are uninfluenced phase, solvability phase, affect phase, confirmation phase, registration phase, exaggeration phase, and extreme phase.**

Index Terms-- **value system, value measures, value-oriented human relationship framework, human relationship influence spectrum**

I. INTRODUCTION

We all live in societies and build some relationships with society members. These relationships will affect each society member's life everyday in many circumstances including working and living cases. In working case, customer relationship management (CRM) for example is an important task for an organization to continuously attaining benefit from customers by maintaining health relationships with their customers. In living case, marriage relationship management, for example, is an important task for couples to persistently obtaine loves and something else from each other by maintaining robust relationships between each other. We all know personal interrelationship is very important. Analyzing human relationships from value perspective is a good way to intrinsically digging out the coherent relationship components and features. In this paper, value-orientation is the intrinsic basis for inspecting the insight of human relationships.

The term of value in current time seems to get into ambiguous usage and interpretation in current days for most of people. Value, most of time, implies the price. For example, that apple 's value is worthing $2, which is considered in cost perspective from seller's viewpoint. However, the insight of value should contains more meanings. A value should be considered both from buyer viewpoints and seller viewpoints, especially for the buyer's viewpoints. For example, the apple's buyer can eat the bought apple to satisfy some buyer's needs, such as, hungry. In fact, needs satisfaction is the kernel of the term of value. In order to reveal the insight of human relationship from value perspective, this paper first inspects the insight of value system.

In most of time, an individual has to obtain something from outside to satisfy his/her needs. In such, in most cases, the exchange or transaction is invoked. When an exchange is occurred, at least two participants involve to interact with each other. In the simplest case, a buyer and a seller come together to take the exchange. The basic reason which causes the exchange is that each participant can get values from this exchange. That is, each participant tries to satisfy each one's needs from this exchange. Therefore, obtaining values from exchange is the kernel of exchange for each participants. In addition, it is obvious that there are two value lines to be produced, where each value line is operated exclusively for each participant. In order to get values from each exchange participant, a connection line has to be constructed, which will finally construct a relationship between participants. In addition, managing the relationship can in fact be considered maintaining these two value lines. This means that maintaining these two value lines are the kernel of human relationships. Therefore, the insight of human relationships from the value line perspective will be emphasized.

When the human relationship is going on progress, the relationship intensity will be changed. Different relationship intensity will affect the will and way to exchange for both two participants. Systematically to analyze the relationship intensity shown by a human relationship influence spectrum then is necessary. Anyone can apply this spectrum to realize the relationship with the others. In order to see the insight of human relationship, the analysis of the relationship influence spectrum is important. Therefore, This paper then inspects the human relationship influence spectrum with some new complements.

Remaining sections are arranged as follows. Section II discusses the insight of value system. Section III creates the models of humans relationship with the emphasis of value line concept. Section IV analyzes the insight of human relationship influence spectrum. Finally, Section V gives some concluding remarks.

II. ISIGHT OF VALUE SYSTEM

In this section, the value definition is discussed first. A

general value system framework is analyzed. After that, value evaluation for positive value and negative value is discussed. Finally, value characteristics are discussed.

A. Value definition

Axiology is a field of study for value in academy and can be divided into three perspectives: subjective, objective, and gestalt perspectives. No matter which perspective, subject (people) and object are discussed implicitly or explicitly.

From subjective perspective, the value is the worth attached to an object judged by a person when s/he evaluates the object as pleasure, desired or interested. Values are not some kinds of quality that exist inside the object, but are determined by the subject's valuation on the object.

From objective perspective, the value represents the criterion to make the valuation. Values are independent of the subject or even the object. The subjectivists claimed that, if values exist outside the domain of our valuation, how can we know the existence of such kind of values? However, the objectivists argued that we must distinguish between valuation and value. The key point is that value is prior to valuation. If there were no values, what would people evaluate [2]? That is, it is meaningful only when there is something (value) that can be found when someone is searching for or evaluating for.

As a gestaltist, Frondizi [2] pointed out that the psychological experience of pleasure, desire, and interested is the necessary condition but not the sufficient condition for values. On the other hand, that the psychological experience does not exclude the objective aspects of values, but instead confirming their existence. Therefore, values have both objective and subjective aspects; they reside in the relationship between the subject of valuation and the object of values, and do not exist apart from them. Therefore, the value is gestalt, which is not only the combination of each component, but it also the uniqueness of the totality composed from these separated parts.

Chen and Chen [1] then defined the value with a similar gestalt perspective as follows.

> *A value is of that a behaving subjective entity attains the subjective confirmation of needs satisfaction by the interaction with objective entities.*

In such, there are four key value factors which are subject needs, subject intention, object competence, and subject satisfaction. That is, a subject should arouse her/his needs and intend to be satisfied by capable objects with results of subject satisfaction.

B. Value system

There are many value classifications from many perspectives and can be found in [3][6]. Liao [4][5] proposed a framework of value system based on above value definition, as shown in Fig. 1. A value system comprises three components which are subject, object, and values. The subject component contains needs and valuation measures. The object component contains its attributes, where each attribute value represents the

object's competence that can potentially satisfy some subject needs. The value component is the results of match or interactions between subject needs and object competences.

The subject needs are categorized into physical needs, spiritual needs, and resources needs. The physical needs include sensory needs such as needs of eye, ear, nose, tongue and skin (or body), and safety needs which represent the needs without suffering hurts for physical body. The spiritual needs include affection needs, knowledge needs, and completeness needs. The affection needs represent, for example, the needs for attaining good emotion and belongingness. Pleasure and happiness are two instances of affection needs. Love need is an example of belongingness. The knowledge needs represent the needs for satisfying curiosity and knowing facts or rules that can be applied to solve some problems. The completeness needs represent the needs to attain the wholly states for her/him-self. The resources needs include the needs of tangible resources such as human, material and financial resources, and the needs of intangible resources such as time, brand equity, human relationships, and so on. The resources initially represent the role of object, which can satisfy subject needs. The resources needs imply that a subject can store these resources into warehouse for conveniently satisfying subject needs in the future.
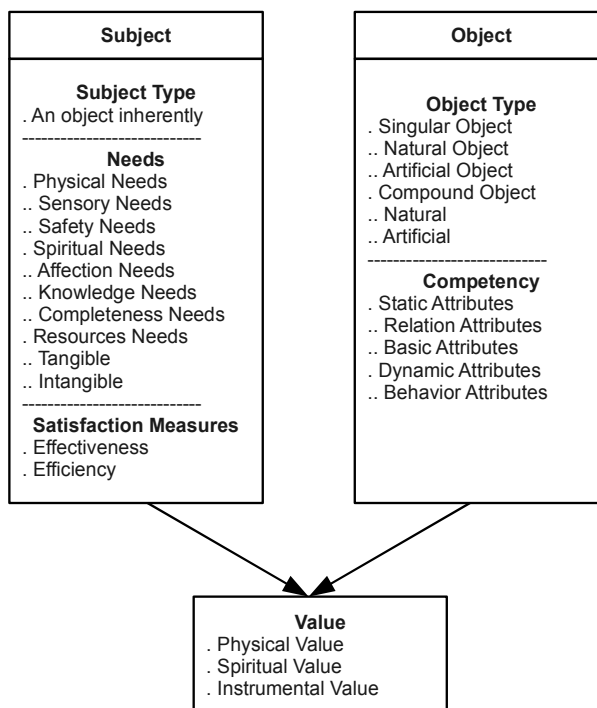

Figure 1. A framework of value system [4][5].

The object competences are classified into static attributes and dynamic (or behavior/ operation) attributes. The static attributes include relation attributes and fundamental attributes. The relation attributes basically depict the relations for external objects and internal components, which further include dependency, generalization, association, and realization attributes. The

association attributes can further be classified into composition and aggregation association attributes. The basic attributes show the object's unique characteristics and are not the characteristics of relation attributes. For example, flying attribute is an airplane's basic attribute which is synthesized by many parts, such as engine, wings, and so on. However, each part can not fly itself. The dynamic attributes show the object's behaving characteristics.

Finally, all object attributes can potentially satisfy some subject needs. The values are produced by the match or interactions between subject needs and object competency. In addition, the values are classified into physical values, spiritual values, and instrumental values. Therefore, the match of physical needs and object competency results in physical needs, and the match of the spiritual needs and object competency results in spiritual needs. However, the instrumental value results from the instrumental role of objects applied by subjects. An instrument means that it can be applied to find terminal objects or other instrumental objects for subject. For example, a fishing net is an instrument which can help a subject to catch a fish (terminal object) to satisfy a subject's physical needs.

C. Value structure

To analyze a value, we also can analyze its structure and characteristics. Here the value structure is discussed first.

A value structure includes its constructed components and the relationships among the components. The value components include subject need and capable object. The relationships among the value components include component connection and interactions between the components.

For component connection, there are one connection between the subject and the object. In this connection, the subject plays the need role which some needs are intended to be satisfied, or from the other viewpoint, the demand role that the subject demands something (ie., objects) to satisfy the needs. On the other side, the object plays the function role that it is capable to satisfy the subject's need.

For interactions between the components, the subject applies or uses the object to satisfy the subject's need by filling the need attributes until reaching the balance state. The object can help the subject to recover the balance state. Generally, the subject is actively to invoke the interactions to the object. In addition, the way of applying object may be confined by the object's features such as the usage guidelines set up by the manufacture, or followed by the subject's preferences.

D. Value characteristics

Value characteristics show the other states, viewpoints, and derived descriptions that the value structure does not address. The value characteristics are then discussed as follows.

1. Gestalt. Gestalt means that the discussed target is composed by its components and can not be described only from any component's viewpoint.

That is, the discussed target should be described by considering all its components in the same time. Discarding any component will destroy the discussed target. In such, a value is consists of its components, which are both the subject need and object capability. There is no value occurred while lacking one of the value components.

2. Subject's intention of need satisfaction. If a subject does not have the intention to satisfy its needs, no values will produced.

3. Object replaceability. The object currently been applied by the subject can be replaced by any other object that is capable to satisfy the subject need.

4. Need variety. This means that a subject has many needs.

5. Need dynamic. This means that the objective of need is always changing. The objective of need can be considered as the balance state of some attributes. A need means that some attributes are under unbalance state and the subject intends to recover from the unbalance state. However, the balance state of an attribute sometimes is defined by the subject with preference. The subject preference may be changed affected by some factors. In such, the balance state of an attribute may be changed under some circumstances.

6. Need grouping. Need grouping means that some subject's attributes are related and then grouped. In addition, the related attributes will affect each other under some ways. Therefore, it is possible that to invoke an attribute's unbalance state will also cause the unbalance state of the other related attributes.

E. Value evaluation

Values are produced by the match of subject needs and object competency. There are two types of measures for valuation: effectiveness measures and efficiency measures. The effectiveness measures show the degree of subject needs satisfaction by object competency, which include sub-measures of value depth, value width, and value quality. The value depth measure can be implemented by absolute values and relative values, eg., ratios. The value width measure includes its sub-measure of individual value width with multiple needs satisfaction, and sub-measure of group value width which many individuals can be satisfied by a single object competency. The value quality measure includes its sub-measures of consistency and longevity (or duration).

The efficiency measures show the degree of the cost that a subject has to pay to attain the values. The efficiency measures include sub-measures of easiness and procedure. The easiness measures include its sub-measures of cost, object scarcity, and total demand of object. The procedure measures include its sub-measures of easiness, stepwise, and quality. The stepwise measures show that whether the subject needs can be satisfied from each step/phase or until the final step. The quality measures also include the consistency and longevity

measures.

In general, the value is a positive term for most people , since that it shows an object can satisfy some subject needs. However, in some cases, when a subject tries to acquire and apply an object to satisfy its own needs, it might not obtain positive results but negative ones. The negative results can mean that a subject has paid some cost to acquire and apply an object but it still can not satisfy one's own needs and even hurt oneself. The negative result can be regarded as a loss or a negative value for subjects. The papers in [4][5] did not consider the situation of negative value. Therefore, this paper complements this negative value concept to the original contents of value system.

If we regard the value as a variable, it might contain positive contents and negative contents. In positive value perspective, as the same as the propositions of previous papers, there are three positive value types, which are positive physical value, positive spiritual value and positive instrumental value. On the other hand, there is the corollary of three negative value types, which are negative physical value, negative spiritual value, and negative instrumental value. The negative physical value therefore implies the suffering of physical loss. The negative spiritual value then implies the suffering of spiritual pain or discomfort. The negative instrumental value here means that the object can not play well the role of instrument to help the subject to find out suitable terminal objects. This object may even hinder a subject to find out the suitable terminal object.

In sum, the term of value generally contains the types of positive value and negative value, although most people regard the term of value as the type of positive value.

In such, the value measures can be divided into positive value measures and negative value measures by dimensions of performance type and value sign type. This can be shown in TABLE I. The performance type can be effectiveness or efficiency with the contents as described above. The value sign type can be positive or negative. Therefore, there are four types of value measures, which are positive value effectiveness measures, positive value efficiency measures, negative value effectiveness measures, and negative value efficiency measures.

TABLE I.
VALUE MEASURES

		Value Sign Type	
		Positive Value	Negative Value
Performance Type	Effectiveness	Positive value effectiveness measures	Negative value effectiveness measures
	Efficiency	Positive value efficiency measures	Negative value efficiency measures

Generally, a subject gets terminal object by interchanging resources with the other subject. That is, each subject contains some resources (objects) for that can satisfy each other subject needs. Therefore, in the

process of value generation, exchange process creates a connected relationship between exchangers. In the meanwhile, value generation has to be considered when discussing the subject relationships.

III. MODELING HUMAN RELATIONSHIP

A. Definition of human relationship

Basically, a relation can be defined as a connection between entities, such as people. However, this definition is too general to depict its contents. There are some features of this connected relation. Connection influence is the key point to describe the relation connection in this paper. Therefore, this paper defines a human relationship as follows.

A human relationship is a reciprocal influence connection between two connected human entities.

In such, there are some features of human relationship.
1. Connectivity. There are two human entities connected by some ways and may be under different circumstances. Without this connection, there is no human relationship created.
2. Interaction. When two human entities are connected each other, this connection line just like a road which can transmit something from one side to the other side. That is , this connection line provides an interaction or exchange channel. Reversely, each entity intends to interact each other and in such that creates this connection line.
3. Influence. Each connection line will help entities interact each other. Each interact will cause some changes of each entity. For example, each entity will both pay something in order to get the other things from the other exchangers. This exchange then will result in an influence between these exchangers.
4. Dynamic. The relationship can be changed. That is, this connection line can be cut off. Conversely, this connection line may be promoted. Therefore, the relationship influence can also be promoted.

In order to get more understanding of relationship, there are some important relationship attributes should be analyzed, which are connection hook, constructed material, and intensity.

There are two connection hooks in an connection line. Each connection hook link to an entity. The major constructed material of a relation connection line is value. The intensity can be measured from dimensions of thickness, flexibility, and loyalty.

B. Framework of value-oriented human relationship

Since a human relationship is an interaction connection, interchange is the essential behavior. The reason of interaction then is the entity's needs satisfaction. With applying the value definition, the purpose of interaction within a relation connection is trying to create values. That is, each human entity tries to

exchange their own resources each other in order to satisfy each one's needs. Fig. 2 shows a framework of value-oriented human relationship.
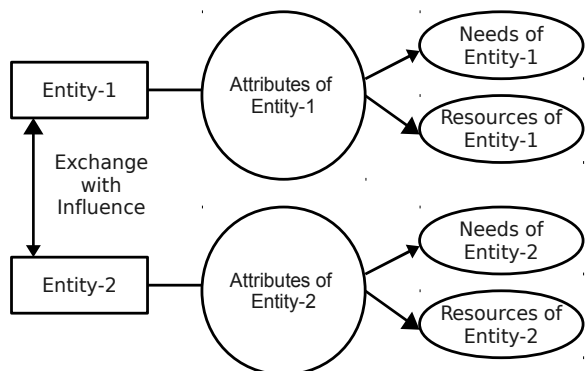

Figure 2. A framework of value-oriented human relationship.

Each human entity contains many attributes. Each entity can be described by two different perspectives, which are subject perspective and object perspective. That is, each entity can concurrently play two roles, which are the subject role and the object role under the same entity attributes. When an entity is analyzed from subject perspective, the discussion focus is on the subject need and all entity's attributes are regarded as the needs attributes. When an entity is analyzed from object perspective, the discussion focus is on the object competence and all entity's attributes are regarded as the resources attributes. An exchange will only occur when each exchanger has some needs intending to be satisfied from acquiring the objects owned by the other exchanger. Therefore, if an entity does not intend to satisfy its needs, the exchange will not occur. As described in the definition of human relationship, the exchange will produce influence for each participants. The features of connection influence in fact can show the types and intensity of human relationship.
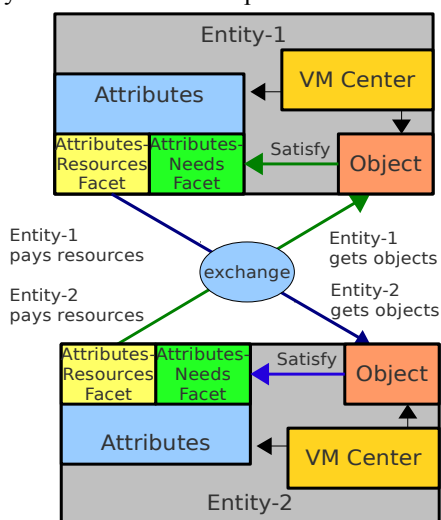
Figure 3. Insight of the framework of value-oriented human relationship.

A relation connection line in fact contains two sub-connection lines. Each sub-connection line can be viewed

from the entity's need facet. That is, how an entity's needs are satisfied by the other entity's resources. Since there are two entities, it then creates two value lines, as the two crossed lines between two entities in Fig. 3. Each value line can show a need-satisfying process. Fig. 3 also points out there is a VM Center which is a value management center. Fig. 4 shows the contents of the VM center. A VM center will search targets and coordinate with the target's holder and then decide whether to exchange with the target's holder. If the decision is willing to have an exchange, then it goes to the attain and pay stages. After attaining the targets, an entity can make the rebalance for needs. After that, evaluation for this whole process will be analyzed and reviewed. The evaluation results will affect the relationship intensity between the entity and the target holders.

A VM center will manage the targets. First the VM center will determine whether some resources play the role of objects. After that, the VM center applies the objects to satisfy entity's needs. After that, a VM center will evaluate the object's performance.
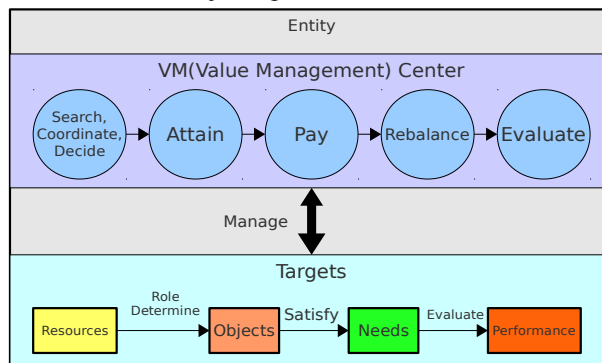

Figure 4. An entity's value management center.

IV. INSIGHT OF HUMAN RELATIONSHIP INFLUENCE SPECTRUM

A. Human relationship influence spectrum

A relationship connection line is mainly constructed by values in an accumulated way. The essential value types are physical values and spiritual values. Physical values are those that physical needs are satisfied. Spiritual values are those that spiritual needs are satisfied. The value content in fact implies the allowed transmitting stuff in a connection line. These constructing material of value for relationship will affect a connection's thickness, fragility, growth rate, etc. This construction features of human relationship can give us more pictures of human relations.

However, it still neglects some insights of human relationships. What we want to know is the connection influence. That is, a relation connection just likes a communication channel which can cause some exchanges between the connected participants. Each exchange will cause some changes and this means that this connection line can influence these exchanging activities.

As analyzed above, the obtaining value is the essential constructing material of relation connection. Logically, more value obtaining will cause more relation influence

and less value attaining will cause less relation influence. But this still can not depict a clear picture of influence degree.

Generally, a human is a compound entity of physical body and spiritual soul. And most of us believe that the soul dominates the physical body even though the soul are often affected by the physical body with continuously changing states. Since the spiritual soul is the master of physical body, we can inference that when a value is more related to soul satisfaction, it will cause more influence to the entity. Therefore, the influence of spiritual values is bigger than the influence of physical values.

In addition, there are many spiritual features or needs attributes with different layers, such as Maslow's needs hierarchy, which are physical layer, security layer, belongingness layer, esteem layer, and self achievement layer. This means that a need satisfaction can be analyzed by the degree or process of relation carving. Therefore, the influence degrees can be described by seven influence phases. This seven influence phases or degrees are the uninfluenced phase, solvability phase, affect phase, confirmation phase, registration phase, exaggeration phase, and extreme phase, as shown in Fig. 5. Since the

influence direction can be positive and negative, the above phases can simultaneously be applied to these influence direction. This results in a human relationship influence spectrum, as shown in Fig. 6. Each influence phase with the influence spectrum is discussed as follows.

1. Uninfluenced phase. In this phase, there is no any exchange occurred between entities. Therefore, there is no any influence between entities.

2. Solvability phase. In this phase, each entity has some problems to be solved. That is, each entity wants some needs to be satisfied or recovered from unbalanced state. In this phase, each entity has little understanding of other entity. Therefore, the focus is only on specific physical resources and the focus scope is narrow. Therefore, the main interchange is of physical material exchange. In this phase, each entity does not put or project their emotion to the other entity. Sometimes, we can regard the human relationship in this phase is functional. That is, the considering focus is whether this relationship is of function or of malfunction.
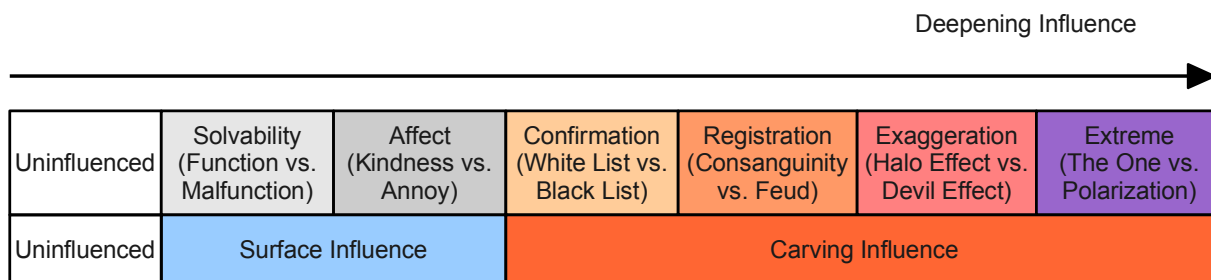

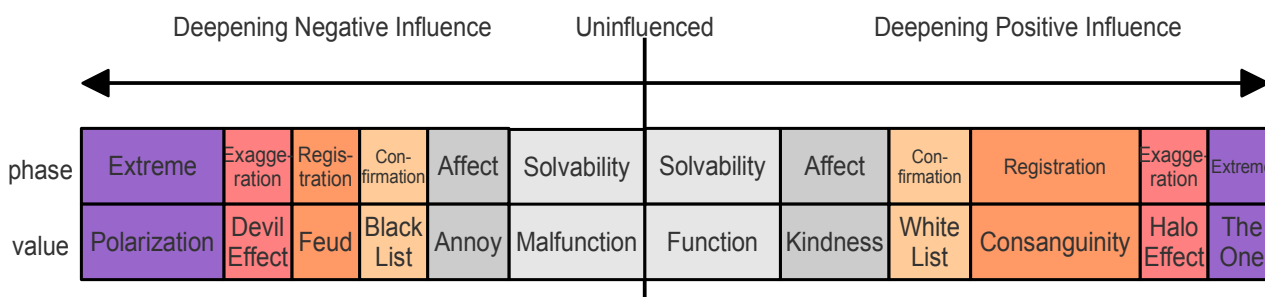Figure 5. Seven general deepening phases of human relationship influence intensity

Figure 6. Human relationship influence spectrum.

3. Affect phase. In this phase, some feelings are exchanged. Each entity recognizes the other guy's gentleness or rudeness in the process of major needs satisfaction. But this feeling of gentleness or rudeness is only surface. This feeling is likely to be emotional or temporal. That is, in this affect phase, each entity only obtained shallow kindness or annoy.

4. Confirmation phase. In this phase, each entity writes down the other entity's name on her/his

white list or black list. For the entity in the black or white list, an entity would like to pay more energy to deal with the relationship with the other one. In this time, each entity will frequently and autonomously scan the other guy's current information and take some responses by then. Enrolling a guy on a white list or black list depends on confirmed benefits or harms for each other. However, in this phase, the affect intensity is still shallow but is stronger

then that in affect phase.

5. Registration phase. In this phase, the affect intensity is increased for the entities on the white list and black list. Registration means that some specific guys on the list are deliberately picked out. From value perspective, this means that this relationship will cause significant benefits or harms each other. Under this big significance perception, the registered guys will be written on a VIP list or criminal wanted list. This will produce the relationships which are similar to consanguinity or feud. In the consanguinity relationship, connected entities are family members, just like each member has the kinship or with blood-relations. On the other hand, in the feud relationship, connected entities are inherited enemies. In this case, each entity's emotion will easily be invoked or provoked. Therefore, responses will be quickly emitted. In this feud case, each connected entity will trap into the prejudice fighting with quick and strong responding for each other.

6. Exaggeration phase. In this phase, perceived values for each connected entity will spread out. This will cause the state of halo effect or devil effect. Basically, the halo effect describes the situation of one guy will amplify the other one's other capability after confirmed the other guy's specific competence. Reversely, the devil effect describes the situation of one guy will also amplify the other one's other drawbacks after confirmed the other guy's specific inability. Therefore, within exaggeration relation, each connected entity will amplify perceived values from each other.

7. Extreme phase. In this phase, connection entities will go through two differently extreme routes. One route with positive direction is that an entity is to be an incarnated agent for each other. Then these two entities can be regarded as the one. That is, what one thinks is the same as what the other one thinks. On the other route with negative direction, perceived recognitions are totally reverse each other within the connected entities. That is, what one thinks is totally the reverse as what the other one thinks.

In addition, these seven phases, can be grouped into three general phases, which are uninfluenced phase, surface influence phase and carving influence phase. The surface influence phase includes the solvability phase and the affect phase. The carving influence phase includes the confirmation phase, registration phase, exaggeration phase, and extreme phase. In the surface influence phase, the relationship intensity is shallow. In the carving influence phase, the relationship intensity will be more and more stronger.

B. Phase transition

The human relationship influence spectrum describes the deepening influences by seven phases. This study proposes that there are two factors will affect the phase transition, which are realization and belief.

Realization means that each connected entity can really get the correct and necessary information about the other guy. More understanding will help an entity to make correct value judgment. More correct value judgment will intensify the connected relationships.

However, the value judgment in fact will be affected by each entity's belief. A belief is that something an entity thinks right or wrong. A belief will cause an entity to likely obtain something perceived correct to satisfy one's needs. On the other hand, a belief will cause an entity to likely discard something perceived wrong to prevent one's attribute from unbalanced state. A belief sometimes can be regarded as an entity's value viewpoint. However, each entity's beliefs for something is perceived different. In addition, each entity's beliefs are generally in persistent state. That is, a belief is not easy to be changed. Changing one's belief will take times. Therefore, the belief will affect the speed of phase transition.

C. Influence scope

The influence scope can be discussed in degree of width and depth.

The influence width implies the number of needs attributes that are to be satisfied. With the connection relationship with higher influence width, it will cause more resources exchange to make more needs satisfaction. In addition, this study proposes that higher phases will show higher influence width.

On the other hand, the influence depth means how much a need is satisfied. A more influence depth implies that it will get more satisfaction for that need. In addition, this study proposes that higher phases will increase higher influence depth.

D. Why relationship is changed?

Basically, when the performance of two value lines in the relationship line is changed, the relationship will be changed. The reasons that makes the relationship changed include the resource quality and the resource shortage.

E. Characteristics of human relationship influence spectrum

The characteristics of human relationship influence spectrum are discussed as follows.

1. Bi-direction. This means that there are two different influence direction of human relationship. One is positive influence direction and the the other one is negative influence direction.

2. Changeability. This means that the influence direction can be changed. The preliminary description of influence changing feature can be reviewed in the sub-section of phase transition. The changeability also shows the features of dynamic. The dynamic means that the influence intensity is not fixed all the time.

3. Stronger influence intensity is not easier to be changed. Such as, when the influence intensity is on the the negative polarization of extreme phase,

an entity already has deepening belief of worst impression or image of the other entity. It is hard to change one's deepening belief. Therefore, when the the influence intensity is stronger, this intensity is not easier to be changed.

V. CONCLUDING REMARKS

We all live in a society and will naturally interact with the other people. These interactions will build some relationship connections among connected people. The essential purpose of interaction among people is to get values. Therefore, it is important to discuss what is a value first while talking about the human relationships. This paper then first discuss the contents of value system. By above value discussions, a value can simply be regarded as a need satisfaction by applying object competences. Therefore, the kernel of human relationship is showing that each connected people makes exchanges with their own resources (objects) in order to satisfy each one's needs. A framework of value-oriented human relationship is built and discussed under the combination of value perspective and interchange perspective, especially the emphasis of two value lines in the relationship connection line.

In addition, this study analyses the human relationship starting from the constructing material for the connection line. Meanwhile, based on above discussions, a human relationship influence spectrum with seven phases is discussed. This spectrum could provide a good insight of human relationships. Finally, the influence phase transition, influence scope, and characteristics of human relationship influence spectrum are discussed.

The human relationship influence spectrum can be applied to the case of agent interaction, such as resonance or interaction among entities. This spectrum can provide an interaction basis and guidelines for everyone.

REFERENCES

[1] B. -C. Chen and X.-M. Chen, *Value Sociology*, Laureate, Taipei.1990. (in Chinese)
[2] R. Frondizi, *What is Value? An Introduction to Axiology*, La Salle, Ill: Open Court, 1971.
[3] P. -Y. Lee, *The Nature of Values and Valuations: Theory, Measurement, and Application*, PhD. Dissertation, University College London, University of London, 1996.
[4] B. P. Liao, "A Framework of Value Management," *The First Conference of Innovation & Management,* Taipei, paper no. 457, 2004.
[5] B. P. Liao, "A framework of value-oriented innovation system," paper no. P0134, 2007.
[6] M. Rokeach, *The Nature of Human Values*, New York: Free Press, 1973.

Benjamin Penyang Liao was earned his PhD of Information Management from National Central University, Taiwan, at 2006. His mainly research fields are value management and herd behavior by applying the information technology.

He is current an associate professor of Department of Information Management, Overseas Chinese University, Taiwan.

Development of Semantic-CBR Framework for Virtual Enterprises in Project Management

Chouyin Hsu

Associate Professor
Department of Information Management,
Overseas Chinese University
Email: hkelly@ocu.edu.tw

Abstract—The information communication technology definitely encouraging global collaboration among enterprises on a project basis makes virtual enterprises possible and popular. However, the existing constraints of virtual enterprises, such as distributed resources, heterogeneous data and different definition, easily resulted in many difficulties as managing a project. Accordingly, the semantic support becomes rather important as solving these problems. Therefore, Semantic-CBR framework is proposed for improving the communication and integrating the project experiences for virtual enterprises. OWL and RDF are applied for the implementation of the semantic structure of the framework. Moreover, case-based reasoning is utilized for selecting the related project experiences according to the user requirement. Therefore, the semantic support increases the understanding for virtual enterprises and enhances the facility of reasoning mechanism for project management in the research.

Index Terms—Case-Based Reasoning; Virtual Enterprises; Semantic Web Service

I. INTRODUCTION

Virtual Enterprise is regarded as the most competitive management model of enterprises that faces the resource of the globe. Globalization leads to an efficient new business paradigm of virtual enterprises, where companies increasingly concentrate on their core competencies and outsource all other functions to their partners on a project basis. A temporary network aims to share

skills, resources, costs and benefits to achieve one or more projects answering to the market opportunities for products and services [16]. The advanced information technology is certainly important as processing a project in virtual enterprises since the distributed project resources and the inevitable communication barriers. Fortunately, Web services are changing the way applications communicate with each other on the web. They promise to integrate business operations, reduce the time and cost of web application development and maintenance as well as promote reuse of code over the Internet. Moreover, semantic descriptions are increasingly being used for exploring the automation features related to web services. The meaningful content and service are helpful for accurate searches as browsing the overwhelming information on Internet. Therefore, the semantic representation is rather emphasized in the research for responding the adequate project experiences for supporting further project management in virtual enterprises.

We propose the Semantic-CBR framework for integrating the previous project experiences with consistent and meaningful descriptions. Moreover, Case-based reasoning is applied for modeling the use and retrieval of previous project experiences. Case-based reasoning is an artificial intelligence methodology that uses specific encapsulated prior experiences as a basis for reasoning about similar new situations [1]. Moreover, the case structure in semantic representation is reliable for organizing the consistent content of project experiences. Therefore, developing reasoning mechanism for accurate search result becomes essential and promising mission in the proposed framework.

II. VIRTUAL ENTERPRISE AND PROJECT MANAGEMENT

Virtual enterprises are temporary networks of independent companies, which cooperate on a short-term basis for a certain project and are perceived to be a single unit from outside.

Internally the companies act as partners and bring in their core competences in a synergetic way [1]. As existing organizations are challenged by new entrants using direct channels to undercut prices and increase market share, solutions have to be found that enable organizations to successfully migrate into the electronic market [3]. The advantage of virtual enterprises provides a form of cooperation of independent market players (enterprises, freelancers, authorities etc.) which combine their core competencies in order to manufacture a product or to provide a service. The innovative concept of virtual enterprises combines large pool of distributed resources and flexibility to adapt to the turbulent markets.

Virtual enterprises are mostly formed to work on a single project. Long term virtual enterprise need to be provided with an own identity and profile to offer their services to the market. A project is a temporary endeavor, having a defined beginning and end, usually constrained by date, but can be by funding or deliverables, undertaken to meet unique goals and objectives, usually to bring about beneficial change or added value. The primary challenge of project management is to achieve all of the project goals and objectives while honoring the preconceived project constraints. The nature of virtual enterprises increases the difficulties as managing the projects. Therefore, modern information and communication technologies are critically important for integrating the project processing among distributed services.

III. CASE-BASED REASONING

Case-based reasoning (CBR) is a problem solving paradigm that in many respects is fundamentally different from other major AI approaches. Instead of relying solely on general knowledge of a problem domain, or making associations along generalized relationships between problem descriptors and conclusions, CBR is able to utilize the specific knowledge of previously experienced, concrete problem situations. In CBR terminology, a case usually denotes a problem situation. A previously experienced situation, which has been captured and learned in a way that it can be reused in the solving of future problems, is referred to as a past case, previous case, stored case, or retained case. Correspondingly, a new case or unsolved case is the description of a new problem to be solved.

Case representation in case-based reasoning makes use of familiar knowledge representation formalisms from Artificial Intelligence (AI) to represent the experience contained in the cases for reasoning purposes. A large variety of representation formalisms have been proposed. However, three major types of case representation have arisen:

1. Feature vector fir propositional cases,
2. Structured formation for relational cases.

3. Textual representation for semi-structured cases.
4. Other specialized representations of cases for specific tasks.

Case-based reasoning is a cyclic and integrated process of solving a problem, learning from this experience, solving a new problem, etc. Figure 1 shows Aamodt & Plaza's (1994) classic model of the problem solving cycle in CBR. The individual tasks in the CBR cycle (i.e., retrieve, reuse, revise and retain) have come to be known as the '4 REs'. The cyclical process is typically described as follows [1]:

● RETRIEVE the most similar case(s).
● REUSE the case(s) to attempt to solve the problem.
● REVISE the proposed solution if necessary.
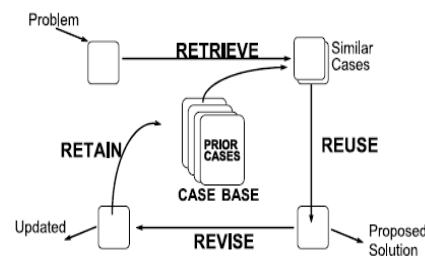● RETAIN the new solution as a part of a new case.

Figure 1. The CBR cycle

A new problem is matched against cases in the case base and one or more similar cases are retrieved. A solution suggested by the matching cases is then reused and tested for success. Unless the retrieved case is a close match the solution will probably have to be revised producing a new case that can be retained.
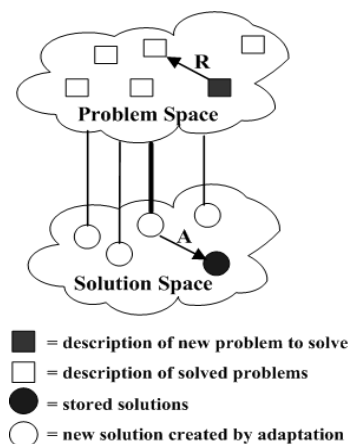

■ = description of new problem to solve
□ = description of solved problems
● = stored solutions
○ = new solution created by adaptation

Figure 2. Relationship between problem and solution spaces in CBR.

As illustrated in Figure 2, Leake (1996) expresses the role of similarity through the concepts of

retrieval and adaptation distances. Also captured in Leake's diagram is the relationship between problem and solution spaces in CBR. In Figure 2, the retrieval distance R increases as the similarity between the input problem description and a stored problem description decreases, lower similarity means greater distance. A common assumption in CBR is that the retrieval distance **R** is commensurate with **A**, the adaptation distance or effort [13] The concept is helpful for constructing the project experiences for supporting further users' requirements in the research.

IV. SEMANTIC WEB AND WEB SERVICES

A web service is a software program identified by an URI, which can be accessed via the internet through its exposed interface. Also, web services can be defined as software objects that can be assembled over the Internet using standard protocols to perform functions or execute business processes. The barriers to providing new offerings and entering new markets will be lowered to enable access for small and medium-sized enterprises.
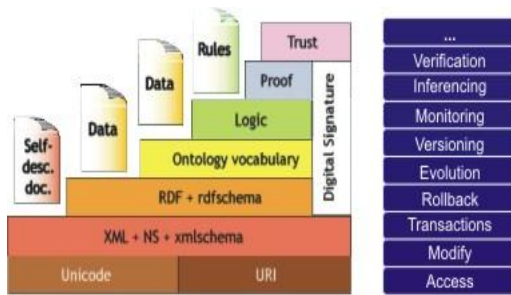
Figure 3. Static and dynamic aspects of the Semantic Web

The Semantic Web layer cake shown in figure 3 was presented by Tim Berners-Lee at XML2000 Conference [2]. Both the static and dynamic parts of the Semantic Web layer are discussed as follows. On the static side, Unicode, the URI and namespaces (NS) syntax and XML are used as a basis. The Resource Description Framework (RDF) may be used to make simple assertions about web resources or any other entity that can be named. RDF Schema extends RDF with the concepts of class and property hierarchies that enable the creation of simple ontologies. The Ontology layer features OWL (Ontology Web Language) which is a family of richer ontology languages that intend to replace RDF Schema. The Logic, Proof and Trust layers aren't standardized yet.

Syntax is the structure of data and semantics is the meaning of data. Two conditions necessary for interoperability: one is to adopt a common syntax and the advantage enables applications to parse the data, and another is to adopt a means for understanding the semantics. The advantage of

semantics is to enable applications to use the data DAML (DARPA Agent Markup Language) was created as part of a research program started in August 2000 by DARPA, a US governmental research organization. It is being developed by a large team of researchers, coordinated by DARPA. DAML and OIL merging in 2001 becomes OWL W3C standard in March 2003 [13].

RDF is a data model the model is domain-neutral, application-neutral and ready for internationalization. RDF model can be viewed as directed, labeled graphs or as an object-oriented model (object/attribute/value). RDF Schema (RDFs) defines small vocabulary for RDF, including Class, subClassOf, type, Property, subPropertyOf, domain and range. Vocabulary can be used to define other vocabularies for your application domain. OIL extends RDF Schema to a fully-fledged knowledge representation language [9].

The dynamic aspects apply to data across all layers. It is obvious that there have to be means for access and modification of Semantic Web data. Like in all distributed environments, monitoring of data operations is needed, in particular for confidential data. Finally, reasoning engines are to be applied for the deduction of implicit information as well as for semantic validation.
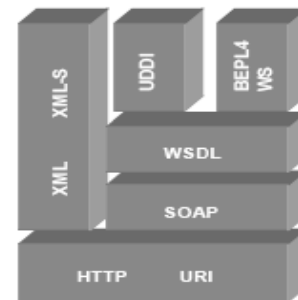

Figure 4. Web Services enabling standards

As shown in figure 4, web is organized around URIs, HTML, and HTTP. Web services require a similar infrastructure around Universal Description, Discovery, and Integration (UDDI), Web Service Definition Language (WSDL), and Simple Object Access Protocol (SOAP).

UDDI provides a mechanism for clients to find web services. Using a UDDI interface, businesses can dynamically look up as well as discover services provided by external business partners.

SOAP is a message layout specification that defines a uniform way of passing XML encoded data. It also defines a way to bind to HTTP as the underlying communication protocol for passing SOAP messages between two endpoints.

WSDL defines services as collections of network endpoints or ports. In WSDL the abstract definition of endpoints and messages is separated from their concrete network deployment or data format bindings.

Semantic web provides intelligent access to heterogeneous, distributed information, enabling software products to mediate between user needs and the information sources available [2][8]. Web services deal with the orthogonal limitation of the current web. Semantics web enabled web services have the potential to change our life to a much higher degree than the current web already has done, identifies the following elements necessary to enable efficient inter-enterprise execution: public process description and advertisement; discovery of services; selection of services; composition of services; and delivery, monitoring and contract negotiation[3]. Accordingly, the advantage of semantic web provides the conveniences to develop customized applications for virtual enterprises based on the infrastructure which are benefited by web services.

V. THE SEMANTIC-CBR FRAMEWORK

We propose the Semantic-CBR framework for improving the semantic communication and understanding for virtual enterprises as processing projects. There are two main roles: *CBR Project Engine* that is responsible for the manipulation of Semantic Project Base and *Semantic Annotation Builder* that described the valued previous projects in the consistent semantic representation. Figure 5 is the schematic diagram for the proposed framework.
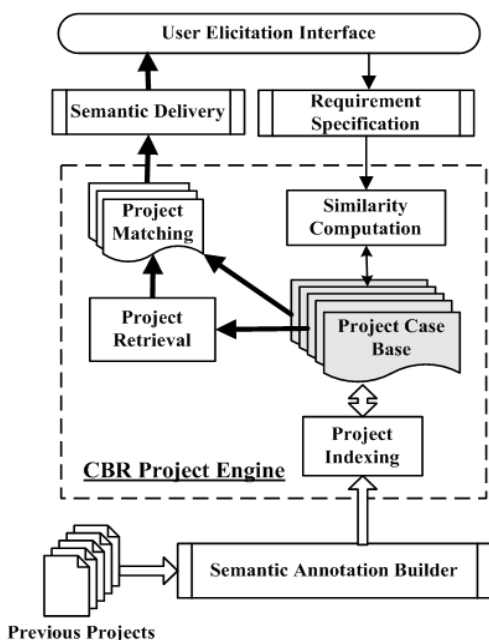

Figure 5. The Semantic-CBR framework

A. The operation of the framework

Semantic-CBR framework allows the requestor to provide project description and search for the

useful references in Semantic Project Base. The basic operation of the framework is explained as follows:

1. First, the administrator settles the Project Case Base with semantic representation formats for previous projects. The representation is used to annotate the execution experiences of web services for project solutions. Moreover, the Project Indexing module in CBR Project Engine systematically organized project descriptions and solutions with proper indexes for further efficient search.

2. Second, the requestor can input the project requirements and service references details. Then, the Requirement Specification module summaries the key requirement for Similarity Computation module to compute the similarity with previous situations in Project Case Base. Similarity Computation module can store different similarity algorithms for comparing the relations among the requirement and the previous project descriptions. Currently, the similarity assessment is performed using the nearest-neighbour technique. It processes retrieval of cases by comparing a list of weighted attributes in the target case to source cases in the CBR library [17]. The similarity measurement is shown as follows.

$$Similarity(T,S) = \sum_{j=1}^{n} f(T_j, S_i) \times Wi_i \quad (1)$$

Where
  T is the target case description
  S is the source case description
  n is the number of attributes in each case
  i is an individual attribute from **1** to **n**
  f is a similarity function for attribute **i** in cases **T** and **S**
  w is the importance weighting of attribute **i**

3. Third, the Project Retrieval module will select the suitable project cases based on the result of Similarity Computation module.

4. Forth, Project Matching module will determine the proper solutions from the previous result according to user requirement. Several operations, such as sorting, ranking or classifying, for the result of Project Retrieval module are such valued for efficient references.

5. Last, Semantic Delivery module will describe the project solutions for the requestor. The advantage of semantic representation is helpful for understanding and accessing the previous project solutions.

6. After the project is finished, the valued project information will be recorded in

Project Case Base with consistent semantic description by Semantic Annotation Builder module. To maintain the semantic Project Case Base is the most important mission of Semantic Annotation Builder module.

Accordingly, the performance result of the framework can integrate the project experiences and help users to reuse the related part of experiences to improve the further projects. The semantic support is rather emphasized since the dynamic environment in virtual enterprise increases the difficulties of resource identification, communication, understand and sharing.

The proper structure of semantic Project Case Base can store the complete project knowledge, as well as improve the understanding among project participants in virtual enterprises. The participants including the people, systems, machines or other devices are distributed in different places.

B. The structure of a project case

The task of project management becomes rather difficult in virtual enterprises since a project is developed for solving a certain problem across different organizations. That is, different definitions, locations, data structures, and even heterogeneous cultures will confuse the participants as processing a project. Therefore, the proposed framework emphasizes the semantic representation for integrating project resources in virtual enterprises.

The major structure of Project Case Base is designed for collecting the project problems and solutions. The entire description of project problem and solution is the essential component in a case. Therefore, we adopt frame structures for case representation since frame is highly structured and modular which allows handling complexity involved in case.

A frame system is a hierarchy of frames and each frame has a name and slots. Slots have dimensions that represent lower level elements of the frame, while fillers are the value range the slot dimensions can draw from. Moreover, the natural transition between semantic web descriptions and frame structure is very important for the semantic support in virtual enterprises.

In order to improving the consistent and interoperable understanding among virtual enterprises, we refer to the project ontology designed publicly for defining the frame structure in the research [20]. Basically, *Goal, Module, Task, Project, Agent, Session* and *Environment* are proposed for slots in a case frame. Many properties are provided for the usage of dimensions, including *name, logo, version, hasGoal* are the basic four dimensions for all slots. More useful defined properties can be applied for the dimensions, including *goalType, taskType, priority, status, duration, submittedDate, startDate, targetDate, finishedDate, dependsOn, isDependOf, isHelpedBy,*

helps, has Agent, owner, report, agentType, role, hasReport and branchTag which are used for describing different slots. The dimensions are useful annotations for slots to describe more project details. An example case frame is shown in Table 1.

Table 1. The Illustration of the Case Frame structure

Slot	Dimension	Filler
Goal	basic 4 *	valid values
	goalType	valid type instance
	priority	positive integer
	status	valid status instance
	duration	length of time instance
Module	basic 4 *	valid values
	priority	positive integer
	status	valid status instance
	duration	length of time instance
Task	basic 4 *	valid values
	tasktype	valid type instance
	priority	positive integer
	status	valid status instance
	duration	length of time instance
Project	basic 4 *	valid values
	priority	positive integer
	status	valid status instance
	duration	length of time
	hasAgent	valid values
	startDate	valid date
	finishDate	valid date
Agent	basic 4 *	valid values
	Owner	literal
	dependsOn	valid values
	agentType	valid type instance
Session	basic 4 *	valid values
	priority	positive integer
	duration	length of time instance
Environment	basic 4 *	valid values
	status	valid status instance
	isDependOf	valid values
	duration	length of time

Basic 4* means *name, logo, version* and *hasGoal* four dimensions.

C. The semantic expression in OWL

OWL is the latest standard in ontology languages (W3C Recommendation February 2004). It is layered on top of RDF and RDFS, and has a rich set of constructs. There are three categories of OWL: OWL-Lite, OWL-DL and OWLFull [24]. The rich expression of OWLFull is useful for describing the complex relationship of project management in the research.

Not only the defined properties are useful, but al so many properties provided by OWL are meaningful for describing the meaningful relationships in the framework. For example, a relation between a project and an agent can be described as follow.

```
<owl:ObjectProperty rdf:about="#hasAgent">
  <rdfs:domain rdf:resource= "#ProjectAA"   />
  <rdfs:range rdf:resource= "#AgentCC"   />
</ owl:ObjectProperty >
```

Next, iff the development of a framework is one of goals in the project, the OWL expression can be described as follows.

```
<owl:Class rdf:ID="Framework">
  <rdf:subClassOf rdf:resource= "#Goal"   />
</ owl:Class
```

Some restriction expressions in OWL are also useful for illustrating the real constraints in the project. For example, the projectKK is exclusively prepared for developing the framework. The syntax of *hasValue* is herein useful as follows.

```
<owl:Class rdf:ID="Framework">
  <rdfs:subClassOf rdfs:resource="#Goal" />
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty df:resource="#dependsOn"/>
      <owl:hasValue  rdf:resource="#ProjectKK"/>
    </owl:Restriction>
  </rdf:subClassOf>
</owl:Class>
```

If some of the programming tasks are helped by AgentK, then the syntax of *someValuesFrom* is useful for the situation.

```
<owl:Class rdf:ID="Programming">
  <rdfs:subClassOf rdfs:resource="#Task" />
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty
          rdf:resource="#isHelpedBy"/>
      <owl:someValuesFrom
              rdf:resource="#AgentK"/>
    </owl:Restriction>
  </rdf:subClassOf>
</owl:Class>
```

Otherwise, if the task of programming is exclusively assigned to AgentK, then the syntax of *allValuesFrom* is suitable for the situation.

Moreover, the usage of owl:*intersectionOf*, owl:*unionOf*, owl:*complementOf* and owl:*oneOf* are properly indicated the relations of intersection, union, complement, and Enumeration. Next, the cardinality constraints in OWL, containing owl:*maxCardinality*, owl:*minCardinality*, and owl:*cardinality* are useful for indicated for a specific number of values for that property, or to insist that a property must not occur.

In addition, OWL property characteristics, containing basic four rules, are useful for reasoning in the research.

1. Functional:- For a given individual, the property takes only one value.
2. Inverse functional:- The inverse of the property is functional.
3. Symmetric:- If a property links A to B then it can be inferred that it links B to A.
4. Transitive:- If a property links A to B and B to C then it can be inferred that it links A to C.

The OWL semantic expression capability is indeed helpful for integrating the heterogeneous project resources for virtual enterprises in the framework.

VI. CONCLUSIONS AND FUTURE WORK

We have illustrate the facility of Semantic-CBR framework of using semantic web description and integrating CBR cycle to allow project experiences sharing and reuse to occur in virtual enterprises. The fundamental idea in the research is to depict the importance of semantic support for virtual enterprises in project management. The proposed framework can help avoid past errors and learn from the errors and success. Also, the system keeps a record of each situation that occurred for future reference. Then, OWL, developed as a vocabulary extension of RDF, is effective for delivering complex project experiences. The semantic and conceptual support is accordingly fulfilled by OWL and RDF in the research.

More implementation and system simulation are the essential work. We will also involve investigating the case adaptation, which is necessary when the available cases can not meet the project problem requirements. Also, the reasoning methods required improving. The fundamental idea is to depict the importance of semantic support for virtual enterprises in project management.

REFERENCES

[1] A. Aamodt and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," *AI Communications*, vol.7, pp. 39-59, 1994.

[2] T. Berners-Lee, J. Handler and O. Lassila, "The Semantic Web," *Scientific American*, May 2001.

[3] C. Bussler, "The Role of Workflow Management Systems in B2B Integration," In *Proceedings of the Fourth International Conference on Electronic Commerce Research* (ICECR-4), Dallas, TX, USA, November 2001.

[4] J. M Burn,. and M. L. Barnett, "Communicating for Advantage in the Virtual Organization," *IEEE Transactions on Professional Communication*, vol.42, no.4, pp.1-8, 1999.

[5] J. Byrne, "The Virtual Corporation," *Business Week,* Feb 8, pp. 98-104,1993.

[6] L. Cabral, J. Domingue, E. Motta, T. Payne, and F. Hakimpour, "Approaches to Semantic Web Services: An Overview and Comparisons," 1st European Semantic Web Symposium, Heraklion, Greece, 2004.

[7] D. Fensel, "The Semantic Web and Its Languages," *IEEE Intelligent Systems*, vol. 15, no. 6, pp. 67–73, 2000.

[8] D. Fensel , J. Hendler, H. Lieberman, and W. Wahlster, (eds.), "Semantic Web Technology," *MIT Press*, Boston, 2002.

[9] F. Gandon and N. Sadeh, "Semantic Web Technologies to Reconcile Privacy and Context Awareness," *Web Semantics Journa*l, Vol.1, No.3, 2004.

[10] F. Gandon and N. Sadeh, "Semantic Web Technologies to Reconcile Privacy and Context Awareness," *Web Semantics Journal*, vol. 1,no. 3, 2004.

[11] H. J. ter Horst, "Completeness, decidability and complexity of entailment for RDF Schema and a semantic extension involving the OWL vocabulary," *Journal of Web Semantics,* vol. 3, pp. 79–115, 2005.

[12] J. Kolodner, "Case-Based Reasoning," Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[13] O. Lassila, and R. Swick, "Resource Description Framework (RDF) Model and Syntax Specification," W3C Recommendation, World Wide Web Consortium, Feb. 1999.

[14] DB. Leake, "CBR in Context: The Present and Furture," *Case-Based Reasoning Experiences, Lessons, & Future Directions*, pp. 3-30, AAAI Press, 1996.

[15] M. Lenz, and H.D. Burkhard, "Case retrieval nets: Basic ideas and extensions," in *KI-96*: *Advances in Artificial Intelligence*, Berlin: Springer, pp. 227–239,1996.

[16] A. Mowshowitz, "Virtual Organization," *Communications of the ACM*, vol.40, No.9, 1997.

[17] Project Management Institute, *A Guide to the Project Management Body of Knowledge*, fourth Edition (PMBOK Guides), 2009.

[18] E. Sirin, J. Hendler, and B. Parsia, "Semi-automatic composition of Web services using semantic descriptions," In *Proceedings of Web Services: Modeling, Architecture and Infrastructure*, 2002.

[19] I. Watson, *Applying Case-Based Reasoning: Techniques for Enterprise Systems*, Morgan Kaufmann Publishers, Inc., San Francisc,1997.

[20] SchemaWeb, **http://www.schemaweb.info/default.aspx**

[21] Semantic Web layer cake: **http://www.w3.org/2004/Talks/0412-RDF-functions/slide4-0.html**.

[22] W3C, Semantic Web, URL: **http://www.w3.org/2001/sw/**,2002

[23] W3C, RDF, **http://www.w3.org/RDF/,2004**

[24] OWL Web Ontology Language Reference, **http://www.w3.org/TR/owl-ref/**

Chouyin Hsu received the M.S. degree in computer science in 1993 from Ohio University, USA and the Ph.D degree in information management in 2005 from National Chiao Tung University, Hsinchu, Taiwan. She is currently an Associate Professor in the Department of Information Management at Overseas Chinese University, Taichung, Taiwan. Her current research interests include knowledge management, data mining, project management and enterprise resource planning.

A Diffusion of Innovations Approach to Investigate the RFID Adoption in Taiwan Logistics Industry

Yu-Bing Wang
University of Northern Virginia, Business School, Annandale, VA 22003, USA
Email: icebbb@gmail.com

Kuan-Yu Lin
Spalding University, Louisville, KY 40222, USA
Email: kylin086@gmail.com

Lily Chang [1]
Department of Information Management, Overseas Chinese University
Taichung 407, Taiwan
Email: lily@ocu.edu.tw

Jason C. Hung
Department of Information Management, Overseas Chinese University
Taichung 407, Taiwan
Email: jhung@ocu.edu.tw

Logistics tasks heavily depend on reliable shipment and accurate tracking information. For this reason, logistics today have evolved into a high-technology industry. Distribution is no longer simply about moving cargo on the road or via air from location A to B, but is a complex process based on intelligent system for sorting, planning, routing, and consolidation that supports faster transportation. [1] The purpose of this study was to investigate how RFID technology was implemented and adopted in Taiwan's logistics industry. Specifically, this study focused on the positive influence of using RFID technology on the industry and the strategical benefits the RFID system had provided to companies, which had accepted and utilized this technology. This study also aimed to determine the concern factors of adopting the RFID system into current company management systems. An integral part of this research was to develop and to empirically test a model of the adoption of RFID in the context of the logistics industry in Taiwan. Based on the concepts of Rogers [2]-- the theory of technology diffusion, this research used a questionnaire to assess Taiwan logistics companies' cognition and perspective of the relative advantage, compatibility, complexity, trialability and observability of the RFID system; as well as to assess their attitudes toward the RFID system and intentions of using the system. Research findings revealed the attributes of innovations mentioned above were significantly positively associated with the adoption of RFID. According to the research results, managerial implications and opportunities for future research were discussed in the final.

1. Corresponding author

Index Terms—**Radio Frequency Identification; Diffusion of Innovations Theory; Attributes of Innovations; Logistics Industry.**

I. INTRODUCTION

Since June 2003, mass consuming markets had demonstrated a significant shift toward Radio Frequency Identification (RFID) technology. This has occurred not only because of RFID mandates imposed by Wal-Mart and other stores, but also the widely used of RFID by government organizations. Its use has the potential to affect an extremely wide spectrum of the population, from technology adopters to vendors, integrators, and users [3]. For this reason, the adoption of RFID technology is expected to increase rapidly in the current and coming years. In 2006, Chu et al. [4] had even speculated that RFID production will increase 25-fold by 2010. One RFID consumer survey, the RFID Consumer Buzz - Special Report, found that 28 percent of the 7,000 U.S. consumers surveyed were aware of RFID, and that most of them could describe the technology to others [5]. Journalists and researchers [6] have found that a growing number of businesses use RFID in their supply chain or manufacturing processes. In addition, there are hundreds of studies that discuss how Wal-Mart uses RFID to reduce its huge logistics costs and improve supply chain efficiency.

In Taiwan, government departments, the medical and pharmaceutical sectors, and private businesses have followed the RFID trend to take advantage of this new

technology to enhance their standard operation processes [7] These companies are attracted by the number of potential benefits RFID offers, such as improved supply chain visibility, reduced labor costs, and enhanced process efficiency. This research study investigated the RFID technology adoption and how RFID has been used in Taiwan's logistics industry.

II. THEORETCAL FOUNDATION

A. Innovation Diffusion Theory (IDT)

According to Rogers [2], "diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social system". (p.5) In Rogers's definition there are four elements of diffusion that constitute the main idea of diffusion in the innovation process: the innovation, communication channels, time, and the social system. In this study of RFID system adoption, RFID system as the innovation is communicated through certain channels over time among the members. In this study, the social system means Taiwan business environment and the members refer to Taiwan logistics industry.

Regarding to innovation-decision process mentioned above, Rogers explained it is "the process through which an individual (or other decision-making unit) passes from first knowledge of an innovation to forming an attitude toward the innovation, to a decision to adopt or reject, to implementation and use of the new idea, and to confirmation of this decision". ([2], p.20) There are five main steps in the innovation-decision process: (1) Knowledge, (2) persuasion, (3) decision, (4) implementation, and (5) confirmation.

In addition, Rogers mentioned in his book "innovations that are perceived by individuals as having greater relative advantage, compatibility, trialability, and observability and less complexity will be adopted more rapidly than other innovations". (p.16) Innovations with the cluster of opposite characteristics require a longer diffusion period. There are many studies that indicate that these five qualities are the most important characteristics of innovation (attributes of innovations) in explaining the rate of adoption.

Relative advantages—"the degree to which an innovation is perceived as being better than the idea it supersedes. ". ([2], p.212) He explained the degree of relative advantage is often expressed as economic profitability, social prestige, or other benefits. The greater the perceived relative advantage of an innovation, the more rapid its rate of adoption will be.

Compatibility—The innovation's consistency with the existing values and norms (social system), past experiences, and the needs of potential adopters. "An idea that is more compatible is less uncertain to the potential adopter, and fits more closely with the individual's life situation". ([2], p.224)

Complexity—The degree to which an innovation is perceived as a difficulty to understand and use.

Trialability—The degree to which an innovation may be experimented on a limited basis. "New ideas that can be tried on the installment plan are generally adopted more rapidly than innovations that are not divisible". ([2], p.243) The higher the trialibility, the more quickly the innovation may be adopted.

Observability—The degree to which the results of an innovation are visible to others. Observability has been discovered to be positively related to adoption of an innovation.

Base on the concepts of Rogers's five steps of innovation-decision process and attributes of innovations, this research used a questionnaire to understand the perceived attributes of innovations and cognitions, decision among Taiwan logistics companies.

B. Radio Frequency Identification (RFID)

Radio Frequency Identification (RFID) technology is defined as a "wireless data collection technology that uses electronic tags for storing data and recognizing data" [8], and then uses radio waves to automatically identify any objects that have RFID tags. Sandip Lahiri, an RFID Solution Architect with IBM Global Services, indicated in his *RFID Sourcebook* [3] that "RFID is an example of automatic identification (Auto-ID) technology by which a physical object can be identified automatically." (p. 1)

The use of Radio Frequency Identification (RFID) in tracking and accessing applications first appeared during the 1980s. As the technology has been refined, more pervasive and possibly invasive uses for RFID systems have quickly gained the retail industry's attention because of RFID's ability to track moving objects.

An RFID system is a system that integrates a collection of components that implement an RFID solution. An RFID system may consist of several technology components — such as a tag, an antenna, a reader, and middleware — that can be embedded into a business environment to improve and transform key transaction processes.

The benefits of adopting an RFID system can be vary. They relate to the company or industry's current performance. Kevan [9] mentioned in Frontline solutions magazine ,Michael Dominy, a senior analyst for the Yankee Group, stated that "if you have world-class, leading capabilities in your logistics functions, the amount of benefit you're going to get out of RFID will be very small and incremental," ([9] ¶ 2). This means that if the logistics system the company adopted is a labor-based system company like most of the small and medium enterprises (SMEs) in Taiwan and the company wants to update to an RFID-enabled system, the benefits from adoption will be dramatic. In addition, the technology investment costs will be huge.

The logistics industry can be a major beneficiary of the RFID system because RFID technology dramatically enhances supply chain management, inventory management, and labor cost reduction. Chris Murphy, Senior Executive Editor of Information Week Web, said that "Today, without RFID, we don't know what's in the back room and what's in the front of customers' hands" [10]. In addition, a white paper published by the New Times Company (2006) pointed out that retailers can

expect great inventory savings and labor cost reduction from the adoption of radio frequency identification (RFID) technology. Furthermore, numerous researchers ([11],[3]) and consulting firms (such as A.T. Kearney, Information Week) have reported benefits in several areas, including inventory management, human resources management, and stock and shelf management.

III. RESEARCH QUESTIONS AND HYPOTHESES

A. Research Questions

Based on the research background and purpose of the study, this investigation attempted to examine various factors of RFID implementation in the logistics industry in Taiwan. The study sought to answer the following research questions:

RQ1: What factors influence RFID technology acceptance attitude in Taiwan's logistics industry?

H1: The Relative Advantage of RFID system positively affects the attitude of RFID adoption in the logistics companies in Taiwan.

H2: The Compatibility of RFID system positively affects the attitude of RFID adoption in the logistics companies in Taiwan.

H3: The Complexity of RFID system negatively affects the attitude of RFID adoption in the logistics companies in Taiwan.

H4: The Trialability of RFID system positively affects the attitude of RFID adoption in the logistics companies in Taiwan.

H5: The Observability of RFID system positively affects the attitude of RFID adoption in the logistics companies in Taiwan.

RQ2: What factors affect RFID implementation intention in Taiwan's logistics industry?

H6: The attitude of using RFID system positively affects the usage intention of RFID adoption in the logistics companies in Taiwan.

RQ3: What are the barriers to RFID adoption in Taiwan's logistics industry?

H7: The logistics environmental factor has amplifies effects on the relationships between attitudes toward RFID adoption and future usage intention.
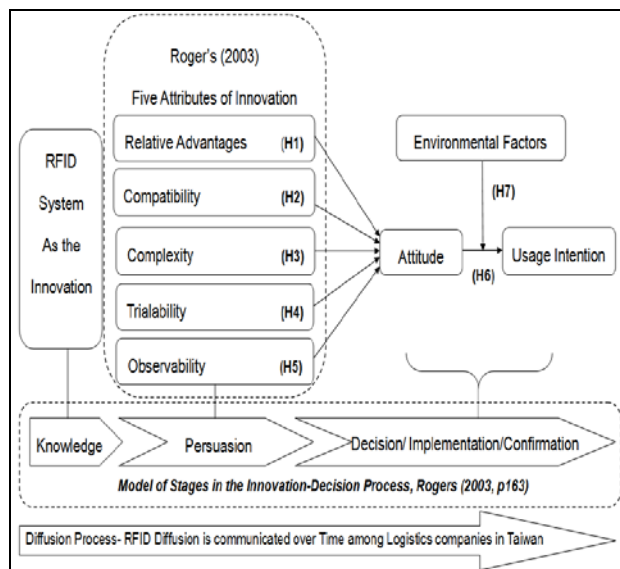
Figure1. Research Framework

IV. RESEARCH INSTRUMENTATION

A. Description of Questionnaire

The survey used in this study was primarily developed by the author to address unique aspects of this study with academic references, while the questionnaire as the data collection instrument for the context of the Taiwan logistics industry and its adoption of RFID technology.

The questionnaire was designed based on a five-point Likert-type scale which asks respondents to select from five scale options: 1 –Strongly disagree; 2 –Disagree; 3 – Undecided; 4 –Agree; 5 –Strongly Agree. It consisted of two sections. Part A aimed to investigate the basic background of the participating companies. Part B was designed to evaluate responding companies' knowledge and cognition of the relative advantage, compatibility, complexity, trialability, observability, and environmental effects regarding RFID technology.

The measurement variables of part B (Rogers's characteristics of innovation), relevant key questions and references are showed n below:

TABLE 1
MEASUREMENT SCALE QUESTIONS
FOR THE QUESTIONNAIRE PART B OF THIS RESEARCH

Variable	Relevant key questions	References
Relative Advantage	Q1: Efficiency Q2: Operation Performance Q3: Reduce Costs of Operations Management Q4: Profit Capability Q5: Effective	[12],[13], [14] ,[15], [2],[16][17] ,[18],[19], [20]
Compatibility	Q6: Value and belief Q7: Compatible Q8: Accord with the demand Q9: Accord with information system	

Complexity	Q10: Easy to use Q11: Easy to learn Q12: Clear and easy to understand Q13: Complicacy Q14: Difficulty	
Trialability	Q15: Popularization Q16: Assistance Q 17:Training programs Q18: Good product technique Q19: Proper post-sale service Q20: Training course	
Observability	Q21: Management support Q22: Management resource and budget Q23: Management sensed the efficiency Q24: Management encourage	
Environmental Factors	Q25: Competition Q26: Substitute products and services Q27: Barrier to entry Q28: Loss customers if not adopted RFID system Q29: More Competitive if adopt RFID system	
Attitude Toward RFID adoption	Q30: Positive toward RFID Q31: Positive view point of management's RFID adoption Q32: Advocate	[18],[21],[14]
Usage Intention	Q33: will and intention Q34: Rate of Usage Q35:Supply chain integration	[19],[20], [15];[14],[22]

Note. Questionnarie scale questions from this study

V. DATA COLLECTION AND ANALYSIS

The research population for this study was the companies list under the logistics-related category "104 Yellow Pages Business Directory" from Taiwan's largest job bank consultant company - 104 Job bank (http://www.104.com.tw) February, 2010, a total of 439 companies were the research population.

A. Validity and Reliability

The instrument of this research was developed with relevant references, key literature reviews and critical theory of Rogers [2] as the fundamental knowledge. Questions are abstracted from the most important points of the fundamental knowledge base. Thus, the questionnaire consisted of valid and crucial questions.

This study used Cronbach's alpha coefficient to examine the reliability of the instrument. There are eight variable factors in questionnaire part B. The reliabilities for those factors ranged from .702 to .923 as listed in Table 2, and the total reliability for part B questions were .931. According to George & Mallery [23, p231], they indicated that Cronbach's alpha reliability coefficient >0.9 means an excellent internal consistency of the questionnaire items. The closer Cronbach's alpha

coefficient is to 1.0 the greater the internal consistency of the items in the questionnaire scale. [28] The smallest value of Cronbach's alpha of this instrument was .702; this explained that this instrument was reliable. [24]

TABLE 2
RELIABLIITY OF CRONBACH ALPHA COEFFICIENT

Variables	Relevant key questions	Cronbach Alpha
Relative Advantage	Q1,Q2,Q3,Q4,Q5	.806
Compatibility	Q6,Q7,Q8,Q9	.812
Complexity	Q10,Q11,Q12,Q13,Q14	.855
Trialability	Q15,Q16,Q17,Q18,Q19,Q20	.779
Observability	Q21,Q22,Q23,Q24	.861
Environmental	Q25,Q26,Q27,Q28,Q29	.702
Attitude	Q30,Q31,Q32	.923
Usage Intention	Q33,Q34,Q35	.905
Total Alpha		.931

Note. [23] provided the following rules of thumb for Cronbach's alpha coefficient): "[alpha] > 0.9--Excellent, [alpha] > 0.8--Good, [alpha] > 0.7--Acceptable, [alpha] > 0.6--Questionable, [alpha] > 0.5--Poor, and [alpha] < 0.5--Unacceptable"

B. Descriptive analysis for participants' background

In this section, a total 439 participated companies' background analyzed by SPSS 16. June 3, 2010 to July 3, 2010 was the mailing survey operation time. After the end of one month survey period, 163 valid questionnaires were returned, yielding a 37.1 percent (37.1%) usable response rate. The background details have shown in the table below: (Table 3)

TABLE 3
PARTICIPATED COMPANY BACKGROUNDS

Background items	Items	Percent
Industry	Logistics Industry	68.10%
	Non-Logistics Industry	31.90%
Participants' Department	Management Dept.	42.41%
	Non-Management Dept.	57.59%
Business operated years	0-10 years	41.10%
	11-20 years	47.85%
	Over 21 years	11.05%
Company's Capitalization	under 51 million	63.19%
	51 million-100 million	20.86%
	Above 100-300 million	15.95%
The usage of Barcode system	Non-Barcode user	62.73%
	Barcode user	37.28%
The usage of RFID system	Non-RFID user	87.04%
	RFID user	12.96%

Note. There were total 163 valid respondent questionnaires.

Table 4-1 below showed the correlation of backgrounds and variables among participating companies (industry, participants' department, age of the firm, company capitalization) and all variables (five attributes of innovations, environmental factors, and attitude and usage intention) were conducted by Pearson correlation analysis. Table 4-1 demonstrated only participant's department has a significant positive correlation with the compatibility and complexity. The correlation coefficients are r =.165 (*) and .248 (**). The statistical result meant the higher working position participants attach the greater importance to the compatibility and complexity of RFID adoption.

TABLE 4-1
CORRELATION BETWEEN
COMPANY BACKGROUNDS AND VARABLES

Variables	Industry	Participant's Department	Business operated years	Company Capitalization
Relative Advantage	.023	.140	-.002	-.018
Compatibility	.024	.165 (*)	.015	.029
Complexity	.139	.248 (**)	-.042	-.051
Trialability	.119	-.004	-.109	-.017
Observability	.101	.065	-.082	-.037
Environmental	-.065	.098	.152	.022
Attitude	-.072	.078	.087	-.005
Usage Intention	.007	.098	-.016	.023

*Note. **. Correlation is significant at the 0.01 level (2-tailed).*

. Correlation is significant at the 0.05 level (2-tailed).

For the company background questions, the experience of using barcode system and RFID system was also examined. The relationship between those two system and variables were also conducted. In table 4-2, the Pearson's correlation represented that the usage of barcode system have significant (p<0.05) positive correlation with trialability and usage intention. Most importantly, the experiences of using RFID system were significantly related to all those variables.

TABLE 4-2
CORRELATION BETWEEN
BARCODE /RFID EXPERIENCES AND VARABLES

Variables	The usage of Barcode system	The usage of RFID system
Relative Advantage	.118	.265 (**)
Compatibility	.027	.305 (**)
Complexity	-.010	.259 (**)
Trialability	.181 (*)	.383 (**)
Observability	-.035	.258 (**)
Environmental	-.050	.175 (*)
Attitude	.151	.317 (**)
Usage Intention	.168 (*)	.449 (**)

*Note. **. Correlation is significant at the 0.01 level (2-tailed).*

. Correlation is significant at the 0.05 level (2-tailed).

C. Relationship between innovation variables/Attitudes; and relationship between innovation variables/usage intentions

As Table 5 illustrated below, the relationships among participating companies' cognition of attributes of innovations toward RFID adoption, environmental factors, attitude toward RFID adoption and future RFID usage intention were examined by Pearson Correlation analysis. All variables had significant (*p*<.01) correlation with each other (from .507 to .669), except the environmental factors. The high coefficients signified that these variables had solid relationships with each other. The strong correlations among all those variables were positive. The statistical analysis confirmed the hypothesis 1 to hypotheses 5 (H1-H5) were accepted. However, the hypothesis 7 has been denied.

TABLE 5
CORRELAITON COEFFICIENT-H1 to H5, H7

Variable	Attitude toward RFID adoption	Usage Intention
Relative Advantage	.581 (**)	.528 (**)
Compatibility	.569 (**)	.540 (**)
Complexity	.361 (**)	.459 (**)
Trialability	.507 (**)	.669 (**)
Observability	.570 (**)	.598(**)
Environmental	.060	.137

*Note. **. Correlation is significant at the 0.01 level (2-tailed).*

D. Relationship between the Attitude toward RFID System and Behavioral Intention

As Table 6 demonstrated below, the Pearson correlation analysis .623 revealed a strong and positive relationship existed between attitude toward RFID system and future usage intention. (*r*=.623**, *p*=.00*0*) The result supported past researches, which indicated a positive association between attitude toward RFID system adoption and behavioral intention. [25], [26], [27] Therefore, the hypothesis 6 was accepted.

TABLE 6
CORRELATION BETWEEN THE ATTITUDE AND USAGE
INTENTION

Correlations (H6)		Attitude toward RFID	Usage Intention
Attitude toward RFID	Pearson Correlation	1	**.623**[**]
	Sig. (2-tailed)		.000

Adoption	Sum of Squares and Cross-products	106.453	75.369
	Covariance	0.657	0.465
	N	163	163
Future usage Intention	Pearson Correlation	.623**	1
	Sig. (2-tailed)	.000	
	Sum of Squares and Cross-products	75.369	137.339
	Covariance	0.465	0.848
	N	163	163

*Note. **. Correlation is significant at the 0.01 level (2-tailed).*

Attitudes toward RFID adoption *Environmental Factor		1.854	.011	Rejected H7

VI. RESULTS AND CONCLUSION

The investigation of this paper served as a feasibility study to determine what factors have influenced RFID future usage intention. To achieve this purpose, the researcher adapted Rogers's [2] innovation diffusion concept and included the following five major variables in the research as a research model of this study: relative advantage, compatibility, complexity, trialability, observability and attitude toward RFID adoption.

The research framework of this study used five steps of Rogers's innovations – decision processes (knowledge, persuasion, decision, implementation, and confirmation), and followed by the main concept of innovation diffusion: an innovation is communicated through certain channels over time among the members of a social system to build up the RFID diffusion model. Using attributes of innovations, environmental factors and attitude toward RFID system as variables to examine potential influence factors toward the RFID future usage intention.

From data collection and analysis section, the researcher took advantage of SPSS 16.0 to analyze the conceptual framework, research questions, and hypotheses of this study. The quantitative research methods involved in this study were described as follow: descriptive statistics, Pearson's correlation, and one-way and two-way ANOVA. Base on the information of the valid questionnaires and the analyzed data, the findings and the conclusion of this study were discussed.

D. Tests of the Hypotheses

In this section, the researcher conducted one-way and two-way ANOVA to examine all seven hypotheses of this study. In one-way ANOVA, the five attributes of innovations toward RFID adoption were independent variables while attitude toward RFID adoption served as one independent variable. The statistical results indicated that the F-value of H1 to H5 were all reached the significant level ($p<.001$); therefore, hypotheses 1 to 5 were accepted. (See Table 7)

Regarding to H6, the purpose of one-way ANOVA was to verify whether there was a significant effect between attitudes toward RFID adoption and RFID usage intention. Attitude was the independent variable and dependent variable was usage intention. The ANOVA was significant for H6 ($F=18.842$, $p=.000$); therefore, hypothesis 6 was accepted. (See table 7)

A two-way analysis of variance was conducted to evaluate the relationship between two independent variables. In this case, they were the attitude toward RFID adoption and environmental factors; and the usage intention was the dependent variable. The statistical results indicated the F value was 1.854 and p-value was .011 made the results were not significant, therefore the hypothesis was rejected.

A: Research question 1: What factors influence RFID technology acceptance attitude in Taiwan's logistics industry?

For the research question one (RQ1), not only the Pearson product-moment correlation (Table 5) was chosen to find the relationship between variables and the attitude toward RFID system adoption, but also one-way ANOVA (Table 7) was applied. The results found significant relationships and differences among these independent factors, relative advantages, compatibility, complexity, trialability, and observability, to the dependent variables, the attitude toward RFID system adoption. The results of this study suggested the five attributes of RFID system were conducive to the attitude toward RFID adoption of Taiwan logistics companies. It can be the influential factors toward RFID acceptance attitudes.

TABLE 7
RESULT OF HYPOTHESIS TESTING

Fixed Factor	Dependant Variable	F-value	P-value (Sig.)	Hypothesis testing
Relative Advantage	Attitudes toward RFID adoption	8.862	.000	Accepted H1
Compatibility		8.109	.000	Accepted H2
Complexity		5.759	.000	Accepted H3
Trialability		10.327	.000	Accepted H4
Observability		6.882	.000	Accepted H5
Attitudes toward RFID adoption	Usage intention	18.842	.000	Accepted H6

B. Research question 2: What factors affect RFID implementation intention in Taiwan's logistics industry?

The relationship between the attitudes toward RFID adoption and RFID usage intention showed significant differences and positive correlated with each other. Base on the results of one-way ANOVA (Table 7) for RQ2, different attitude toward RFID adoption had a different usage intention. Pearson product-moment correlation (Table 6) proved that attitudes toward RFID adoption and RFID usage intention had positive relationship; therefore, attitudes toward RFID adoption affected RFID implementation intention in Taiwan's logistics industry.

C. Research question 3: What are the barriers to RFID adoption in Taiwan's logistics industry?

Base on the two-way ANOVA, the relationship between the attitudes toward RFID adoption and environmental factors showed no significant difference to the RFID usage intention which meant the environmental factors such as competition inside the logistics industry, substitute products and services, entry barrier were not very influential to RFID usage intention.

VII. SUGGESTIONS

The results of this study have demonstrated the value and the feasibility of RFID adoption for Taiwan's logistics industry. The diffusion of innovation model of Rogers in this study also revealed the importance of the attitude toward RFID adoption and usage intension in terms of potential users' perceptions of attributes of RFID system.

Base on the research questions and the findings from the data analysis, the five perceived attributes of RFID system have the most significant predictive power on potential users' intentions of adopting RFID system. According to the research results, the researcher offered the following contributions and suggestions to this study.

The most prominent finding of this study was the five attributes of the innovations - RFID system. RFID provider companies should try their best to promote this technology's attributes of innovations: the relative advantages of RFID system to their potential customers, such as the costs reduction, profit capability and improvement of operation performance, design RFID system more friendly and more compatibility for quick adoption; reduce RFID system's difficulty and make it easy to learn and use. RFID providers can provide a trial-able project toward their potential customers. (Korteweg, et al, 2006) Moreover, the pilot project can identify aspects of important before deciding to totally implement the technology, which improves the observability of the RFID system.

REFERENCES

[1] Jakobs, K., Pils, C., Wallbaum, M. (2001), Using the internet in transport logistics – the example of a track and trace system, *Lecture Notes in Computer Science. No. 2093,* 194-203

[2] Rogers, E. M. (2003). Diffusion of innovations (5th ed.). New York: Free Press.

[3] Lahiri, Sandip (2006) *RFID Sourcebook*. New Jersey: International Business Machines Press.

[4] Chu, Wai Lang (2006) RFID production to increase 25-fold by 2010, Labtechnologist.com. Retrieved November 25, 2008, from http://www.labtechnologist.com/Industry-Drivers/RFID-production-to-increase-25-fold-by-2010

[5] Roberti, M. (2005). EPC Reduces Out-of-Stocks at Wal-Mart. *RFID Journal.* Retrieved March 19, 2009, from http://www.rfidjournal.com/article/view/1927/1

[6] Sabbaghi, A., & Vaidyanathan, G. (2008). Effectiveness Supply Chain Management: Strategic values and Challenges. *Journal of Theoretical and Applied Electronic Commerce Research*, 3(2), 71-81.

[7] Chow, K., Choy, K. & Lee, W. (2006) A dynamic logistics process knowledge-based system – An RFID multi-agent approach; knowledge-Based Systems, Volume 20, Issue 4, May 2007, Pages 357-372

[8] Expert Barcode & RFID, Inc., 2009, ¶ 5. Retrieved July 20,2010 from http://www.expertbarcode.com/autotech.htm

[9] Kevan, T. (2004). "Calculating RFID's benefits: retailers can expect significant inventory and labor savings, but manufacturers face slim returns. However, the benefits aren't always measured in dollars". *Frontline Solutions.* Retrieved July 23, 2010, from http://findarticles.com/p/articles/mi_m0DIS/is_1_5/ai_112563027/

[10] Murphy, C. (2005). Real-World RFID: Wal-Mart, Gillette, And Others Share What They're Learning. *Information Week.* Retrieved August 08, 2009, from http://www.informationweek.com/news/mobility/RFID/showArticle.jhtml?articleID=163700955

[11] Palmer, M.: Principles of Effective RFID Data Management. Enterprise Systems (2004)

[12] Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice-Hall.

[13] Chau, P.Y. K., (1996) An empirical assessment of a modified technology acceptance model. *Journal of Management Information System, 13,* p.185-204

[14] Davis, F. D. (1989). "Perceived Usefulness, Perceived Ease-of-Use and User Acceptance of Information Technology," *MIS Quarterly, Vol. 13, No. 3*, 319-339.

[15] Premkumar G. & M., Roberts, (1999), Adoption of New Information Technologies in Rural Small Businesses. *The International Journal of Management Science*, 27(4): 467-484.

[16] Gildas A. and Philippe O. RFID traceability: A multilayer problem. In Andrew Patrick and Moti Yung, editors, *Financial Cryptography – FC'05*, volume 3570 of *Lecture Notes in Computer Science*, P125–140, Roseau, The

Commonwealth Of Dominica, February 2005. IFCA, Springer-Verlag.

[17] Amororso, D. L. (2004) , Development of an instrument to measure the acceptance ofinternet technology by consumers, Proceedings of the 37th Hawaii International Conference on System Science, *IEEE*

[18] Lee, J. , Discriminant analysis of technology adoption behavior: a case of internet technologies in small businesses, The Journal of Computer Information Systems, 2004, 44(4), p.57-66

[19] Ajzen, I., & Fishbein, M. (1980). Understanding attitudes and predicting social behavior. Englewood Cliffs, NJ: Prentice Hall.

[20] Ajzen, I. (2005). *Attitudes, personality, and behavior* (2nd. Edition).Milton-Keynes, England: Open University Press / McGraw-Hill.

[21] Premkumar, G., and Ramamurthy, K.,(1995) "The role of interorganizational and organizational factors on the decision mode for adoption of interorganizational systems," Decision Sciences, 26(3), May/Jun 1995, pp.303-336.

[22] Bagchi, P.K., Chun Ha, B., Skjoett-Larsen, T., Soerensen, L.B. (2005), "Supply chain integration: a European survey", *International Journal of Logistics Management*, Vol. 16 No.2, pp.275-94.

[23] George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.). Boston: Allyn & Bacon.

[24] Nunnally (1978), J.C. (1978) Psychometric theory (2ed ed.). New York: McGraw Hill.

[25] Agarwal, R., and Prasad, J. (1997). The Role of Innovation Characteristics and Perceived Voluntariness in the Acceptance of Information Technologies, *Decision Science, Vol. 28*, No. 3, pp. 557-582.

[26] Davis, F. D., & Venkatesh, V. (1996). A critical assessment of potential measurement biases in the technology acceptance model: three experiments. *International Journal of Human-Computer Studies,* 145, 19-45.

[27] Venkatesh, V., Morris, M.G., Davis, G.B., & Davis, F.D. (2003). User acceptance of information technology: toward a unified view. *MIS Quaterly,* 27 (3), 425-478.

[28] Korteweg, B. (2007). Crossing the Chasm: exploring RFID adoption characteristics in South African industries. University of Twente Student Theses
Gliem, J. A., & Gliem, R.R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Midwest Research to Practice Conference in Adult, Continuing, and Community Education, retrieved July 10, 2010 from https://scholarworks.iupui.edu/bitstream/handle/1805/344/Gliem+&+Gliem.pdf?sequence=1

Yu-Bing Wang earned her Master of Science in Electronic Business Management from the Department of Engineering, the University of Warwick, UK in 2002. From 2002 to 2003, she was a lecturer of Department of Information Management of Overseas Chinese University and Department of Marketing Logistics Management of Ling Tung University. She is now a candidate of Doctor of Business Administration at University of Northern Virginia in the USA. Her research interests include e-commerce, logistics, and business administration.

Kuan-Yu Lin is currently a doctoral candidate of Leadership Education at Spalding University in the USA. His research interests include English as a foreign language acquisition, Human Resources Management and Chinese-English translation.

Lily Chang, a lecturer of Dept. of Information Management, Overseas Chinese University, Taiwan. Her research interests include Microeconomics, Macroeconomics, and Management Science.

Jason C. Hung is an Associate Professor of Dept. of Information Management, Overseas Chinese University, Taiwan. His research interests include Multimedia Computing and Networking, Distance Learning, E-Commerce, and Agent Technology. From 1999 to date, he was a part time faculty of the Computer Science and Information Engineering Department at Tamkang University. Dr. Hung received his BS and MS degrees in Computer Science and Information Engineering from Tamkang University, in 1996 and 1998, respectively. He also received his Ph.D. in Computer Science and Information Engineering from Tamkang University in 2001.

Incremental Mining of Across-streams Sequential Patterns in Multiple Data Streams

Shih-Yang Yang

Department of Media Art, Kang-Ning Junior College of Medical Care and Management, Taipei, Taiwan 114, R.O.C.
Email : shihyang@knjc.edu.tw

Ching-Ming Chao

Department of Computer Science and Information Management, Soochow University, Taipei, Taiwan 100, R.O.C.
Email : chao@csim.scu.edu.tw

Po-Zung Chen and Chu-Hao Sun

Department of Computer Science and Information Engineering Tamkang University, Tamsui, Taiwan 25137, R.O.C.
Email : pozung@mail.tku.edu.tw, 894190320@s94.tku.edu.tw

Abstract—**Sequential pattern mining is the mining of data sequences for frequent sequential patterns with time sequence, which has a wide application. Data streams are streams of data that arrive at high speed. Due to the limitation of memory capacity and the need of real-time mining, the results of mining need to be updated in real time. Multiple data streams are the simultaneous arrival of a plurality of data streams, for which a much larger amount of data needs to be processed. Due to the inapplicability of traditional sequential pattern mining techniques, sequential pattern mining in multiple data streams has become an important research issue. Previous research can only handle a single item at a time and hence is incapable of coping with the changing environment of multiple data streams. In this paper, therefore, we propose the IAspam algorithm that not only can handle a set of items at a time but also can incrementally mine across-streams sequential patterns. In the process, stream data are converted into bitmap representation for mining. Experimental results show that the IAspam algorithm is effective in execution time when processing large amounts of stream data.**

Index Terms— **Multiple data streams, Data stream mining, Sequential pattern mining, Incremental mining**

I. INTRODUCTION

In the era of knowledge economy, the creation, communication and application of knowledge are the driving force behind the growth of industry. As a result, knowledge has become one of indispensable requirements to success. Therefore, how to acquire valuable knowledge is very important. Data mining is to discover useful knowledge from a great amount of data. Sequential pattern mining is to find out sequential patterns that frequently happen with time or specific sequence. For example, in terms of web pages access, the access patterns of web surfers can be explored to predict the next possibly accessed web page for advance access to expedite the web access.

There are many user behaviors in our daily life that can generate stream data such as the signal data of mobile communication, stock transaction data, web pages browsing records, etc. These stream data have the following characteristics [1]: (1) unlimited data input, (2) limited memory capacity, (3) one-time processing of input data, (4) fast data arrival, (5) inability of system to halt data inflow. As the characteristics of data streams are different from those of traditional databases, database mining algorithms are inapplicable in the data stream environment. Data streams can be classified into single data stream and multiple data streams. Sequential pattern mining in multiple data streams is an important research issue.

With the change of environment in multiple data streams, data will not only grow larger and larger but also get more and more complicated. Existing algorithms can only handle a single item at a time and hence are incapable of coping with the changing environment of multiple data streams. Besides, the sequential patterns mined may cross different data streams but are regarded as existing in a single data stream. To solve these problems, in this paper we propose the IAspam (Incremental Across-streams Sequential Pattern Mining) algorithm to cope with more complicated multiple data streams environment. IAspam not only can handle a set of items at a time but also can incrementally mine across-streams sequential patterns.

In the process of mining sequential patterns in multiple data streams, the efficiency in calculating the support and searching for frequent sequential patterns must be taken into account. To improve efficiency, we use a bitmap representation, in which each item is converted into one bit, to reduce execution time.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the ASPAMDAS method, which covers data sampling and the IAspam algorithm. Section 4 evaluates and compares the performance of the IAspam algorithm. Section 5 concludes this paper.

II. RELATED WORK

Due to the advent of data stream environment, traditional data mining algorithms are not applicable. Consequently, many researches were conducted on modifying algorithms to be applicable to data stream environment, e.g. mining frequent itemsets, sequential patterns, maximal sequential patterns, and closed sequential patterns in data stream environment. On the other hand, there are also many researches on the types and characteristics of stream data. The MSDD (Multi-Stream Dependency Detection) algorithm proposed by Oates and Cohen [2] aims to find out the dependency rules among stream data in a multiple data streams framework. Whereas, Golab and Ozsu [3] explore the characteristics, models, and query semantics of data stream as well as explain the application of data stream in real life. After the presentation of data stream framework, there are many researches on mining sequential patterns in data stream environment.

In data stream environment, incremental mining retains previous mining results. However, if not updated for a long time, previous mining results might become incorrect, for which, many studies put forward decay mechanisms for data stream mining to prevent data from becoming obsolete and save memory space. Chang and Lee [4] proposed a decay mechanism for data stream environment, using (1) to decay the mining results not updated for a long time to ensure correct mining results. The "d" in the equation stands for "decay rate", "b" for "decay-base", and "h" for "decay-base-life".

$$d = b^{-(1/h)} \ (b > 1, h \geqq 1, b \leqq d < 1) \qquad (1)$$

The MILE algorithm proposed by Chen et al. [5] is based on the prefix-projection concept of PrefixSpan. It is used for handling multiple data streams in a time period and mining sequential patterns in time sequence data stream environment. It identifies the prefix subsequence and suffix subsequence in stream data to reduce the generation of surplus candidate sequences and combine the sequential patterns with the same prefix. The main purpose is to avoid repeated data searching and speed up the mining of new sequential patterns. However, as MILE is a one-time fashioned algorithm that cannot retain previous mining results, it will take more time in re-mining. IncSPAM mentioned in the previous subsection is a sequential pattern mining algorithm for single data stream environment. It adopts a sliding window to sample stream data and incrementally mines sequential patterns. As sequential patterns will be outdated if not updated for a long time, it uses the weight to decay the support. This concept is based on the decay mechanism proposed in [6], which uses (1) to decay the support of sequential patterns.

Later the SPEED (Sequential Patterns Efficient Extraction in Data Streams) algorithm was proposed by Raissi et al. [7] for mining maximal sequential patterns in data stream. It uses a new data structure to maintain frequent sequential patterns and a fast pruning strategy to allow users to locate at any time the maximal sequential patterns in arbitrary time intervals. The SSM algorithm proposed by Ezeife and Monwar [8] differs from those mentioned above. It uses a D-list structure to effectively store

and maintain the support of all items in data stream environment. Meanwhile it continuously builds a PLWAP tree to effectively mine in batches sequential patterns in data stream, after which an FSP-tree is used to incrementally maintain the mining results. Also, there are some algorithms that use graphs for mining sequential patterns. The GraSeq algorithm proposed by Li and Chen [9] adopts a directed weighted graph structure to store the synopsis of sequences and only needs to scan the data stream once. It also makes use of subsequence matching to reduce the handling time of longer sequences and adopts the concept of valid node to improve the correctness of mining results.

The behavior patterns of mobile user in mobile communication environment are also a kind of sequential patterns. Lee et al. [10] proposed the T-Map-Mine algorithm for mining the behavior patterns of mobile user. It mines mobile sequential patterns from mobile user's signal data within each time interval at each location. A mobile sequential pattern includes a sequential pattern and a moving path. T-Map-Mine uses T-Map-Tree to store sequential patterns and adds the moving locations and customer data to T-Map-Tree one by one to increase the support of locations and sequential patterns. The mining results are the sequential patterns with frequent moving paths found in T-Map-Tree. In the mining process, T-Map-Mine must continuously search and insert nodes. Besides, the tail node of each branch of T-Map-Tree is connected to a RST (Requested Services Tree) to store sequential patterns. Therefore, in processing large amount of data, more time is needed for the construction and traversal of the tree.

III. THE ASPAMDAS METHOD

In multiple date streams environment, data will grow larger in size and become more complicated with the change of real-world environment. However, existing algorithms can only handle a single item at a time and are hence incapable of coping with the changing environment of multiple data streams. Based on commercial consideration, mobile operators may need to find out the association among streams produced at different locations to provide their customers with better services. In this paper, therefore, we propose the ASPAMDAS (Across-streams Sequential Pattern Mining in Multiple Data Streams) method to solve these problems. ASPAMDAS mainly consists of two stages: data sampling and incremental mining. In the stage of data sampling, we adopt the sliding window model to sample stream data. The size of the sampled data will be the window size. In the stage of incremental mining, we convert customer transaction data in the sliding window into a bitmap to find across-streams sequential patterns hereby to improve mining efficiency. After that, new and old across-streams sequential patterns are either updated or eliminated to produce precise mining results. Figure 1 shows the overall process of the ASPAMDAS method.

A. Data Sampling

In data stream environment, due to the increasing need of real-time knowledge and the limitation of data streams, the newer the data is, the more important it becomes. Accordingly, we use a sliding window to sample new stream data and dump the old ones that have been processed. As shown in Figure 2, the sliding window only maintains the newest N pieces of transaction data. N stands for the window size. After the transaction data within the window is processed, the window will move forward one time point to sample the newest stream data to process.
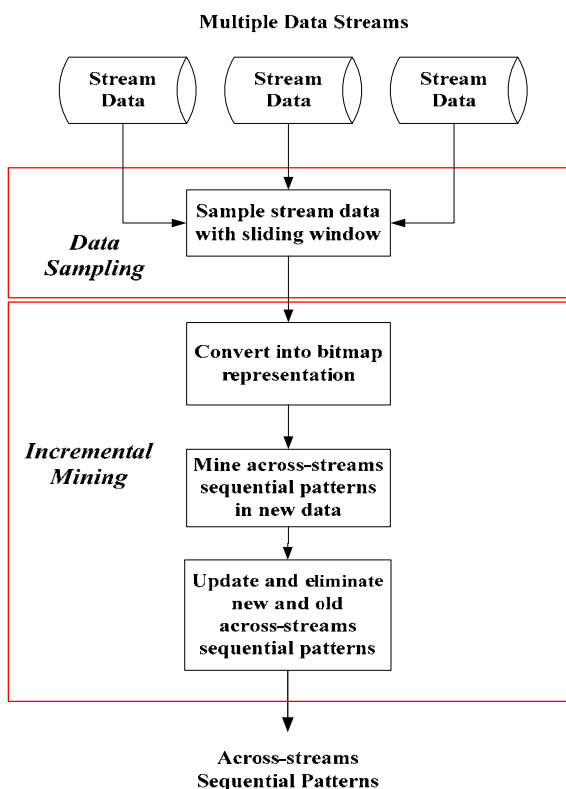
Figure 1. Overall process of ASPAMDAS
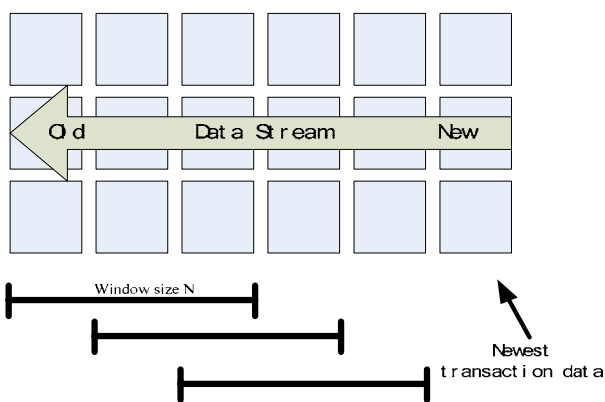
Figure 2. Sliding window in data stream environment

Figure 3 shows an example of multiple data streams. S1, S2, and S3 represent three data streams. Data from three data streams will be received at each time point. Wi stands for the ith window. Each data stream can be regarded as a location of customer transaction; whereas the stream data at each time point can be viewed as a piece of customer transaction data. These customer transaction data ordered by their arrival sequence become the transaction sequence data. The window size is 3, which means each time the sliding window samples transaction data of the latest 3 time points. For example, the customer transaction data sampled at the first time point are C1:(a,c), C3:(b,c,d), and C4:(b,c). The sliding window will move in the direction of time. After the transaction data in W1 are processed, the sliding window will move forward one time point to process the transaction data in W2. In this way, stream data will be mined incrementally.
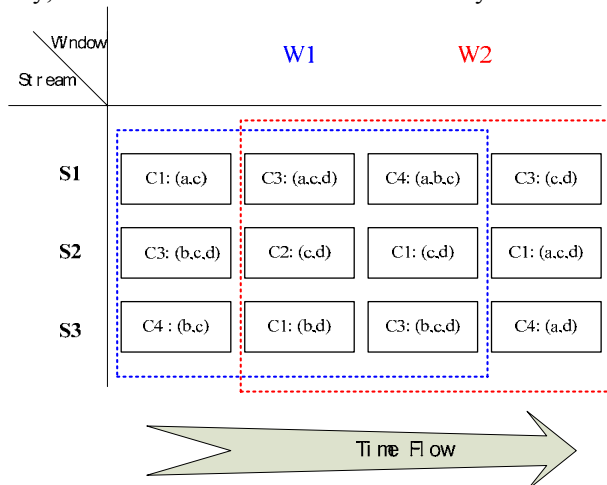
Figure 3. Example of multiple data streams

B. Incremental Mining

As described in previous subsection, a sliding window is used to sample stream data for mining. But due to the characteristics of data stream and the limitation of memory capacity, not all the mining results will be completely stored. Therefore, the concept of incremental mining is introduced to prevent the problem of imprecise mining results. As new stream data are produced continuously, the mining results will be continuously updated with the mining of new stream data. Below are possible situations of sequential pattern updating:

1. New data bring in new sequential patterns, which are frequent sequential patterns.
2. After updated, frequent sequential patterns remain frequent sequential patterns.
3. Frequent sequential patterns may become infrequent sequential patterns.
4. Infrequent sequential pattern may turn into frequent sequential patterns.

Situation 1: A frequent sequential pattern mined in current sliding window does not appear in previous mining results. This sequential pattern is a new frequent sequential pattern.

Situation 2: A frequent sequential pattern mined in current sliding window already exists in previous mining results. After updating its support, this sequential pattern remains a frequent sequential pattern.

Situation 3: A previously mined frequent sequential pattern does not appear in later sliding windows. When

its support decays to less than the minimum support, this sequential pattern will become an infrequent sequential pattern.

Situation 4: An infrequent sequential pattern may turn into a frequent sequential pattern in current sliding window if in previous mining results each of its subsequences is a frequent sequential pattern and does not have any supersequence.

Definition 1 (**support**): Let α be a sequential pattern. The support of α is the ratio of the number of transactions containing α to the number of all transactions.

Definition 2 (**frequent sequential pattern**): Let α be a sequential pattern. Then α is a frequent sequential pattern if and only if the support of α is greater than or equal to the minimum support.

Definition 3 (**frequent across-streams sequential pattern**): Let α be a sequential pattern, β be a stream sequence, and the sequence of streams containing the subsequences of α be a subsequence of β. Then α is a frequent across-streams sequential pattern if and only if β is a frequent stream sequence.

Figure 4 shows the IAspam algorithm whose parameter definitions are as follows:

- SW' stands for the data in current sliding window and SW stands for the data in previous sliding windows.
- min_sup stands for minimum support.
- S stands for stream number, I for itemset, CID for customer identification, and SD for stream data.
- k-item represents the item named k.
- CP stands for candidate sequence.
- CS stands for customer sequence table.
- $sup_{sw}(p)$ stands for the support of sequential pattern p in SW and $sup_{sw'}(p)$ stands for the support of sequential pattern p in SW.
- L-tree stands for lexicographical sequence tree.
- FASP represents frequent across-streams sequential pattern and StP represents stream sequence.
- $Sup_{SD}(p)$ represents the support of sequential pattern p in SD.
- Sup(FASP) represents the support of sequential pattern FASP in L-tree.

Algorithm: IAspam

Input: SW', SW, min_sup, CID, S, I, k-item, CS, L-tree
Output: FASP, StP

1. **For each** transaction data of SW' **do**
   store into CS as <S, I>;
2. **For each** transaction sequence of CS **do**
   convert into bit-vector and store into bitmap matrix;
3. **For each** 1-item **do**
   add 1-item and sup_{SD}(1-item) into L-tree;
   If sup_{SD}(1-item) $\geqq$ min_sup **then**
   generate CP with I-step and S-step and map to bitmap matrix;
4. **For each** CP **do**
   If $sup_{sw'}$(CP) $\geqq$ min_sup **then**
   If CP exists in L-tree **then** update sup_{sw}(CP);
   else add FASP and StP into L-tree;
5. **For each** FASP in L-tree **do**
   If sup(FASP) not updated **then**
   decay sup(FASP);
   If sup(FASP) < min_sup **then** delete FASP;
6. **return**;

Figure 4. IAspam algorithm

The IAspam algorithm consists of three parts. The first part is the conversion into bitmap representation (Step 1-2). Customer transaction data will be converted into a bitmap matrix. The second part is the mining of across-streams sequential patterns in new data (Step 3). I-step and S-step are conducted on the 1-item to generate candidate sequences, which will be mapped to the bitmap matrix. The third part is the integration of new and old sequential patterns (Step 4-5). The newly mined patterns will be compared with previous patterns to update sequential patterns and decay the support of the sequential patterns not updated. Below is the detailed explanation of the steps of the IAspam algorithm.

Step 1: Store customer transaction data in current sliding window into the customer sequence table, in which data are ordered by customer identification. Data format is <S, I>.

Step 2: Convert each transaction sequence of the customer sequence table into a bitmap vector of the customer bitmap matrix. Judge if the itemset at each time point contains k-item. If yes, store bit 1 into the column of that time point of k-item; otherwise, store bit 0.

Step 3: Count the support of each 1-item in each stream and store each 1-item and its supports into L-tree. Judge if every support of a 1-item is greater than or equal to the minimum support. If yes, conduct I-step and S-step to generate candidate sequences and map them to a bitmap matrix. In the mapping process, the candidate itemsets generated by I-step will be used to search for the '1' bits that are at the same time point and in the same stream but not in the same item, and count the support. Whereas, the candidate sequences generated by S-step will be used to search, in different items and at different time points, for the '1' bits whose prefix subsequences are at the same time point and in the same stream and the '1' bits whose suffix subsequences are at the same time point and in the same stream, and count the support.

Step 4: Judge if the support of each candidate sequence in SW' is greater than or equal to the minimum support. If yes, check to see if the frequent sequential pattern exists in the lexicographical tree. If yes, update the support of this sequential pattern. Otherwise, insert a frequent across-streams sequential pattern (FASP) and a stream sequence (StP) into the lexicographical tree.

Step 5: Judge if the support of each FASP in the lexicographical tree has been updated. If not, decay the support of FASP and judge if the decayed support is less than the minimum support. If yes, delete FASP to get rid of the FASP that has not been updated for a long time.

Step 6: Output FASP and StP.

The steps of sequence extension are to conduct I-step before S-step. I-step is to insert a new item into an itemset. For example, after going through I-step, <(a,b)> and <(c)> will become <(a,b,c)>, which can be expressed as <(a,b)> $\diamondsuit_I$ <(c)> = <(a,b,c)>. S-step is to append a new itemset to an itemset. For instance, after undergoing S-step, <(a,b)> and <(c)> will turn into <(a,b)(c)>, which can be expressed as <(a,b)> $\diamondsuit_S$ <(c)> = <(a,b)(c)>. According to the Apriori property, the supersequences of an infrequent sequence cannot be frequent. Consequently, it

is not necessary to conduct I-step and S-step on infrequent sequences and hence no surplus candidate sequences will be generated.

Each leaf node of a lexicographical tree records a frequent across-streams sequential pattern plus its support and stream sequence. The purpose of constructing a lexicographical tree is to store frequent across-streams sequential patterns in order to insert new across-streams sequential patterns or to combine or update new and old across-streams sequential patterns. The nodes of a lexicographical tree have a lexicographical order. The order of constructing a lexicographical tree is from top to bottom and from left to right. According to the characteristics of lexicographical order, the possible lexicographical orders of sequences are as below: [8]:

- If s' = s $\diamondsuit$ p, then s < s'; e.g. <(a,b)> < <(a,b)(b)>.
- If s = α $\diamondsuit_I$ p, s' = α $\diamondsuit_I$ p', and p < p', then s < s'; e.g. <(a,b,c)> < <(a,b,d)>.
- If s = α $\diamondsuit_S$ p, s' = α $_S$ p', and p < p', then s < s'; e.g. <(a,b)(c)> < <(a,b)(d)>.
- If s = α $_I$ p and s' = α $_S$ p', then s < s' regardless the order of p and p'; e.g. <(a,b,c)> < <(a,b)(a)>.

After the construction of the lexicographical tree, the across-streams sequential patterns in the tree will be updated with the production of new mining results. The across-streams sequential patterns that have not been updated for a long time may become obsolete. In maintaining the lexicographical tree, it is therefore necessary to decay their supports to delete those sequential patterns that have not been updated for a long time. The decay equations are as follows:

$$weight = d^p, \ 0 \le d \le 1 \qquad (2)$$

$$p = w - u \qquad (3)$$

$$w\_\sup = weight \times \sup \qquad (4)$$

In (2), d stands for the decay rate that is defined by the user to control the speed of decaying sequential patterns. In (3), p is the decay period of a sequential pattern, w is the number of times of sliding window movement, and u is the number of times of updating a sequential pattern in the lexicographical tree. The decay period of a sequential pattern is obtained by deducting the number of times of updating the sequential pattern from the number of times of sliding window movement. The longer a sequential pattern has not been updated, the bigger its p value becomes, which means its weight will become lower accordingly. In (4), w_sup stands for the weighted support a sequential pattern, which is obtained by multiplying the support of the sequential pattern by its weight. An across-streams sequential pattern will be deleted when its weighted support is less than the minimum support.

We will use an example to illustrate the IAspam algorithm. Table I is the CID-ordered customer sequence table of the data sampled for the first execution of IAspam

(W1). In a customer sequence table, the transaction data of each customer transaction sequence are sequenced according to the order of their arrival time. Because the size of the sliding window is set to 3, customer transaction data at 3 time points in each data stream are sampled. Each customer transaction data is represented as <S_i, α>, in which S_i is a stream number and α is an itemset. <S1, (a,c)> means that itemset (a,c) comes from stream S1. The empty sets in the first and third columns whose CID is 2 mean that there is no transaction at these 2 time points for customer #2.

TABLE I.
CUSTOMER SEQUENCE TABLE OF W1

CID	Customer transaction sequence		
1	<S1, (a,c)>	<S3, (b,d)>	<S2, (c,d)>
2	<∅>	<S2, (c,d)>	<∅>
3	<S2, (b,c,d)>	<S1, (a,c,d)>	<S3, (b,c,d)>
4	<S3, (b,c)>	<∅>	<S1, (a,b,c)>

After the data of W1 is mined, the sliding window will move forward one time point to sample data for the second execution of IAspam (W2). As shown in Table II, three pieces of new customer transaction data, <S2, (a,c,d)>, <S1, (c,d)>, and <S3, (a,d)>, are appended respectively to the customer transaction sequences whose CIDs are 1, 3 and 4. These three pieces of customer transaction data are the new stream data in W2. The customer transaction sequence whose CID is 2 does not have any new transaction data, which means this customer did not engage in any transaction at that time point.

TABLE II.
CUSTOMER SEQUENCE TABLE OF W2

CID	Customer transaction sequence		
1	<S3, (b,d)>	<S2, (c,d)>	<S2, (a,c,d)>
2	<S2, (c,d)>	<∅>	<∅>
3	<S1, (a,c,d)>	<S3, (b,c,d)>	<S1, (c,d)>
4	<∅>	<S1, (a,b,c)>	<S3, (a,d)>

Figure 5 is the bitmap matrix converted from the customer sequence table of W1, in which a, b, c, and d represent items and CID=1 represents customer #1. For each customer, there is a bit list for each item and these bit lists are converted from the transaction sequence of this customer. The bit list (1,0,0) for item a means that item a appears in the transaction data at the first time

point and does not appear in the transaction data at the second and third time points. The stream sequence <S1,S3,S2> means that the transaction data at the first, second, and third time points come respectively from streams S1, S3, and S2.
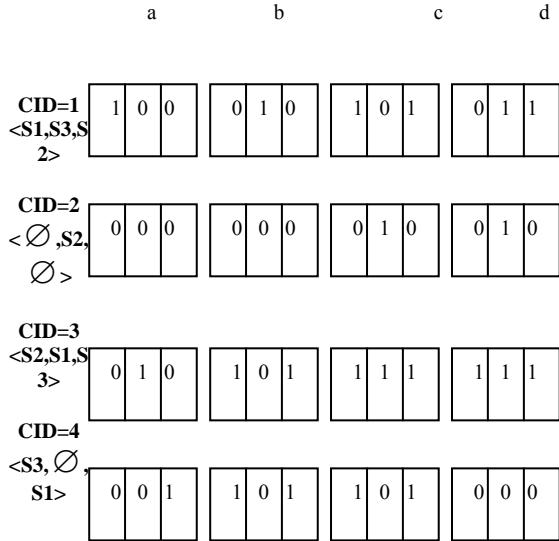

Figure 5. Bitmap matrix of W1

After the customer sequence table is converted into a bitmap matrix, the support of each 1-item will be counted to generate candidate sequences. Figure 6 shows the support of each 1-item in each stream. Suppose the minimum support is set to 2. The supports of c-item in S1, S2, and S3 are 3, 3, and 2 respectively, all of which are greater than or equal to the minimum support. Therefore, c-item is a frequent item in each of these three streams.

<a>   <b>   <c>   <d>
[S1:3,S2:0,S3:0] [S1:1,S2:1,S3:3] [S1:3,S2:3,S3:2] [S1:1,S2:3,S3:2]
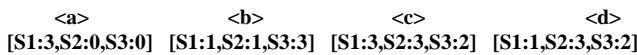
Figure 6. Support of 1-item

Based on the 1-item in Figure 6, sequence extension will go through I-step and S-step to generate candidate sequences, which will be mapped to the bitmap matrix. As shown in Figure 7, different extension steps use different ways of mapping, which are described below.

- I-step: The candidate itemset <(a,c)> extended using I-step will be mapped to the bitmap matrix and its support will be counted. As shown in Figure 8, a-item and c-item have bit 1 (the 1 in bold type) at the same time point for customers #1, #3, and #4 and all of them appear in S1. Therefore the support of <(a,c)> is 3. In Figure 6, as a-item is infrequent in S2 and S3, it will remain infrequent after I-step. Accordingly, it needs not to be processed.

- S-step: After going through S-step, <(a,c)> and <(b,d)> will become <(a,c)(b,d)>, which will be mapped to the bitmap matrix. Due to the time sequence of a sequential pattern, <(b,d)> must be mapped to a time point after that of <(a,c)>. As

shown in Figure 7, bit 1 (the underlined 1) appears at the first time point of a-item and c-item as well as the second time point of b-item and d-item for customer #1. Also, bit 1 (the underlined 1) appears at the second time point of a-item and c-item as well as the third time point of b-item and d-item for customer #3. In the meantime, <(a,c)> and <(b,d)> show up respectively in S1 and S3. Therefore, the support of <(a,c)(b,d)> is 2 and the stream sequence is <(S1)(S3)>.
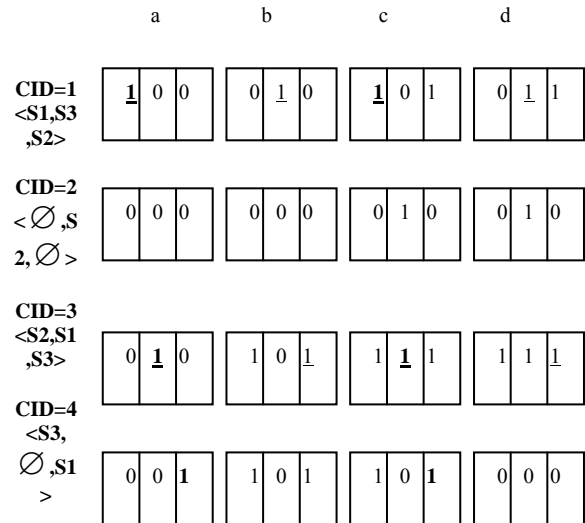
Figure 7. Mapping candidate sequences to bitmap matrix

Suppose the minimum support is set to 2. Figure 8 and Table III show separately the lexicographical tree and result of the first execution of IAspam. The across-streams sequential patterns in the lexicographical tree include <(a)(b)>, <(a)(d)>, <(c)(b)>, <(c)(d)>, and <(a,c)(b,d)>. The stream sequence of all of these across-streams sequential patterns is <(S1)(S3)>, which means that their moving path goes from stream S1 to stream S3. As <(a,c)(b,d)> is the supersequence of all the other across-streams sequential patterns, it will be the only frequent across-streams sequential pattern.
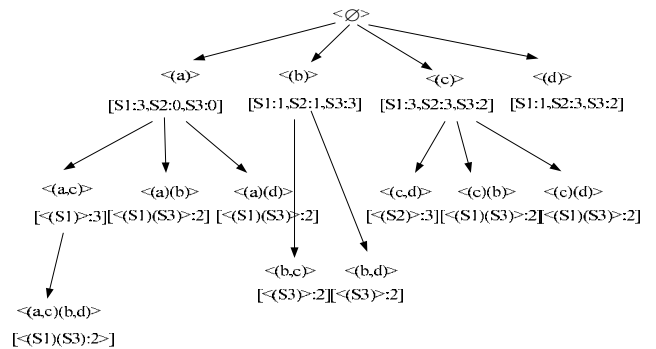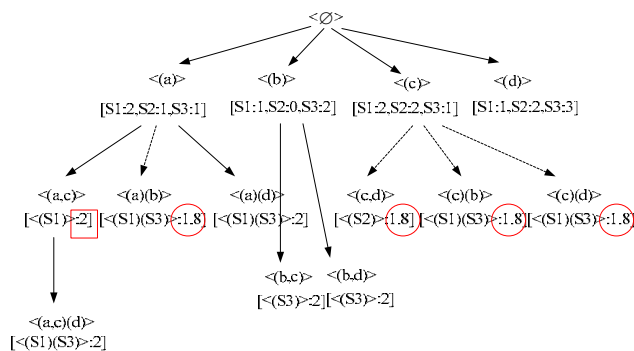

Figure 8. Lexicographical tree of the first execution

TABLE III.
RESULT OF THE FIRST EXECUTION

Frequent across-streams sequential pattern	Stream sequence
<(a,c)(b,d)>	<(S1)(S3)>

Figure 9 and Table IV show separately the lexicographical tree and result of the second execution of IAspam. In Figure 9, the support within a red square is an updated support, which indicates the sequential pattern remains frequent in W2 and its support is updated from 3 to 2. The support within a red circle is a weighted support, which is decayed from 2 to 1.8 under the condition that the decay rate is set to 0.9. Accordingly, the sequential pattern will be deleted because its support is less than the minimum support, and the link will be represented as a dotted line.

Figure 9. Lexicographical tree of the second execution

TABLE IV.
RESULT OF THE SECOND EXECUTION

Frequent across-streams sequential pattern	Stream sequence
<(a,c)(d)>	<(S1)(S3)>

IV. PERFORMANCE EVALUATION

A. Experimental Environment and Data

Table V shows the experimental environment. In Net-Beans IDE 6.1 compilation environment, Java JDK 6.0 programming language is used to write programs. A Pentium D 2.8 GHz processor and 1 GB memory are used. Transaction data generated by the synthetic data generator developed by IBM Almaden Research Center are used as experimental data. Table VI shows the parameters of experimental data. For instance, the expression S6T100I3 stands for a multiple data streams environment in which there are 6 data streams, 100 time points, and averagely 3 items per transaction.

TABLE V.
EXPERIMENTAL ENVIRONMENT

Equipment	Specification
Processor	Pentium D 2.8 GHz
Memory	1 GB
Hard Disk	160 GB
Operating System	WINDOWS XP

Programming Language	Java JDK 6.0

TABLE VI.
PARAMETERS OF EXPERIMENTAL DATA

Parameter	Parameter Description
S	Number of data streams
T	Number of time points
I	Average number of items per transaction
C	Number of customers
D	Number of item varieties

B. Performance Comparison

Both IAspam and T-Map-Mine are used for mining sequential patterns of moving paths in data stream environment. IAspam uses a bitmap representation for mining frequent across-streams sequential patterns; whereas T-Map-Mine uses a T-Map-Tree structure for mining behavior sequence patterns. In the mining process, T-Map-Mine needs to repeatedly search T-Map-Tree to insert new nodes and increase the support, and hence may take a great amount of time in constructing and maintaining T-Map-Tree. On the contrary, IAspam maps candidate sequences to a bitmap matrix based on the stream, which can save a lot of execution time while searching for across-streams sequential patterns. We compare the execution and memory usage of these two algorithms.

(1) Execution Time

Figure 10 shows the comparison of execution time between IAspam and T-Map-Mine in the S5T100C100I1 multiple data streams environment. The vertical axis is execution time in second and the horizontal axis is the minimum support in percentage. To be able to compare with T-Map-Mine, the average number of items per transaction is set to 1. The execution time of IAspam and T-Map-Mine are compared under various minimum supports. As seen in Figure 10, the execution time of IAspam is less than that of T-Map-Mine under various minimum supports. This is because T-Map-Mine must repeatedly search T-Map-Tree and RST for each customer data as well as insert new nodes and increase the support, which will take more time in the mining process.
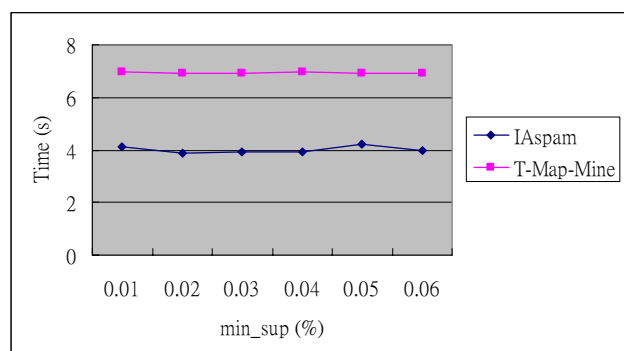
Figure 10. Comparison of execution time

(2) Memory Usage

Figure 11 shows the comparison of memory usage between IAspam and T-Map-Mine in the same environment as that for the comparison of execution time. The vertical axis is the maximum usage of memory in kilobyte (KB). As seen in Figure 11, the memory usage of IAspam is about the same as that of T-Map-Mine under various minimum supports. The memory usage of IAspam will be affected by the number of item varieties and the number of customers. The branching of T-Map-Tree of T-Map-Mine will depend on the complexity of customer transaction data. The more chaotic the data is, the bigger the T-Map-Tree becomes.
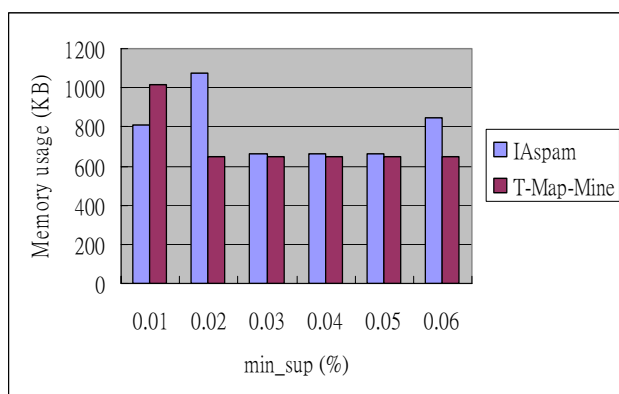
Figure 11. Comparison of memory usage

V. CONCLUSION

The primary difference between the IAspam algorithm proposed in this paper and existing algorithms for mining sequential patterns in multiple data streams is that IAspam is for mining across-streams sequential patterns, which are a new type of sequential patterns. To avoid inaccurate mining results caused by the re-mining of data, IAspam uses a sliding window to sample stream data and incrementally mines across-streams sequential patterns to maintain the newest mining results. To solve the inefficiency problem caused by mining a massive amount of data, IAspam uses a bitmap representation to search for across-streams sequential patterns and count their supports. As sequential patterns may cross different streams, IAspam searches for sequential patterns based on the stream, which enables the identification of not only across-streams sequential patterns but also stream sequences. Experimental results show that the IAspam algorithm is effective in execution time when processing large amounts of stream data.

ACKNOWLEDGMENT

REFERENCES

[1] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and Issues in Data Stream Systems. The 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. 2002 June 1-16; Madison, Wisconsin, USA.

[2] T. Oates and P. R. Cohen. Searching for Structure in Multiple Streams of Data. The 13th International Conference on Machine Learning. 1996 July 346-354; Bari, Italy.

[3] L. Golab and M. T. Ozsu. Issues in Data Stream Management. ACM SIGMOD Record. 2003 June Volume 32 Issue 2: 5-14.

[4] J. Chang and W. Lee. Decaying Obsolete Information in Finding Recent Frequent Itemsets over Data Stream. IEICE Transaction on Information and Systems, June, 2004; No. 6, Vol. E87-D.

[5] G. Chen, X. Wu, and X. Zhu. Sequential Pattern Mining in Multiple Streams. The 5th IEEE International Conference on Data Mining, Washington. 2005 Nobember 27-30; USA.

[6] J. H. Chang and W. S. Lee. Finding Recent Frequent Itemsets Adaptively over Online Data Streams. The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2003 August 487-492; Washington, USA.

[7] C. Raissi, P. Poncelet, and M. Teisseire. SPEED: Mining Maximal Sequential Patterns over Data Streams. The 3rd IEEE International Conference on Intelligent Systems. 2006 September 546-552; London, UK.

[8] C. I. Ezeife and M. Monwar. SSM：A Frequent Sequential Data Stream Patterns Miner. The IEEE Symposium on Computational Intelligence and Data Mining, Honolulu. 2007 March 120-126; USA.

[9] H. Li and H. Chen. GraSeq: A Novel Approximate Mining Approach of Sequential Patterns over Data Stream. The 3rd International conference on Advanced Data Mining and Applications. 2007 May 401-411; Harbin, China.

[10] S.-C. Lee, E. Lee, W. Choi, and U. M. Kim. Extracting Temporal Behavior Patterns of Mobile User. The 4th International Conference on Networked Computing and Advanced Information Management. 2008 September 455-462; Gyeongju, Korea.

Shih-Yang Yang received his Ph.D. degree in Computer Science and Information Engineering from Tamkang University, Taiwan, in January 2008. Since January 2008, he is an Associate Professor with the Department of Media Art at Kang-Ning Junior College of Medical Care and Management (Taipei, Taiwan). His research interests include parallel & distributed systems, web technology, and multimedia.

Ching-Ming Chao received his B.S. degree in Computer Science from Soochow University, Taipei, Taiwan in 1982 and his Ph.D. degree in Computer Science from The University of Iowa, Iowa City, Iowa, U.S.A. in 1990. From 1990 to 1992, he was an assistant professor in the Department of Computing Sciences at the University of Scranton, Scranton, Pennsylvania, U.S.A. He joined the faculty of the Department of Computer and Information Science at Soochow University in 1992 as an

associate professor. From 1992 to 1996, he served as the department chair. Since 2003, he has been a professor in Department of Computer Science and Information Management at Soochow University. During the past few years, he has published more than 90 refereed journal and conference papers. His research interests include data mining, data warehousing, database, and web technology.

Po-Zung Chen received his Ph.D. degree in Computer Science from the University of Iowa in December 1989. From November 1989 to May 1990, he was a visiting Assistant Professor at Michigan Technological University (Houghton, Michigan). Since August 1990, he is an Associate Professor with the Department of Computer Science and Information Engineering at Tamkang University (Taipei, Taiwan). His research interests include object-oriented distributed programming, parallel & distributed systems, and simulation & modeling.

Chu-Hao Sun is currently a candidate of Ph.D. student in department of Computer Science and Information Engineering in Tamkang University (Taipei, Taiwan). He received his B.E. and M.E. degree from the same university in 1995 and 1998. His research interests include database management system, parallel processing, web technology, and data mining.

The Effect of Online Training on Employee's Performance

Shu Ching Yang
Department of Accounting Information
Aletheia University, Taiwan
Au4335@mail.au.edu.tw
Chin Hung Lin
Department of Information Management
Chung Chou Institute of Technology, Taiwan
chlin@dragon.ccut.edu.tw

Abstract—The new telecommunication technology has opened educational opportunities to learners who are having difficulty participating in traditional instruction. Enhanced educational technology has played a critical role in professional development of employees in the business field, as online distance learning instruction is one of the popular options for so many instructional designers in business. This study sought to determine the effect of online distance-learning instruction on employees' learning achievement those who took a training program of online distance-learning instruction and those who did not. Since many factors will affect employees' learning achievement, this study also explored the effect of self-efficacy, gender, computer experience, and socioeconomic status, on the learning achievement of employees in a Taiwanese manufacturing company. Additionally, this study investigated students' satisfaction with online distance-learning instruction. The major findings were that learning achievement was similar for online distance learning and traditional face-to-face instruction. Also, it showed a significant relationship between self-efficacy and learning achievement. An implication is that a judicious embedding of self-efficacy consideration is the design and implementation of online distance learning courses might well enhance learning achievement.

Index Terms—online distance-learning, self-efficacy, employees' training

I. INTRODUCTION

In accordance to the global economy development and the need, educational training provides the transformation ability to employees individually and the whole organization. The enterprise's development depends on fostering the talent as well as displaying the ability of the talented person. The educational training is the key work fostering talented employees for an enterprise to continue indefinitely [21].

Without doubt, the target of educational training is to prompt personal skill, mold personal independence, and develop self-confidence. Moreover, regardless of theory or practice in educational training, the educational training links closely with the external environmental trend, the enterprise growth, and the professional development of employees [21]. Although the importance of educational training is continuously mentioned, when knowledge and skill of employees are no longer suitable for today' need, educational training will be urgenting needed, especially under the pressure of transformation.

Advances of educational technology have impacted curricula and the ways content is delivered and received in today's educational world. The new telecommunication technology has opened educational opportunities to learners who are having difficulty participating in traditional instruction. Enhanced educational technology has played a critical role in professional development of employees in the business field, as online distance-learning instruction is one of the popular options for so many instructional designers in business. Many educational departments in business organizations have been attempting to build an educational system accessible through distance learning in order to serve more employees in different branches efficiently.

The purpose of this study was to examine the effectiveness of online distance-learning instruction on the performance of employees training in a Taiwanese industry of manufacturing. Concurrently, this study sought to determine the effect of online distance-learning instruction on the learning achievement of employees who took the training program via online distance-learning instruction and those who did not. Since many factors affect employees' learning achievement, this study also explored the effect of self-efficacy, gender, computer experience, and socioeconomic status, on the learning achievement of employees in a Taiwanese manufacturing industry. Additionally, this study explored the issue of students' satisfaction with online distance-learning instruction.

II. THEORETICAL FRAMEWORK

A. Self-Efficacy

Bandura [1] defined self-efficacy as the personal judgment of one's capabilities to organize and execute courses of action to attain designated goals. Bandura's research identified self-efficacy as a significant predictor of a student's learning achievement.

Learning achievement can be influenced by various factors, both internal and external. However, self-

efficacy is rooted in the core belief that one has the power to produce change by one's actions [1]. Self-efficacy can be extended beyond specific tasks or attainment goals – a person can have self-confidence in general.

Bandura held that beliefs about one's likelihood of success were better predictors of success [2]. Salanova *et al.* have shown that computer self-efficacy plays a moderating role between technology training[24]. Computer training and use may be associated with increased efficacy beliefs about computer use, which may contribute to an increased motivation to use technology.

From such a fundamental and sweeping view it can be concluded that self-efficacy is a paramount consideration in education. Levine and Donitsa-Schmidt [17] conducted a study, which demonstrated that high computer self-efficacy correlated with high computer-related knowledge. An investigation by Joo, Bong and Choi [14] led to the conclusion that computer self-efficacy is one of the critical variables determining the success of online learning. Similar results were obtained in studies by [12].

Experience using computers will strongly influence computer self-efficacy. Britner and Pajares found that after a computer training course participants exhibited a higher level of computer self-efficacy [3]. Smith found a high correlation between computer experience and computer self-efficacy[26]. Cassidy and Eachus reported a positive relationship between self-efficacy and computing experience[4]. In the context of prior researches, women have lower levels of self-efficacy/computer self-efficacy towards computers or the Internet [9-11]. The researches strongly suggested that in the promotion of self-efficacy in students by teachers, particular attention should be devoted to computer self-efficacy.

B. Learning Achievement in Distance Learning

It was stated that the literature generally reached a favorable comparison (i.e., roughly equal) when such types of learning were compared to traditional face-to-face classroom instruction. A comparison of students' learning achievement in distance learning courses versus traditional courses was done by Lockyer et al. [18]. There is no different in knowledge acquisition and learners' achievement did not show any significant difference. Clark based on reviewing hundreds of studies, concluded that there was no significant difference in learners learning achievement [6]. There are evidences showing different conclusion in learning achievement. Learners have better performance in face-to-face instruction, but web learners worked more in groups [28]. Learners use web practice quizzes improving their performance [13]. Meelissen and Drent indicated that learners from more privileged SES (socioeconomic status) background tend to have more positive computer attitude and better performance than les privileged peers[19].

C. Experience with Computer in Distance Learning

Experience with technology, including computer experience, was identified as a crucial element of success for online distance learners [25]. The Internet has become widely and frequently used in a wide variety of human activities, with its role often being vital. But use of the Internet is inseparable from computer usage; use of the first requires use of the second. The importance of computer experience is therefore unavoidably implied in a study of 152 high school students in Korea [14] found Internet self-efficacy to be an important variable with respect to student success in a web-based learning environment. The Joo, Bong and Choi study shows that experience of using computer is critical for further research and experience in using computer can be defined as the years and ability in using the computer [24]. In a study of 122 college students taking a course in research methodologies it was found that, in addition to self efficacy and experience with computer respecting course content, technological self-confidence on the part of students were good predictors of learner performance in the class [25]. From context of Meelissen and Drent shown that learners from higher socioeconomic status have more positive computer attitudes than their lower socioeconomic status peers[19]. This may be because learners from privileged families have more opportunities to use computers.

D. Learner Satisfaction in Distance Learning

To be truly effective, learning results should include student satisfaction. By the end of a course, the student should not only have acquired the course-targeted knowledge and skills, but also have the belief and feeling that the course successfully met his/her expectations. Perhaps the most immediate and obvious measure of program effectiveness focuses on the quality of the individual learning experience [8]. Chute stated that learner satisfaction is significant to all facets of distance learning [7]. It relates to design, development, and delivery. Technology-based learning is becoming increasingly popular, particularly by means of the World Wide Web (Internet). It is therefore becoming increasingly important that educators determine and analyze the expectations and experiences of the learner. Consequently, a number of studies have been conducted aimed at investigating the level of students' satisfaction associated with distance learning courses.

One such study was done by Sahin indicated that distance-learning students generally had a high level of satisfaction across-the-board [23]. The satisfaction applied to the use of technology, course content, and the support they received from instructors and mentors. Drennan, Kennedy, and Pisarski conducted a study to examine the factors that affect student satisfaction. The students had little of any prior experience with learning via the Internet [9]. The results showed that the course was very successful in introducing students to Internet-based instructional material. However, the students were subsequently reluctant to disassociate themselves from the familiar forms of tutor contact. Drennan, Kennedy,

and Pisarski suggested that although the students were accepting of supporting material from the Web, they might be unwilling to replace a good instructor with technology-based instruction [9]. In the context of Lim, computer self-efficacy has a positive relationship between learners' satisfaction with their Web-based distance education courses and their intent to participate in future Web-based courses [15]. The following hypotheses have been developed to explore research.

H1: There is a significant difference in employees' learning achievement between employees who receive online distance-learning instruction and employees who receive face-to-face instruction.

H2: There is a significant difference in employees' learning achievement between gender, employees' experience of using computer, and socioeconomic status for employees who receive online distance-learning instruction and employees who receive face-to-face instruction.

H3: There is a significant relationship in employees' learning achievement and self-efficacy between employees who receive online distance-learning instruction and employees who receive face-to-face instruction.

H4: There is a significant difference in satisfaction between employees who receive online distance-learning instruction and employees who receive face-to-face instruction.

H5: There is a significant combined effect of self-efficacy and/or gender and/or satisfaction and/or computer experience and/or socioeconomic status on the employees' learning achievement in Taiwanese manufacturing industry.

III METHOD

A. Design

This study is based on a quasi-experimental nonequivalent control group design. The pretest/posttest procedure was applied. Random assignment technique was not used in this study. The mean gain from the posttest was utilized to test for significant differences between experimental and control groups. The independent variable in this study is distance learning instruction and the dependent variable is learning achievement. The subjects in this study are employees of the Yu-Yi Ltd. Co. in Taiwan.

Online distance-learning provides a cost-effective solution to the demand for training at the company headquarters and branches at the same time. There were two groups of subjects in this study. One in Mainland China is the experimental group receiving online distance-learning instruction. The other one in Taiwan is a control group receiving face-to-face instruction. The experimental group consisted 14 employees and the control group consisted 13 employees selected randomly from volunteers. In addition, these two groups met at exactly the same time and share the same trainer. There were 27 participants in this study. Their ages range from

24 to 33; they are all full- time employees, who have earned at least a bachelor's degree.

Data was collected from employees of Yu-Yi Ltd. Co. who enrolled in the required training program offered both at online distance-learning instruction and face-to-face instruction. Both groups met on the same day for two hours once a week, over a time period of four weeks.

The Online Technologies Self-Efficacy Scale (OTSES) and Self-Efficacy Toward Classroom Learning instrument was given as pretests at the beginning of the course to the both of the experimental group and the control group. A Background Information Survey was administered as a pretest to both groups as well. The training is titled How to Use Entry System of Yu-Yi (HUESYY) and the posttest referred to how employees have learned to utilize HUESYY. The posttest was titled Testing Skill of HUESYY. It also served as a pretest for both the experimental and control groups for check equivalence of the knowledge of HUESYY in the beginning of the training program. The Satisfaction Instrument served as a posttest to both experimental and control groups to evaluate how satisfied employees were with this training program. The posttest was scheduled for both groups at the end of the training program. The test results were analyzed to see if there is a significant difference in learning achievement between the experimental group taking online distance-learning instruction and the control group taking face-to face instruction. It was necessary to note that the effects of variables such as teaching style would not be confounding variables because the same trainer taught both the experimental and the control groups.

B. Instrument

To measure self-efficacy for employees who receive online distance-learning instruction and face-to-face instruction the researcher used The Online Technologies Self-Efficacy Scale (OTSES) and Self-Efficacy Toward Classroom Learning. The Online Technologies Self-Efficacy Scale (OTSES) developed by Miltiadou and Yu [20]. Internal consistency reliability (Cronbach's coefficient alpha) is .95 for all the 29 items. Self-Efficacy Toward Classroom Learning developed by Quinones and The acceptable internal consistency reliability (Cronbach's coefficient alpha) is .76 for all the 9 items[22]. Satisfaction Instrument developed by Lee is used to measure the satisfaction of employees who received instruction by distance learning and face-to-face [16]. The acceptable internal consistency reliability (Cronbach's coefficient alpha) is .93 for all the 19 items. The learning achievement pretest and learning achievement posttest were designed by the instructor to measure the knowledge that participants, should gain through the instruction provided in How to Use the Entry System of Yu-Yi. The background information survey included employees' name, gender, computer experience and socioeconomic status.

IV DATA ANALYSIS AND RESULTS

The researcher used the Statistical Package for the Social Sciences (SPSS) 16.0 for Windows to organize and test data. An analysis of the data gathered by a demographic questionnaire, OTSES scores, Self-Efficacy Toward Classroom Learning scores, satisfaction Instrument score and the score of achievement test of employees in a Taiwanese manufacturing industry.

A. Description of Population Sample

Employees' Characteristics included gender, Computer Experience, and Socioeconomic Status in this sample. The Frequencies and Percentages for Employees' Characteristics including experimental group and control group are showed in Table 1.

In Hypothesis One, the independent simple t-test was conducted. The perfect score of learning achievement test is 100. The means of the achievement pretest for the control and experimental groups were 51.46 and 44

respectively. The learning achievement pretest revealed no statistically significant difference between the two groups, F= .03, p= .170>.05. The descriptive statistics for these test scores are presented in Table II. The means of the achievement posttest for the control group and experimental group were 79.92 and 79.57 respectively. The learning achievement posttest revealed no statistically significant difference between the two groups, F= .017, p= .928>.05. The descriptive statistics for these test scores are presented in Table II.

One-way ANOVA was conducted in order to evaluate Hypothesis One. The descriptive statistics for learning achievement posttest are shown in Table III. The learning achievement posttest revealed no statistically significant difference between the two groups, $F_{(1,25)}$=.008, p= .928>.05. Hypothesis One was rejected since there was no significant difference in employees' learning achievement between employees who received online distance-learning instruction and employees who received face-to-face instruction.

TABLE I.
Frequencies and Percentages for Employees' Characteristics

Descriptor	Experimental N (%)	Control N (%)	Total N (%)
Gender:			
Males	8(57.1%)	6(46.1%)	14(51.9%)
Females	6(42.9%)	7(53.9%)	13(48.1%)
Total	14(100%)	13(100%)	27(100%)
Level of Computer Experience:			
1 year and below	2 (14.29%)	2(15.39%)	4(14.81%)
1 to 3 year	4 (28.57%)	5 (38.46%	9(33.33%)
3 to 5 year	4 (28.57%)	5(38.46%)	9(33.33%)
5 year and above	4 (28.57%)	1 (7.69%)	5(18.59%)
Total	14 (100%)	13 (100%)	27 (100%)
Socioeconomic' Status:			
24,000 to 95,999	2 (14.29%)	1 (7.69%)	3 (11%)
36,000 to 47,999	2 (14.29%)	6 (46.15%)	8 (29.62%)
48,000 to 59,999	3 (21.43%)	3 (23.07%)	6 (22.22%)
60,000 to 71,999	3 (21.43%)	1 (7.69%)	4 (14.81%)
72,000 to 95,999	3 (21.43%)	2 (15.38%)	5 (18.51%)
96,000 and above	1 (7.14%)	0 (0%)	1 (3.70%)
Total	14 (100%)	13 (100 %)	27 (100%)

TABLE II
INDEPENDENT T-TEST ACHIEVEMENT PRETEST SCORES AND ACHIEVEMENT POSTTEST SCORES FOR EMPLOYEES BY GROUP

Group	N	Mean	SD	F	p
Pretest Scores					
Experimental	14	44	12.83	.03	.170
Control	13	51.46	14.62		
Posttest Scores					
Experimental	14	79.57	9.95	.017	.928
Control	13	79.92	10.00		

TABLE III
ONE-WAY ANOVA LEANDERING ACHIEVEMENT POSTTEST SCORES FOR EMPLOYEES

Source	SS	df	MS	F	p
Group	0.843	1	0.834	.008	.928
Error	2488.352	25	99.534		
Total	174171.00	27			

TABLE IV
ANCOVA OF GENDER, COMPUTER EXPERIENCE AND SOCIOECONOMIC STATUS FOR EMPLOYEES

Factors	DF	F	p	Error
Gender	1	2.017	.251	3
Computer experience	3	.262	.85	3
Socioeconomic status	5	1.64	.363	3

TABLE V
CORRELATIONS OF SELF-EFFACACY SCORES FOR EMPLOYEES WHO RECEIVE FACE-TO-FACE INSTRUCTION AND DISTANCE-LEARNING INSTRUCTION

Instrument	r	p
Face-to-face Instruction		
OTSES	.922	.000
Self-Efficacy Toward Classroom Learning	.844	.000
Distance-Learning Instruction		
OTSES	.930	.000
Self-Efficacy Toward Classroom Learning	.946	.000

In Hypothesis Two, in order to test for differences in employees' learning achievement between gender, employees' experience of using computer, and socioeconomic status for employees who receive online distance-learning instruction and employees who receive face-to-face instruction, a one-way analysis of Covariance (ANCOVA) was conducted. The independent variable was distance-learning instruction and the dependent variable was employees' learning achievement. Table IV shows that there was no statistically significant difference between gender, employees' experience of using computer and socioeconomic status of employees and learning achievement. F (1,3)=2.017, p=.251>.01; F(3,3)=.262, p=.850>.01; F(5,3)=1.64, p=.363>.01.

In Hypothesis Three, In order to test for a relationship of employees' learning achievement between employees who received online distance-learning instruction and employees who receive face-to-face instruction, the Pearson correlations was conducted. Correlation coefficients were computed for the relationship between employees' learning achievement and self-efficacy. Results revealed a statistically significant relationship between learning achievement and The Online Technologies Self-Efficacy Scale (OTSES), (r = .922) in control group. Results also show a statistically significant relationship between learning achievement and Self-Efficacy Toward Classroom Learning, (r = .844) in the control group. The descriptive statistics for these test scores are presented in Table V.

Results revealed a statistically significant relationship between learning achievement and The Online Technologies Self-Efficacy Scale (OTSES), (r = .930) in the experimental group. Results also show a statistically significant relationship between learning achievement and Self-Efficacy Toward Classroom Learning, (r = .946) in the experimental group. The descriptive statistics for these test scores are presented in Table V.

In Hypothesis Four, in order to test for difference of employees' satisfaction between employees who receive online distance-learning instruction and who employees receive face-to-face instruction; the independent simple t-test was conducted. The .05 level of significance was selected for analysis of data. Results of the analysis revealed no statistically significant differences in satisfaction between two groups, F=1.91, P=.693>.05. The means of experimental and control group were 87.57 and 88.46. The descriptive statistics for these test scores are presented in Table VI.

In order to test for difference in employees' satisfaction between employees who received online distance-learning instruction and employees who received face-to-face instruction. A one- way ANOVA was conducted. The .05 level of significance was selected for the analysis of data. Results of the analysis revealed no statistically significant differences in satisfaction between two groups, F (1,25)=0.159, P=.693>.05. The descriptive statistics are presented in Table VII. Hypothesis Four was rejected because there was no significant difference in satisfaction between employees who received online distance-learning instruction and employees who received face-to-face instruction.

In Hypothesis Five, Correlation coefficient was conducted to test the combined effects of self-efficacy and/or gender and/or satisfaction and/or computer experience and/or socioeconomic status on the employees' learning achievement in the Taiwanese manufacturing industry. The Bonferroni approach control for Type I error across the correlations. The descriptive statistics for these test scores are presented in Table VIII.The correlation between satisfaction and computer experience was significant, r (27)=. 603.The correlation between socioeconomic status and satisfaction was significant, r (27)=. 541. Results revealed there is a significant relationship between satisfaction and self-efficacy, r (27)=. 946; r (27)=. 902. (See Figure 1 and 2). The correlation between gender and socioeconomic status was no significant, r (27)=-.077. Results revealed there is no significant relationship between gender and self-efficacy, r (27)=-.078; r (27)=-.198. The correlation between computer experience and socioeconomic status was significant, r (27)=. 620. Results revealed there is no significant relationship between socioeconomic status and self-efficacy, r (27) =. 497; r (27) =. 461.

TABLE VI
INDEPENDENT T-TEST OF SATISFACTION SCORES FOR EMPLOYEES BY GROUP

Group	N	Mean	SD	F	p
Experimental	14	87.57	6.58	1.91	.693
Control	13	88.46	4.78		

TABLE VII
ONE-WAY ANOVA OF SATISFACTION SCORES FOR EMPLOYEES

Source	SS	df	MS	F	p
Group	5.341	1	5.341	.159	.693
Error	838.659	25	33.546		
Total	209932.00 27				

TABLE VIII
COLORATION OF COMBINED EFFECT SELF-EFFACACY AND/OR GENDER AND/OR SATISFACTION AND/OR COMPUTER EXPERIENCE AND/OR SOCIOECONOMIC STATUS FOR EMPLOYEES

Factors	OTSES	SETCL	Gender	CE	SS	Satisfaction
OTSES			-.078	.511	.497	.946
SETCL			-.198	.559	.461	.902
Gender	-.078	-.198		-.172	-.077	-.053
Computer Experience(CE)	.511	.559	-.172		.620	.603
Socioeconomic Status(SS)	.497	.461	-.077	.620		.541
Satisfaction	.946	.902	-.053	.603	.541	
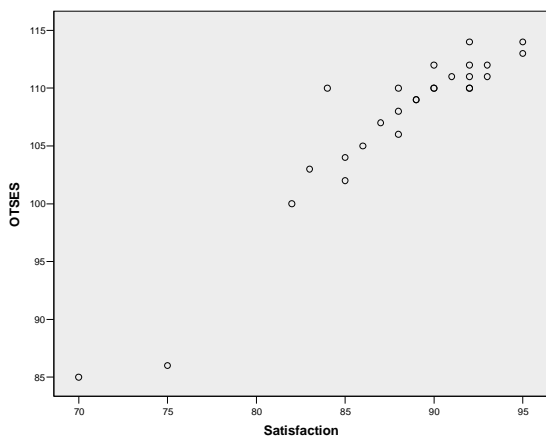
Figure 1

Slopes between Satisfaction and OTSES of Employees
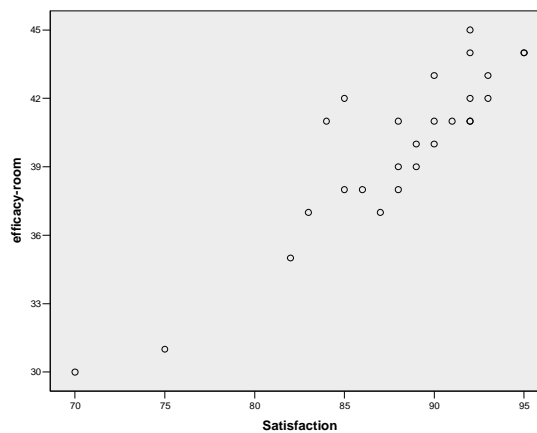
Figure 2.

Slopes between Satisfaction and Self-Efficacy Toward Classroom Learning of Employees

V. DISCUSSION

Results of Hypothesis One indicated that there was no significant difference in learning achievement between employees who received online distance-learning instruction and employees who received face-to-face instruction. The achievement pretest revealed no statistically significant difference between two groups. The statistics showed that the groups had equivalent ability in the beginning of this training program. The findings supports several previous studies conducted by Lockyer et al., and Clark , which indicated that there is no significant difference between distance-learning instruction and face-to-face instruction [5,18].

Results of Hypothesis Two indicated that there is no significant difference in employees' learning achievement between gender, employees' experience of using computer, and socioeconomic status for employees who receive online distance-learning instruction and employees who receive face-to-face instruction. Data showed that there was no statistically significant difference between gender, employees' experience of using computer and socioeconomic status of employees and learning achievement. This finding, which is opposite of what Meelissen and Drent found, are a relatively small sample size and the short-term training program. A relatively sample size is a possible reason for this finding [19].

Results of Hypothesis Three indicated that there is a significant relationship in employees' learning achievement and self-efficacy between employees who received online distance-learning instruction and employees who received face-to-face instruction. The results showed a statistically significant positive relationship between employees' learning achievement score and self-efficacy scores (Self-Efficacy Scale (OTSES) and Self-Efficacy Toward Classroom Learning). The self-efficacy score had an effect on employees' learning achievement scores. It might be inferred that employees who had higher self-efficacy scores than the other employees tended to achieve higher learning achievement scores by the end of the training program. The findings support several previous studies [2,14,12,24], in which a positive relationship between self-efficacy and employees' learning achievement was found the higher employees' self-efficacy, the greater the possibility of higher employees' leaning achievement at the end of the training program. Researcher accepted the research hypothesis.

Results of Hypothesis Four indicated that there is no significant difference in satisfaction between employees who received online distance-learning instruction and employees who received face-to-face instruction. Results of the analysis revealed no statistically significant differences in satisfaction between two the groups. Although the mean difference in satisfaction scores was not statistically significant, the satisfaction scores for the control group were slightly higher than those of the experimental group. This finding, which is opposite of what Sahin and Drennan, Kennedy, and Pisarski found,

are a relatively small sample size and the short-term training program [9, 23].

Results of Hypothesis Five indicated that there is a significant combined effect of self-efficacy and/or gender and/or satisfaction and/or computer experience and/or socioeconomic status on the employees' learning achievement in Taiwanese manufacturing industry.

Results from data organized revealed there was a significant relationship between satisfaction and self-efficacy. The correlation between satisfaction and computer experience was significant. The correlation between socioeconomic status and satisfaction was significant. The correlation between gender and socioeconomic status was no significant. Results revealed there is no significant relationship between gender and self-efficacy. The correlation between computer experience and socioeconomic status was significant. Results revealed there is no significant relationship between socioeconomic status and self-efficacy. The findings showed satisfaction has a positive relationship with self-efficacy supported by Lim, computer experience and socioeconomic status supported by Meelissen and Drent [15, 19]. It might be inferred that employees who have higher self-efficacy, computer experience and socioeconomic status have higher satisfaction with instruction. The results also inferred that employees who have higher level of socioeconomic status have more computer experience. There are two possible factors contributing to this finding: a sample size and the other short-term training program.

VI. CONCLUSIONS

This study referred that the importance of the way in which technology is employed in the design the curriculum to meet learners' needs. The use of online distance learning in training has had a great impact on learners. Understanding the self-efficacy for learners may allow distance educators and training developers to develop and consider strategies that better meet the learners' needs. This study could provide the basis for better understanding the needs of distance learners and lead to the design and the development of more effective distance-learning environments. The outcome of this study has widespread implications for educational institutions and individualized instruction.

REFERENCES

[1] A. Bandura, "Social cognitive theory: An agentic perspective," Asian Journal of Social Psychology, vol. 2, pp. 21-41., 1999.

[2] Bandura, A. (2008).Self-Efficacy in Changing Societies. Cambridge , UK:Cambridge University Press

[3] S. L. Britner, and F. Pajares, "Sources of science self-efficacy beliefs of middle school science," Journal of Women and Minorities in Science Enginerrin. vol. 7, pp. 271-285. , 2006.

[4]   S. Cassidy, and P. Eachus, "Developing the Computer User Self-Efficacy (CFUSE) Scale: Investigating the relationship between computers," Journal of Educational Computing Research, vol. 26(2), pp.133-153., 2002.

[5]   R. E. Clark, "Reconsidering research on learning from media," Review of Educational Research, vol.53, pp.445-459., 1983.

[6]   R. E. Clark, "What works in distance learning: Motivation strategies," In H. O'Neill (Ed.), What works in distance learning: Guidelines. Greenwich, CT: Information Age Publishers., 2004.

[7]   A. Chute, "Evolving from distance education to knowledge management and e-learning." Presented at Ohio Learning Network 2002 Conference. Retrieve December 20, 2004, from http://www.oln.org/conferences/OlN2002/OLN2002webcasts.php

[8]   A. Chute, M. Thompson, and B. Hancock, "The McGraw-Hill Handbook of Distance Learning". New York: McGraw-Hill, 1999.

[9]   J. Drennan, J. Kennedy, and A. Pisarski, "Factors affecting student attitudes toward flexible online learning in management education," The Journal of Educational Research, vol30, pp.331-338., 2005.

[10]   A. Durndell, and A. Hagg, "Computer self-efficacy, computer anxiety, attitudes towards the Internet and reported experience with the internet, by gender, In an European sample," Computers in Human Behavior, vol. 18(5), pp. 521-535., 2002.

[11]   A. Durndell, A. Hagg, and H. Laithwaite, "Computer self-efficacy and gender: a cross culture study of Scotland and Romania," Personality and Individual Differences, vol. 28(6), pp.1037-1044., 2000.

[12]   M. S. Eastin, and R. LaRose, "Internet self-efficacy and the psychology of thedigital divide," Journal of Computer Mediated Communication. RetrievedMarch8,2005,fromhttp://www.ascusc.org/jcmc/vo16/issue1/eastin.html.Galagan,P. (2000). Getting started with e-learning: Training and development. ASTD, vol. 43 (4), pp.62-64., September, 2000.

[13]   G. Huon, B. Spehar, P. Adam, and W. Rifkin, "Resource use and academic performance among first year psychology students," Higher Education, vol.53, pp.1-27., 2007.

[14]   Y. Joo, M. bong, and H. Choi, "Self-efficacy for self-regulated learning, academic self-efficacy, and Internet self-efficacy in Web-basd instruction," Educational Technology Research and Development, vol.48 (2), pp.5-17., 2000.

[15]   C. K. Lim, "Computer self-efficacy, academic self-concept, and other predictors of satisfaction and future participation of adult distance learners," American Journal of Distance Learning, vol. 15(2), pp.41-51., 2001.

[16]   C. Y. Lee, "The impact of self-efficacy and task value on satisfaction and performance in a Web-based course," Unpublished doctoral dissertation, Florida Central University, Orlando. 2002.

[17]   T. Levine, and S. "Donitsa-Schmidt, Computer use, confidence, attitudes, and knowledge:A causal analysis," Computer in Human Behavior, vol. 14, pp.125-146., 1998.

[18]   L. Locyker, J. Patterson, and B. Harper, "ICT in higher education: Evaluating outcomes for health education," Journal of Computer Assisted Learning, vol.17, pp.275-283., 2001.

[19]   M. R. M. Meelissen, and M. Drent, "Gender differences in computer attitudes: Does the school matter?" Computers in Human Behavior, vol., 24(3), pp.969-985., 2008.

[20]   M. Miltiadou, and C. H. Yu, "Validation of online technologies self-efficacy scale (OTSES) (n.d.)," Dr. Alex Yu. Retrieved November 4, 2004,fromhttp://seamonkey.ed.asu.edu/~alex/pub/efficacy.pdf , 2000.

[21]   P. Pont, "Developing Effective Training Skills," London, UK :Chartered Institute of Personnel & Development, (2003).

[22]   M. A. Quinones, "Pretraining context effects: Training Assignment as feedback," Journal of Applied Psychology, vol.,80, pp.226-236., 1995.

[23]   I. Sahin, "Predicting student satisfaction in distance education and learning environments," Journal of distance Education, vol.,19(3), pp.14-26., (2006).

[24]   M. Salanova, R. Crau, E. Cifre, and S. Llorens, "Computer training, frequency of usage and burnout: The moderating role of computer self-efficacy," Computers in Human Behavior, 16(6), pp.575-590., 2000.

[25]   L. Schrum, and S. Hong, "Dimensions and strategies for online success: voice from experienced educators," Journal of Asynchronous Learning Networks, 6 (1). Retrieved November 15,2004,fromhttp://www.aln.org/alnweb/journal/Vo16_issue1/6_lschrum.htm, (2002).

[26]   S. Smith, "An examination of computer self-efficacy and computer-related task performance relationship," The Journal of continuing Higher Education, vol. 53(4), pp.9-18., 2005.

[27]   A. Y. Wang, and M. H. Newlin, "Characteristics of students who enroll and succeed in psychology Web-based classes," Journal of Educational Psychology, vol., 92(1), pp.137-143., 2000.

[28]   T. Waits, and L. Lewis, "Distance education at degree granting post secondary institutions:U.S," Department of Education, National Center for Educational Statistics. 2003.

[29]   Z., Yang, and Q. Liu, "Research and development of web-based virtual online classroom, "Computer and Education, vol., 48(2), pp.171-184., 2007.

A New Intelligent Topic Extraction Model on Web

Ming Xie

Computer School of Wuhan University, Wuhan, China; Guangxi Economic Management Cadre College, Nanning, China

Email: wolfgangtse@gmail.com

Chanle Wu* and Yunlu Zhang

Computer School of Wuhan University, Wuhan, China

Email: wuchle@whu.edu.cn, zhang.yunlu850527@gmail.com

Abstract—**We tackle the problem of topic extraction on Web. In this paper, we propose an approach to implementing ontology-based data access in WordNet with the distinguishing feature of optimizing density-based clustering OPTICS algorithm (DBCO) to extract topics. Our solution has the following two desirable properties: i) it uses WordNet for word sense disambiguation of words in the learning resources documents and ii) it mapping the data space of the original method to a vector space of sentence, improving the original OPTICS algorithm. We outline the interface between our scheme and the current data Web, and show that, in contrast to the existing approaches, no exponential blowup is produced by the DBCO. Based on the experiments with a number of real-world data sets of 310 users in three study sites, we demonstrate that topic extraction in the proposed approach is efficient, especially for large-scale web learning resources. According to the user ratings data of four learning sites in the 150 days, the average rate of increase of user rating after the system is used reaches 25.18%.**

Index Terms—**Topic Extraction, E-learning, Semantic, Ontology**

I. INTRODUCTION

With the rapid development of the Internet, there are more and more course information published, which brings in the coexistence problems of "information overload" and "knowledge poor". It costs much time and manpower to manually extracting topic information from unstructured text or multimedia with existing methods. How to implement topic auto-extract in unstructured learning resources, so that users quickly and accurately obtain the knowledge they want, is the key issues that the next generation e-learning theory and technology should focus on. The paper discusses the existing topic auto-extraction technologies, through which, large amount of course information is presented to user concisely and accurately. In order to eliminate redundant information in curriculum, this paper proposes a clustering algorithm-based topic auto-extraction knowledge model. This model applies information fusion technology into the extraction process of the contents of

topics. This paper designs an improved OPTICS algorithm (NOP)-based Multi-document automatic summarization system, which contributes in two aspects: (1) improving the sentence similarity computing method, computing sentences semantic similarity according to the semantic relations between among words in sentence, and based on this clustering sentence, complete the division of the sub-themes, and finally extract a number of sentences in various sub-themes as a summary sentence with certain strategy, extract a certain number of sentences as topic description sentence. colleting the theme concept rather than the word form, using semantic resources WordNet for word sense disambiguation of words in the learning resources documents, and then extracting the theme concept to build vector space model for topic extraction; (2) improving the original OPTICS algorithm, mapping the data space of the original method to a vector space of sentence, and gives an approach of repositioning sparse nodes. Proposing a density-based clustering algorithm, applying it into web course systems; It divides multi-document collections into different sentence clusters; then extracts a certain number of sentences from different sub-themes to produce the digest.

Definition1. Given a specified norm topic *T* is defined as a triple: *T={T_id, T_contenet, T_tag}*.

For example, we can extract the *T* of "software" from the segment of document such as "*Software is a set of items or objects that form a "configuration" that includes programs, documents and data*". According to that, the *T(software)= {T_id(software), T_contenet(software), T_tag(software)}*. *The T_id(software)* is given by the system; the *T_contenet(software)* is "*a set of items or objects that form a configuration" that includes programs, documents and data*", that will be extracted from the document; *T_tag(software)* is "Software=program+ document".

Definition2. **Intelligent Topic Extraction on Web** is defined as a process to get the content of topic from the web resource automatically.

For example, we can use the proposed model to extract

the content of topic from the primitive documents on web and store it in the triple: **T={T_id, T_contenet, T_tag}.**
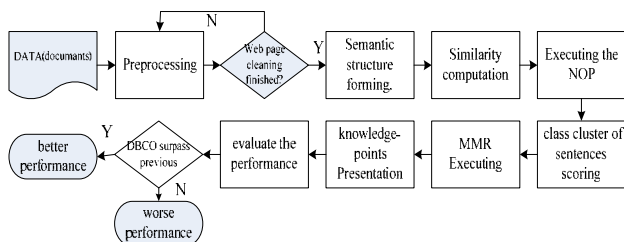

Figure 1. The scheme of topic extraction

II. RELATED WORK

Kim et al. used a special word sense disambiguation technique [1]. In this method, only the 25 most original meanings (Root sense) in WordNet are considered, each word is assigned with a meaning so as to ensure the accuracy of disambiguation. Although they did not give the accuracy of disambiguation, the experimental results on TREC7 and TREC8 data showed over 10% of the increase in the disambiguation performance. Their attempt to add semantic information in the BM25 formula also has been successful. Liu et al. tried to eliminate the ambiguity of the query from an opposite viewpoint [2]. The disambiguation experiment is based on 250 queries evaluated by TREC13Robust, the results show that, Liu's disambiguation method can disambiguate all the 333 ambiguous words in queries, the accuracy was 90%, and the retrieve results is increased by 10% to 25% in five TREC data sets[3]. Michael Ankerst proposed the OPTICS algorithm with the characteristic of density-based clustering across data sets. This algorithm enables digging data of corpus structure [4]. Since the cosine similarity is more concerned about the consistency of the vector direction, which is the characteristic of consistency, the word feature vectors and the cosine similarity are more suitable for calculation of the sentence similarity. We can choose which features the word feature vectors and meaning to both the analysis and calculation, the original method and vocabulary words from the vector were replaced with the concept of distance vector and concepts. Newsblaster is a system developed by Columbia University in multi-document summarization [5], which is a news tracking tool to make daily major news-related abstracts. NeATS [6] is another system of multi-document summarization technology developed by University of Southern California. This system extracts important concepts through reliable statistical information, according to the information of the beginning word in sentences and location information of sentences. Because some sentences begin from a conjunction or verb phrase, if those sentences are extracted in abstract sentence, then the inconsistency will be weaken, so it is needed to filter out these sentences by using the MMR method and reasonable sort of the abstract sentence (such as in chronological order) in order to produce abstracts. Compared with the multi-document summarization technology, the extraction of topics put more requirements on the themes

concept clustering. Research such as [7] targeted on specific content extracting, which integrated knowledge extraction and knowledge mining to extract knowledge from text, and then discover knowledge by association rules. Reference [8] determined directive words from the symbol level. Its extraction objects, however, must be structured and do not meet the requirements of the topic extraction. Paper [9] proposed a reference point and density-based fast clustering algorithm, but it is not appropriate to the dimension reduction of high-dimensional document vector space. Zeng et al. applied the OPTICS algorithm into text clustering [10]. Zhao proposed a similarity measure approach based on different characteristics of sentences [11]. However, the multi-theme documents are major part in multi-document collection, if the sentence is taken by traditional methods; it is likely to ignore certain information in the document collection. There are multiple documents in similar classes, in a multi-document collection, different authors descript the same knowledge in different angles sometimes, or even in the opposite angles, which makes more than one themes appears in multi-document collection. Documents in multi-document collection associate within each other through a common theme, which is taken as central theme. Sub-theme of multi-document collection is the sentences combination with same meaning. these sub-themes present various local information of the document collection. Therefore sub-theme division will be a specific issue in multi-document summarization.

III. THE FRAMEWORK OF INTELLIGENT TOPIC EXTRACTION IN WEB LEARNING RESOURCES

Aiming at providing solutions for the important problems of heterogeneous, discrete, "information overload" and "knowledge poverty" in Web learning resources, we propose a topic auto-extraction method and construct topic auto-extraction system for Web learning resources. By improving and optimizing density-based clustering OPTICS algorithm, we propose the NOP algorithm by mining the internal relations in document collection, divide the document theme more accurately, and improve the topic extraction results. This paper designs a sentence clustering-based topic auto-extraction system by improving the density-based OPTICS algorithm. The system cluster by computing the similarity of sentences, together sentences with the same theme so that each class is represented as a sub-theme document collection in learning resources, and then extract certain number of sentences from each sub-theme to generate description sentence of topic content, finally presents topic contents to the learners. The system's global architecture is shown in Fig. 1.
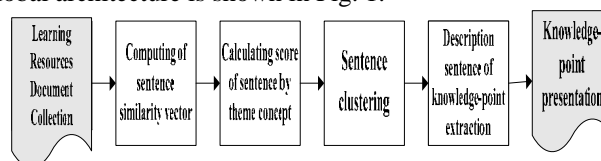

Figure 2.The global structure of topic extraction

Problem Definition

1. There is ambiguity in topic descriptions and topic information inputs of learning resources. To solve the problems of topic expression ambiguity by WordNet description of semantic structure, merging layer by layer, realize semantic disambiguation.

2. The OPTICS algorithm for the data space modeling is applied to the document space model, and the traditional word form-based similarity computing cannot discover the deep information among words. By analyze and calculate both the word characteristic vector and the word meaning vector, we can develop a new approach to replace the measure of word vector and words distance, which are used in the traditional extraction. We need to solve the problem of feasibility and accuracy of Topic extraction, and provide a semantic-level data mining method to avoid the high cost and dispersion of manual extraction.

The steps of topic extraction in learning resources are as following.

Step1.Preprocessing. For example, to accomplish web page cleaning in web learning resources, the resources should be stored as a text document.

Step2.Web page cleaning. The sentences of the of text documents are filtered.

Step3.Semantic structure forming. The semantic structures are described by WordNet, merged layer by layer, to complete semantic disambiguation.

Step4. Similarity computation. This work is done through the concept characteristic vector and concept distance vector among sentences.

Step5. Executing the NOP (an improved OPTIC algorithm) to clustering.

Step6. Scoring the class cluster of sentences.

Step7. Executing the MMR (Maximal Marginal Relevance) method, to choose the sentences of describing topic content, that have high relevance degree to themes while the redundancy among the sentence and other chosen ones is as small as possible.

Step8. Presentation of the content of topics.

Step9. Reference to Edmundson evaluation method, designing two Category of experiments, which evaluate the performance of topic extraction system from subjective, and objectively aspects. The global technology roadmap is shown in Fig. 2.
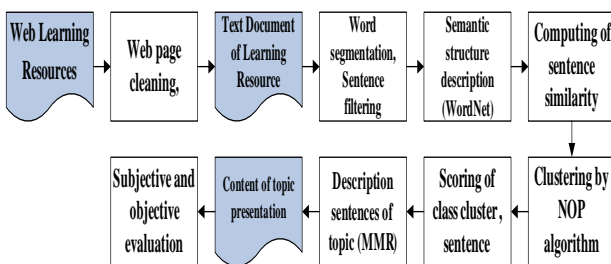
Figure 3. The global technology roadmap

IV. ALGORITHM AND COMPUTING

Notation:

R_u user's query; T topic; $C( )$ web page cleaning; R web learning resources; D text document; $F( )$ sentence filtering; $S( )$ Semantic structure described by WordNet; $M( )$ Merging layer by layer; $Disa( )$ semantic disambiguation; Sen_{CFT} the sentences after the procedures of web page cleaning, sentence filtering and semantic disambiguation, which is stored as text documents; $Sim( )$ similarity computation; $V_{cc}( )$ concept characteristic vector; $V_{cd}( )$ concept distance vector; $NOP( )$ the NOP clustering(improved OPTIC algorithm); Clu_c class cluster; $Scor( )$ Scoring; $Rank( )$ Ranking; Sen_R the sentences after the procedures of Ranking; $Rel_{MMR}( )$ relevance degree computed by the MMR (Maximal Marginal Relevance) method; $Red_{MMR}( )$ redundancy degree computed by the MMR (Maximal Marginal Relevance) method; T_{Rel} threshold of relevance degree to themes; T_{Red} threshold of redundancy degree to themes; Sen_d the chosen sentences of describing the topic content; $P( )$ the content; $Eval_s( )$ a procedure to evaluate the performance of the topic extraction system from the subjective aspect; $Eval_o( )$ a procedure to evaluate the performance of topic extraction system from the objective aspect.

Algorithm1. The Global Algorithm of Topic Extraction

Input: web learning resources R
Output: topic present $P(T)$
1. Begin
2. While(input(R)==True|| PP!=False){
3. $Sen_{CFT} \leftarrow Disa(M(S(F(C(R)))))$
4. $NOP \leftarrow Sim(V_{cc}(Sen_{CFT}), V_{cd}(Sen_{CFT}))$
5. $Clu_c \leftarrow NOP(Sim(V_{cc}(Sen_{CFT}), V_{cd}(Sen_{CFT})))$
6. $Sen_R \leftarrow Rank(top\ i,\ Scor(Clu_c),\ Scor(Sen_{CFT}))$, $1 \leq i \leq n$
7. $Sen_d \leftarrow \forall\ Sen_R\{Sen_R\ |$ $Rel_{MMR}(Sen_R) \geq T_{Rel} \wedge Red_{MMR}(Sen_R) \leq T_{Red}\}$
8. $P(T) \leftarrow P(Sen_d)$ }
9. End

Parameter:

C_i theme concept of a sentence; V vector space; V_1 vector of sentence S_1; ω_i the numbers that the theme concept C_i occurrences in S_1; V_2 vector of sentence S_2; φ_i the numbers that the theme concept C_i occurrences in S_2; $similarity( )$ the similarity of word characteristics by word meaning distance; $Similarity_d( )$ the concept of distance vector similarity of two sentences S_1 and S_2; $\{X_1, X_2, \ldots, X_i\}$, $(1 \leq i \leq m)$ the concept set of sentences S_1; $\{Y_1, Y_2, \ldots, Y_j\}$, $(1 \leq j \leq n)$ the concept set of sentences of S_2; $Distance(X_i, Y_j)$ the distance between the concepts X_i and Y_j; $SIM( )$ sentence similarity computation; $\alpha=0.7$ the coefficient of the concept characteristic vector; $\beta=0.3$ that of the concept distance vector;

Algorithm 2. Sentence similarity computing

Input：Sentence S_1, S_2

Output：$SIM(S_1, S_2)$ similarity of S_1 and S_2

1. Begin

2. $Similarity(S_1, S_2) = \overline{V_1} \cdot \overline{V_2} = \sum_{i=1}^{n} \omega_i \phi_i \left/ \sqrt{\sum_{i=1}^{n} \omega_i^2} * \sqrt{\sum_{i=1}^{n} \phi_i^2} \right.$

3. $Similarity_d(S_1, S_2) = 1 \left/ \sum_{i=1}^{m} \sum_{j=1}^{n} Dis\tan ce(X_i, Y_j) \right.$

4.

$SIM(S_1, S_2) = \alpha * Similarit(S_1, S_2) + \beta * Similarity_d(S_1, S_2)$

5. End

In the original DBCO algorithm, the ordered queue is always sorted in ascending order according to reachable distances. Therefore, the algorithm always selects to deal with the points with the smallest distance.

Parameter：

V a threshold value, when sentences in the class cluster is greater than or equal to V, the class cluster is taken as valid cluster; otherwise as invalid cluster. V is notated as $0.8 * \text{Avg}_{stNum}(C)$; $NUM_v(Sen)$ the number of valid sentences in C_i; $NUM_{vc}(Sen)$ the number of valid sentences in all class clusters ; C_i the ith sub-theme; $Num_i(sen)$ the number of sentences in the class; $Num_i(doc)$ the number of the original documents whose sentences belong to the sub-theme; $Num(sen)$ the total number of sentences in the original document; $Num(doc)$ the number of documents contained in the original document; $SCORE(\ )$ the class cluster score; $Len(\ )$ length; $Rank_{rel}(\ )$ ranking of the relevance degree of the sub-themes and the central theme; Sen sentence; C sub-theme; Sen_d the chosen sentences of describing topic content; $Len_{inx}(T)$ the minimum length requirements of topic description summery; and $P(\ )$ the content.

A. Experiment

In this section we describe the experiments that we carried out on real-world web learning resources data sets. We chose these sets because they are publicly available. We take Edmundson as the evaluation method; design 2 types of experiments to evaluate the performance of our topic extraction system: (1) Subjective evaluation: In order to the subjective evaluation, we used the method of artificial extraction (by the domain experts and authorized teachers jointly) and the method of auto-extraction to extract the content of topics inspective in two courses"the interface technology of computer" and "the architecture of computer"(http://grid.whu.edu.cn/). The domain experts compare the extracted information with the two methods, and then give a review score. The rating levels are as follows: totally not similar, similar, very similar, completely similar to the other. (2) Objective evaluation: in the experimental evaluation, six tasks are tested, including DUC2005 and ROUGE 1.5.5 data sets, select

ROUGE-N (where N take 1 to 4), ROUGE-L and ROUGE-W-1.2. There are voting systems in our topic extraction System, which are designed to collect the users' feedback. The users' review is presented in scores (0-100). If the user thinks the platform is really helpful for them in topic extraction, they evaluate it with a high score. We collect the users' remarks in 3 learning web sites in 150 days, and compare the average evaluation scores, before and after the system are used. The results are showed in Table 1.

In order to evaluate the performance of the topic extraction system, we record 8 experts' reviews on the topic extraction performance of 4 learning sites. The experiment result is showed in Table 2. As showed in Fig.4, different experts present their evaluations on the performance of our topic extraction system in the 3 learning sites. In general, the reviews on site 2 are better than those on other sites. However, the experts' reviews on site 3 are consistent with those on others.

B. Evaluation

As shown in Fig. 4,5,6,7,8,9,10,11, the scores of the 3 sites after using our system are better than those before using the proposed platform. The user ratings of the four learning sites are raised compared to the previous system before using the topic extraction system. The maximum increase is 36% in site 3; while the minimum increase is 17.91% in site 2; the average increase rate of user ratings after the system is used reaches to 25.18%. The result for the real-world data sets verifies the effectiveness of the proposed method. However, we notice that the improvement of the users' reviews in the experiment may be affected by other factors. For example, site 3 is the worst one in users' reviews before using the proposed approach. The suddenly increase of the scores might be partly from the contribution of the advertisement promotion during the days experiments are carried out. At least the results of 2 sites in our experiment clearly demonstrate that the proposed topic extraction system significantly outperforms the previous one. The results for the real-world data sets support this conclusion. All the dataset can be downloaded in the webpage http://grid.whu.edu.cn/rainbow/

We can see from Fig 12,13,14,15, there are some differences among the experts' reviews on the topic extraction performance, which illustrate that the stability of the topic extraction is still needed to be enforced. On the other hand, more objective work is needed in the future.

TABLE 1. THE AVERAGE EVALUATION SCORES OF 3 LEARNING SITES, (S1:SITE1, THRESHOLD VALUE $V=0.48$)

	S1	S2	S3
Pre	63	67	71
Cur	76	79	84

TABLE 2. THE AVERAGE EVALUATION SCORES OF 3 LEARNING SITES, (S1:SITE1, THRESHOLD VALUE $V=0.52$)

	S1	S2	S3
Pre	60	73	79
Cur	78	82	86

TABLE 3. THE AVERAGE EVALUATION SCORES OF 4 LEARNING SITES,
(S1:SITE1, THRESHOLD VALUE *V=0.56*)

	S1	S2	S3
Pre	76	74	82
Cur	84	87	90

TABLE 4. THE AVERAGE EVALUATION SCORES OF 4 LEARNING SITES,
(S1:SITE1, THRESHOLD VALUE *V=0.6*)

	S1	S2	S3
Pre	72	79	67
Cur	81	87	84

TABLE 5. THE AVERAGE EVALUATION SCORES OF 4 LEARNING SITES,
(S1:SITE1, THRESHOLD VALUE *V=0.64*)

	S1	S2	S3
Pre	75	78	70
Cur	80	83	82

TABLE 6. THE AVERAGE EVALUATION SCORES OF 4 LEARNING SITES,
(S1:SITE1, THRESHOLD VALUE *V=0.68*)

	S1	S2	S3
Pre	68	71	74
Cur	75	78	80

TABLE 7. THE AVERAGE EVALUATION SCORES OF 4 LEARNING SITES,
(S1:SITE1, THRESHOLD VALUE *V=0. 72*)

	S1	S2	S3
Pre	70	72	69
Cur	74	78	80

TABLE 8. THE AVERAGE EVALUATION SCORES OF 4 LEARNING SITES,
(S1:SITE1, THRESHOLD VALUE *V=0.76*)

	S1	S2	S3
Pre	78	76	75
Cur	83	85	81

TABLE 9. EXPERTS REVIEWS ON THE TOPICS EXTRACTION
(S1:SITE1, E1:EXPERT1, *V=0.48*)

	E1	E2	E3	E4	E5	E6	E7	E8
S1	76	74	68	73	72	71	67	68
S1'	82	76	71	79	77	80	75	81
S2	75	77	74	76	72	70	71	69
S2'	83	89	90	78	79	82	83	78
S3	74	73	72	75	73	72	68	70
S3'	84	86	85	88	80	81	79	76

TABLE 10. EXPERTS REVIEWS ON THE TOPICS EXTRACTION
(S1:SITE1, E1:EXPERT1, *V=0.56*)

	E1	E2	E3	E4	E5	E6	E7	E8
S1	65	67	69	71	74	72	64	68
S1'	75	78	73	79	76	84	78	77
S2	67	70	71	72	68	66	69	73
S2'	78	82	83	77	75	80	81	77
S3	70	68	66	72	76	71	67	65
S3'	81	79	78	82	84	85	80	74

TABLE 11. EXPERTS REVIEWS ON THE TOPICS EXTRACTION
(S1:SITE1, E1:EXPERT1, *V=0.64*)

	E1	E2	E3	E4	E5	E6	E7	E8
S1	72	74	67	70	73	68	69	71
S1'	78	76	71	77	79	81	80	76
S2	70	78	75	76	71	75	73	74
S2'	81	82	87	84	80	79	80	81
S3	76	80	74	73	77	75	74	72
S3'	84	82	83	78	80	82	84	76

TABLE 12. EXPERTS REVIEWS ON THE TOPICS EXTRACTION
(S1:SITE1, E1:EXPERT1, *V=0.72*)

	E1	E2	E3	E4	E5	E6	E7	E8
S1	68	67	69	74	70	66	73	71
S1'	74	73	76	82	77	67	78	75
S2	72	74	78	75	71	73	76	77
S2'	80	78	83	82	85	80	77	79
S3	70	72	75	78	71	69	74	68
S3'	73	78	79	82	76	79	77	73
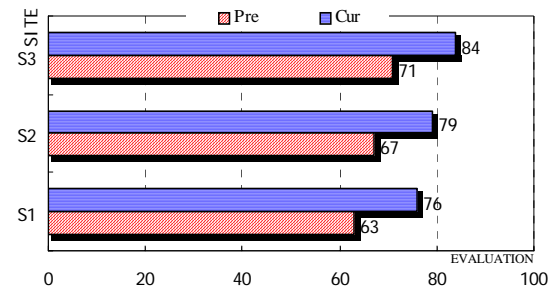
Figure4. Average evaluation scores of 3 learning sites (*V=0.48*)
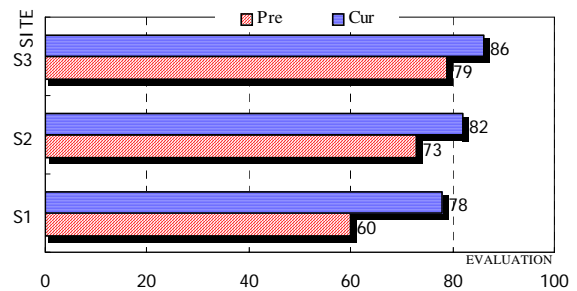
Figure5. Average evaluation scores of 3 learning sites (*V=0.52*)
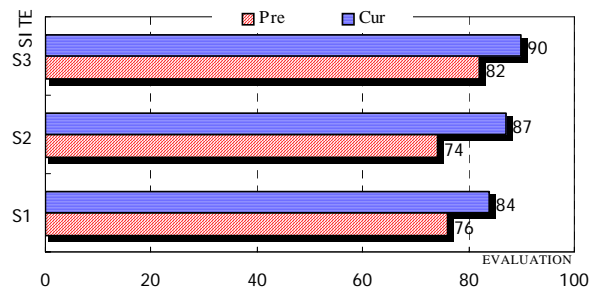
Figure6. Average evaluation scores of 3 learning sites (*V=0.56*)
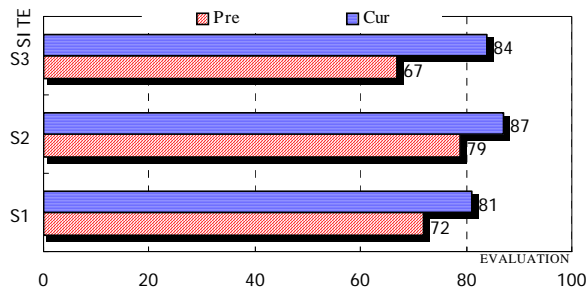
Figure7. Average evaluation scores of 3 learning sites (*V=0.6*)
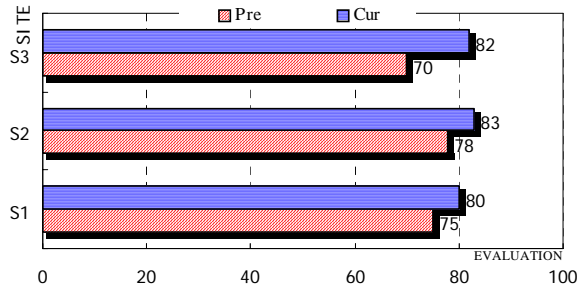
Figure7. Average evaluation scores of 3 learning sites (*V=0.64*)
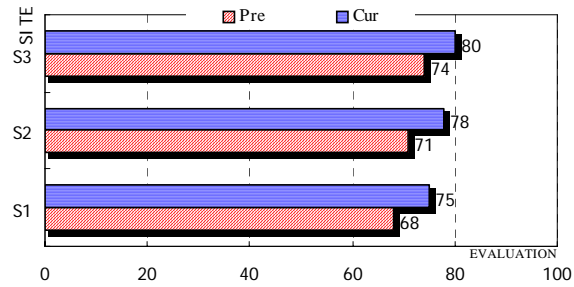
Figure8. Average evaluation scores of 3 learning sites (*V=0.68*)
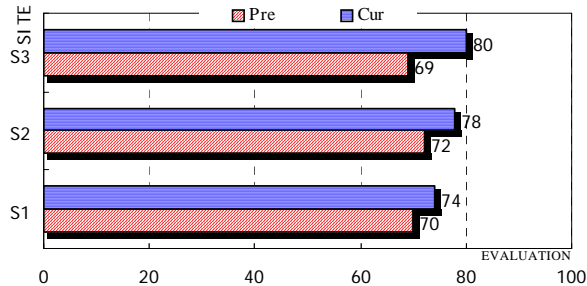
Figure9. Average evaluation scores of 3 learning sites (*V=0.72*)
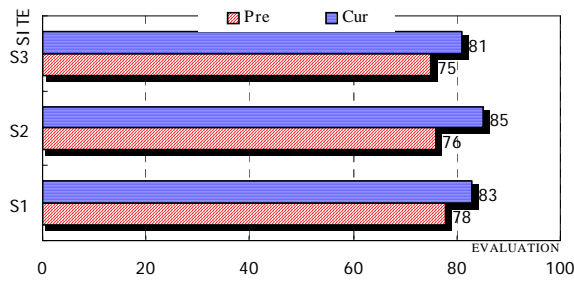
Figure10. Average evaluation scores of 3 learning sites (*V=0.76*)
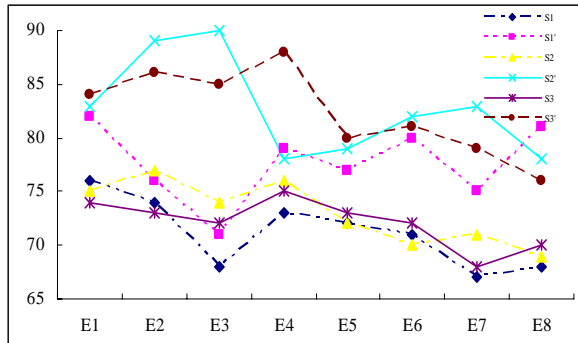
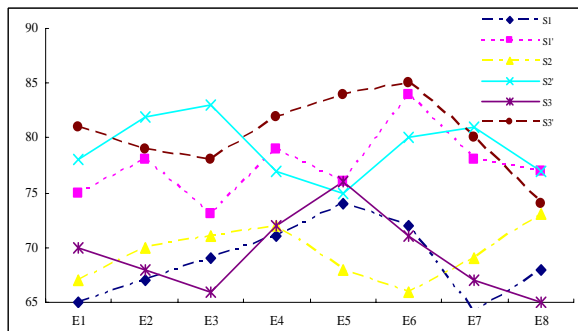Figure11. Experts reviews on the topic extraction (*V=0.48*)

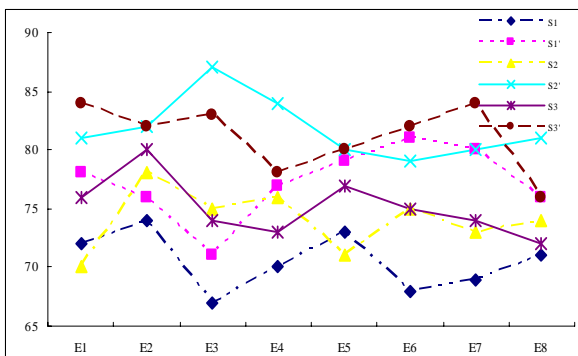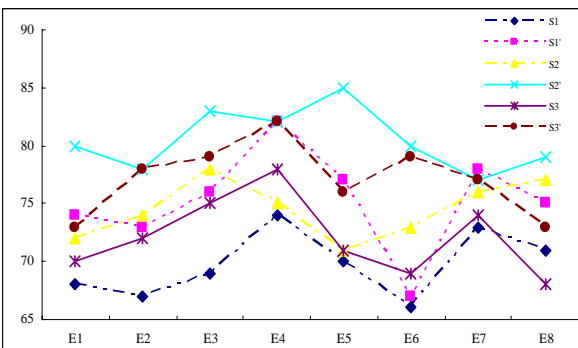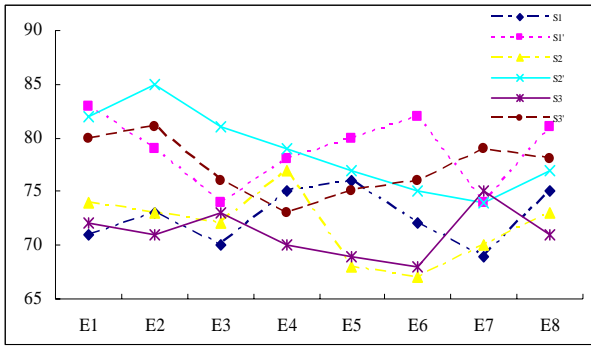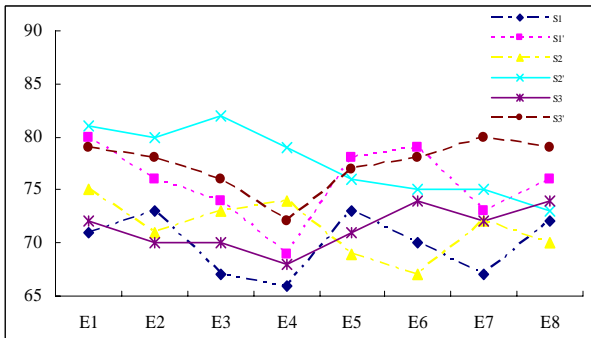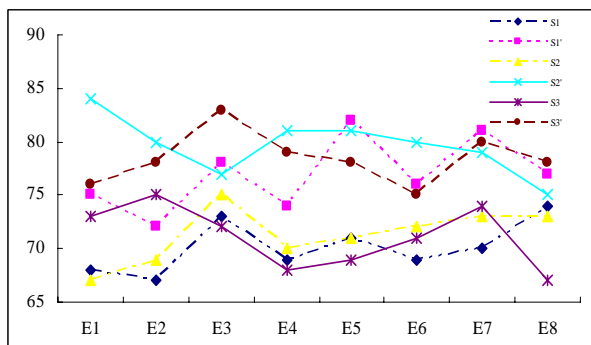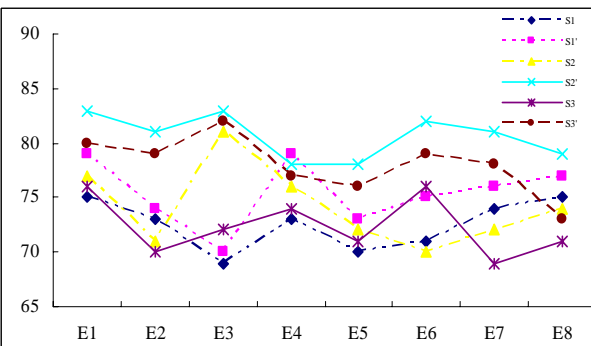Figure12. Experts reviews on the topic extraction (*V=0.56*)

Figure13. Experts reviews on the topic extraction (*V=0.64*)

Figure14. Experts reviews on the topic extraction (*V=0.72*)

Figure15. Experts reviews on the topic extraction (*V=0.64*)

Figure16. Experts reviews on the topic extraction (*V=0.68*)

Figure17. Experts reviews on the topic extraction (*V=0.72*)

Figure18. Experts reviews on the topic extraction (*V=0.76*)

V. CONCLUSIONS AND FUTURE WORK

The project "Computer System and Interface Technology" funded by the National High Technology

Research and Development Program ("863"Program) of China" Research and Development of E-learning Platforms and Education Resource", has carried out the artificial extraction and semantic mark of topic. On this basis, the study of topic extraction method is a feasible new idea. Facing with massive learning resources, we apply the improved OPTIC-based NOP algorithm (DBCO), computing the correlation between topic and the documents in learning resources. The method has been verified to be effective in plain text information retrieval and data mining. We believe that it will also be feasible in multimedia learning resources after being preprocessed. By studying the effective algorithm of topic extraction, and analyzing the limitations of complexity of time and space with a large number of experiments, we verify that the algorithm is effective and it can meet customers' requirements on efficiency and accuracy of topic presentation. This paper's contribution includes: (1) colleting the theme concept rather than the word form, using semantic resources WordNet for word sense disambiguation of words in the learning resources documents, and then extracting the theme concept to build vector space model for topic extraction; (2) proposing a density-based clustering algorithm, applying it into web course systems; It divides multi-document collections into different sentence clusters; then extracts a certain number of sentences from different sub-themes to produce the digest; and (3) sorting the output to generate topic content.

This work could be extended in several directions. Our data sets and experiments are derived from real-world resources. Although the proposed approach requires training data, it may not be necessary to retrain the model frequently. As long as the web learning resources used to compute the class clustering reflect the changes in the sub-themes' relevance to a topic, this method is effective. Future work might include empirically evaluating the robustness of large–scale learning resources derived in a dynamic environment; and binding the user information protection with user's trust links into the topic extraction model.

REFERENCES

[1] S.B. Kim , H. C. Seo and H. C. Rim. Information Retrieval using Word Senses: Root Sense Tagging Approach [A]. In : Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. ACM Press, 2004. 258-265

[2] S. Liu, F. Liu, C. Yu and W. Meng. An Effective

Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases[A]. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval[C]. ACM Press, 2004. 266-227

[3] S. Liu, C. Yu and W. Meng. WordSense Disambiguation in Queries [A]. In: Proceed2ings of the 14th ACM International Conference on Information and Knowledge Management[C]. ACM Press, 2005. 525-532

[4] M. Ankerst, M. M. Breunig, H. P. Kriegel, J. Sander. OPTICS: ordering points to identify the clustering structure. International Conference on Management of Data [A]. In: Proceedings of the 1999 ACM SIGMOD international conference on Management of data[C]. ACM Press,1999. 49-60

[5] R. Barzilay, N. Elhadad, K. McKeown, Inferring Strategies for Sentence Ordering in Multidocument News Summarization. in Journal of Artifical Intelligence Research (JAIR) [J], 2002, Vol. 17, pp 35-55

[6] C. Y. Lin, C.Y. Lin and E. Hovy. Automated Multi-document Summarization in NeATS. In Proceedings of the Human Language Technology Conference[C]. 2002

[7] Y. Li, Q. Zhong, J. Li, J. Tang, Result of ontology alignment with RiMOM at OAEI'07. In Proc. of International Workshop of Ontology Matching on the 6th International Semantic Web Conference[C].

[8] W. Hu, Y. Zhao, Y. Qu, Partition-based Block Matching of Large Class Hierarchies[C] Proc. of the 1st Asian Semantic Web Conference. Beijing, China, 2006: 72-83.

Ming Xie

Hunan, China
Ph.D. in Computer Science
Computer School of Wuhan University,
Wuhan, Hubei, China
430072

Email: wolfgangtse@gmail.com

Chanle Wu

Hubei, China
Professor of Computer Science
Computer School of Wuhan University,
Wuhan, Hubei, China
430072

Email: wuchle @whu.edu.cn

Web Page Classification Using Relational Learning Algorithm and Unlabeled Data

Yanjuan Li[1,2]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]School of Information and Computer Engineering, North-East Forestry University, Harbin, China
Email: lyjuan5@gmail.com

Maozu Guo
School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
Email: maozuguo@hit.edu.cn

Abstract—**Applying relational tri-training (R-tri-training for short) to web page classification is investigated in this paper. R-tri-training, as a new relational semi-supervised learning algorithm, is well suitable for learning in web page classification. The semi-supervised component of R-tri-training allows it to exploit unlabeled web pages to enhance the learning performance effectively. In addition, the relational component of R-tri-training is able to describe how the neighboring web pages are related to each other by hyperlinks. Experiments on Web-Kb dataset show that: 1) a large amount of unlabeled web pages (the unlabeled data) can be used by R-tri-training to enhance the performance of the learned hypothesis; 2) the performance of R-tri-training is better than the other algorithms compared with it.**

Index Terms—**web page classification, relational tri-training, relational learning, tri-training, co-training**

I. INTRODUCTION

In recent years there has been a great deal of interest in applying machine-learning methods to a variety of problems in classifying and extracting information from text [1]. In large part, this trend is sparked by the explosive growth of the World Wide Web. An interesting aspect of the Web is that it can be thought of as a graph in which pages are the nodes of the graph and hyperlinks are the edges. The graph structure of the Web makes it an interesting domain for relational learning [2]. Moreover, Craven, Slattery, and Nigam demonstrated that for several Web-based learning tasks, a relational learning algorithm can learn more accurate classifiers than a common statistical approach [3]. Therefore, many researchers have been done to apply relational learning algorithms to web page classification.

Relaxation labeling [4] is one of the algorithms that work well in web page classification. Relaxation labeling was originally proposed as a procedure in image analysis by Rosenfeld, Hummel, and Zucker. Later, it became widely used in image and vision analysis, artificial intelligence, pattern recognition, and web mining [5]. In the context of hypertext classification, the relaxation labeling algorithm employs a text classifier to assign class probabilities to each node (page) firstly. Then each

page is considered in turn and its class probabilities are reevaluated in light of the latest estimates of the class probabilities of its neighbors.

Based on a new framework for modeling link distribution through link statistics, Lu and Getoor [6] proposed a variation of relaxation labeling, in which a combined logistic classifier is used based on content and link information. This approach not only showed improvement over a textual classifier, but also outperformed a single flat classifier based on both content and link features. Angelova and Weikum [7] proposed another variation of relaxation labeling, not all neighbors are used, that is, only neighbors that are similar enough in content are considered.

Besides relaxation labeling, other relational learning algorithms can also be applied to web classification. Loopy belief propagation and iterative classification are compared and analyzed with relaxation labeling by Sen and Getoor [8]. Their performance on a web collection is better than textual classifiers. A toolkit for classifying networked data is implemented by Macskassy and Provost [9], it utilized a relational classifier and a collective inference procedure [10], and its powerful performance is demonstrated on several datasets including web collections.

Although these relational learning algorithms have been successfully applied to web page classification, they need large number of labeled web pages acting as training set in order to guarantee generalization ability. However, labeled web pages are fairly expensive to obtain because they require artificial marking. In contrast, unlabeled web pages are readily available. Therefore, in relational learning domain, how to exploit unlabeled web pages to enhance the performance of the learned hypothesis is an urgent challenge.

In this paper, we apply R-tri-training algorithm, which is proposed in our previous work [11], to web page classification. R-tri-training, as a relational semi-supervised learning algorithm, combines semi-supervised learning and relational learning. Three different relational learning systems are initialized according to the labeled data and the background knowledge, and then the three classifiers are refined by iterating using the unlabeled

data under certain condition. Experiments on Web-Kb using R-tri-training show its outstanding performance.

The remainder of the paper is organized as follows. Section 2 gives the brief description of relational tri-training. In section 2, the selection of base classifier and the sufficient condition for re-training is introduced. Sections 3 discuss the experiment. In section3, the dataset used in this paper is given firstly, and then the effect of unlabeled web pages is discussed, finally the comparison with other systems is proposed. Conclusion is given in section 4.

II. RELATIONAL TRI-TRAINING

Relational learning is a research area which investigates the inductive construction of first-order clausal theories from examples and background knowledge. By using first-order logic as the representational mechanism for hypotheses and examples, relational learning can overcome the two main limitations of classical machine learning techniques, 1) the use of a limited knowledge representation formalism(essentially a propositional logic), and 2) difficulties in using substantial background knowledge in the learning process [12]. However, current relational learning algorithm such as Nfoil [13], Kfoil [14] and Aleph [15] are all supervised learning algorithms. Therefore, they need a large number of labeled examples acting as training set in order to guarantee generalization ability.

Existing semi-supervised learning algorithms attempt to exploit the additional information provided by the large amount of unlabeled data to guide the learning process, and enhance the final performance. A prominent method known as tri-training was proposed by Zhou and LI [16]. However, current semi-supervised learning algorithm belongs to classical machine learning, use propositional logic to represent training examples, therefore, it cannot overcome the two main limitations of classical machine learning techniques mentioned above.

Since relational learning and semi-supervised learning have their own pros and cons, it motivates the integration of the two approaches. Relational tri-training [11] is a framework for combining the power of enhancing the learning performance by exploiting the unlabeled examples of tri-training and the power of knowledge representation of relational learning. This algorithm combined the tri-training based on propositional logic representation and relational learning algorithm based on first-order logic representation. Three different relational learning systems (base classifiers) are initialized according to the labeled data and the background knowledge, and then the three classifiers are refined by iterating using the unlabeled data. That is, under special condition, for one unlabeled data, it is going to be labeled to one classifier as the new training data when the same labeled result are given by the other two classifiers. The final hypothesis (the learned model) is produced via majority voting of the three base classifiers.

A. The selection of base classifier

Relational tri-training algorithm employs Nfoil, Kfoil and Aleph as three different base classifiers. Nfoil was the first system in literature to tightly integrate feature construction and Naive Bayes. Such a dynamic propositionalization was shown to be superior compared to static propositionalization approaches that employ Naive Bayes only for post-processing the rule sets. Kfoil system integrates FOIL with kernel methods. The feature space is constructed by leveraging FOIL search for a set of relevant clauses. The search is driven by the performance obtained by a support vector machine based on the resulting kernel. KFOIL can naturally handle classification and regression tasks. ALEPH is a publicly available relational learning system implemented in a Prolog compiler called Yap (version 4.3.22) that is a generalization of PROGOL [17].

The reason for selecting Nfoil, Kfoil and Aleph as the base classifiers is that the three relational learning algorithms have different search bias. Nfoil and Kfoil belong to top-down relational learning algorithm that learns program clauses by searching a space of possible clauses from general to specific. The learned final classifier is the most general hypothesis and is prone to classify the unlabeled data into positive. Aleph belongs to bottom-up relational learning algorithm that searches for program clauses by starting with very specific clauses and attempting to generalize them. The learned final classifier is the most specific hypothesis and is prone to classify the unlabeled data into negative.

B. The sufficient condition for re-training

Let L denote the initial labeled example set, and U denote the unlabeled example set. Background knowledge is denoted by B. Nfoil, Kfoil, Aleph system is respectively trained from L and B, three different classifiers h_1, h_2, h_3 is obtained. If h_2 and h_3 agree on the labeling of an example x in U, then x can be labeled for h_1. It is obvious that in such a scheme if the prediction of h_2 and h_3 on x is correct, then h_1 will receive a valid new example for further training; otherwise, h_1 will get an example with noisy label. Therefore, in previous work [11], we analyses the PAC theory of concept class $C_{B,d,M}$, then gives the sufficient condition to ensure that the classification accuracy of individual classifier could be improved by re-training on the new training set, and use the sufficient condition as criteria to decide whether the newly labeled example set should be used for re-training. To guarantee the integrity of the article, this paper briefly describes the sufficient condition.

According to the conclusion of Horuath, Sloan and Turan [18], The concept class $C_{B,d,M}$ is polynomially PAC learnable with random classification noise with sample size

$$s = O\left(\frac{1}{\varepsilon^2 (1-2\eta_b)^2} \left((2m|M|)^{2a^{d+1}} \ln\left(\frac{(2m|M|)^{a^{d+1}}}{\delta} \right) + \ln\left(\frac{(2m|M|)^{a^{d+1}}}{(1-2\eta_b)^2 \varepsilon\delta} \right) \right) \right)$$

, where ε is the hypothesis worst-case classification error rate, η (<0.5) is an upper bound on the classification noise rate, δ is the confidence , m is the arity of target

predicate, $|M|$ is the number of modes in M, a is the maximum arity of any predicate in background knowledge, d is depth bound, $C_{B,d,M}$ is concept class with background knowledge B that are represented by Horn clauses of depth at most d satisfying M.

Let

$$s' = \frac{1}{\varepsilon^2(1-2\eta_b)^2}\left((2m|M|)^{2a^{d+1}}\ln\left(\frac{(2m|M|)^{a^{d+1}}}{\delta}\right)+\ln\left(\frac{(2m|M|)^{a^{d+1}}}{(1-2\eta_b)^2\,\varepsilon\delta}\right)\right),$$

$\mu=s/s'$, then

$$s = \mu s' = \frac{\mu}{\varepsilon^2(1-2\eta_b)^2}\left((2m|M|)^{2a^{d+1}}\ln\left(\frac{(2m|M|)^{a^{d+1}}}{\delta}\right)+\ln\left(\frac{(2m|M|)^{a^{d+1}}}{(1-2\eta_b)^2\,\varepsilon\delta}\right)\right)$$

(1)

For a specific concept class $C_{B,d,M}$, m, $|M|$, a, d are constants, δ is confidence and is also a constant.

Let $p = (2m|M|)^{2a^{d+1}}\ln\left(\frac{2m|M|^{a^{d+1}}}{\delta}\right)$, $q = \frac{(2m|M|)^{a^{d+1}}}{\delta}$,

then p, q are constant, Eq. (1) can be rewritten as:

$$s = \frac{\mu}{\varepsilon^2(1-2\eta_b)^2}\left(p+\ln\left(\frac{q}{(1-2\eta_b)^2\varepsilon}\right)\right) \quad (2)$$

In each round of R-tri-training, the classifiers h_2 and h_3 choose some examples in U to label for h_1. Since the classifiers are refined in the R-tri-training process, the amount as well as the concrete unlabeled examples chosen to label may be different in different rounds. Let L_t and L_{t-1} denote the set of examples that are labeled for h_1 in the tth round and the $(t-1)$th round, respectively. Then, the training set for h_1 in the tth round and the $(t-1)$th round are $L\cup L_t$ and $L\cup L_{t-1}$, whose sample size are $|L\cup L_t|$ and $|L\cup L_{t-1}|$, respectively. Let $e_{1,t}$ (<0.5) denote the upper bound of the classification error rate of $h_2\&h_3$ in the tth round, i.e., the error rate of the hypothesis derived from the combination of h_2 and h_3. Thus, the number of examples in L_t that are mislabeled is $e_{1,t}|L_t|$. Therefore, the classification noise rate in the tth round is:

$$\eta_t = \frac{e_{1,t}\,|L_t|}{|L_t|+|L|} \quad (3)$$

Let ε_{t-1}, ε_t denote the classification error of h_1 in $(t-1)$-th iteration and t-th iteration, respectively. Eq. (2) could respectively be reformed as Eq. (4) and Eq. (5) in $(t-1)$-th iteration and t-th iteration.

$$|L\cup L_{t-1}| = \frac{\mu}{\varepsilon_{t-1}^2(1-2\eta_{t-1})^2}\left(p+\ln\left(\frac{q}{(1-2\eta_{t-1})^2\varepsilon_{t-1}}\right)\right)(4)$$

$$|L\cup L_t| = \frac{\mu}{\varepsilon_t^2(1-2\eta_t)^2}\left(p+\ln\left(\frac{q}{(1-2\eta_t)^2\varepsilon_t}\right)\right) \quad (5)$$

With Eq. (4) and Eq. (5), the sufficient condition of $\varepsilon_t<\varepsilon_{t-1}$ can be given

$$|L_t|\geq|L_{t-1}| \text{ and } \eta_t<\eta_{t-1}$$

Considering Eq. (3), we have

$$|L_t|\geq|L_{t-1}| \text{ and } \frac{e_{1,t}\,|L_t|}{|L_t|+|L|} < \frac{e_{1,t-1}\,|L_{t-1}|}{|L_{t-1}|+|L|}$$

It is the sufficient condition for re-training used in relational tri-training.

III. EXPERIMENT

A. Dataset

The Web page dataset, we use for our experiments is the Web-Kb [19] dataset. This data set consists of 4,167 Web pages collected from Web sites of Computer Science departments of four universities: Cornell University, University of Texas, University of Washington, and University of Wisconsin. These pages have been labeled into a number of categories, among which the category student home page is regarded as the target. That is, student home pages (13 percent) are the positive examples and all other pages are negative examples.

B. Effect of unlabeled web page

Experiment settings: we randomly selected 40% of all web pages as test examples to evaluate the performance of the learned hypothesis, while the rest are used as the pool of training examples, i.e., $L\cup U$. In each pool, L and U are partitioned under different unlabeled rates including 80 percent, 60 percent, 40 percent, and 20 percent. For instance, assuming a dataset contains 1,000 examples, 400 examples are selected as test examples, the remaining 600 examples are the pool of training examples. When the unlabeled rate is 80 percent, 120 examples are put into L with their labels while the remaining 480 examples are put into U without their labels. Here, the pos/neg ratio of L, U, and test set are similar to that of the original dataset.

Under each unlabeled rate, three different random partitions of L and U are generated and one independent run is performed for each random partition, then the average classification error rate of three runs is computed and as the final performance evaluation.

Experiment results: The average predictive error rates of the final hypothesis and of each base classifier (Nfoil, Kfoil, Aleph) under different unlabeled rate are respectively shown in table I, II, III, IV, respectively. Where *initial* denotes the performance at round 0, i.e., the performance obtained using only the labeled training examples, *final* denotes the performance when algorithm finished, *improve* denotes the corresponding improvements, and *Hypothesis* denotes the performance of the hypotheses, i.e., the learned models. Table I, II, III, and IV show that on the Web page classification task, relational tri-training can effectively utilize unlabeled data to enhance the learning performance.

TABLE I. CLASSIFICATION ERROR RATES OF THE INITIAL AND FINAL CLASSIFIERS UNDER 80 PERCENT UNLABELED RATE

classifier	Classification error rate on test dataset		
	initial	*final*	*improve*
Nfoil	0.0650	0.0442	32.00%
Kfoil	0.0804	0.0684	14.93%
Aleph	0.0788	0.0729	7.49%
Hypothesis	0.0470	0.0394	16.17%

TABLE II. CLASSIFICATION ERROR RATES OF THE INITIAL AND FINAL CLASSIFIERS UNDER 60 PERCENT UNLABELED RATE

classifier	Classification error rate on test dataset		
	initial	*final*	*improve*
Nfoil	0.0450	0.0402	10.68%
Kfoil	0.0840	0.0716	14.76%
Aleph	0.0792	0.0658	16.92%
Hypothesis	0.0402	0.0342	14.93%

TABLE III. CLASSIFICATION ERROR RATES OF THE INITIAL AND FINAL CLASSIFIERS UNDER 40 PERCENT UNLABELED RATE

classifier	Classification error rate on test dataset		
	initial	*final*	*improve*
Nfoil	0.0428	0.0392	8.41%
Kfoil	0.0632	0.0316	50.00%
Aleph	0.0868	0.0859	0.94%
Hypothesis	0.0356	0.0254	28.65%

TABLE IV. CLASSIFICATION ERROR RATES OF THE INITIAL AND FINAL CLASSIFIERS UNDER 20 PERCENT UNLABELED RATE

classifier	Classification error rate on test dataset		
	initial	*final*	*improve*
Nfoil	0.0579	0.0504	13.66%
Kfoil	0.0759	0.0612	19.37%
Aleph	0.0963	0.0747	22.43%
Hypothesis	0.0449	0.0384	14.48%

Figure 1, 2, 3, and 4 depicts the error rates change of the final hypothesis and three base classifiers (Nfoil, Kfoil, Aleph) along with the re-training iterations. Note that since the relational tri-training algorithm may terminate in different iterations, the error rates at termination are used as the error rates of the iterations after termination.
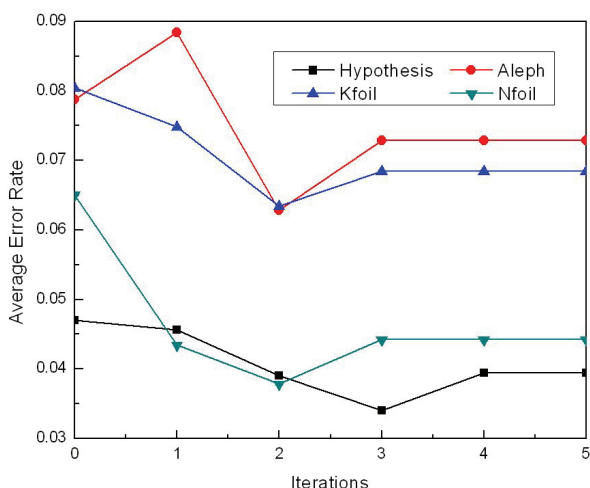
Figure 1. Iterative change of the error rates under 80 percent unlabeled data
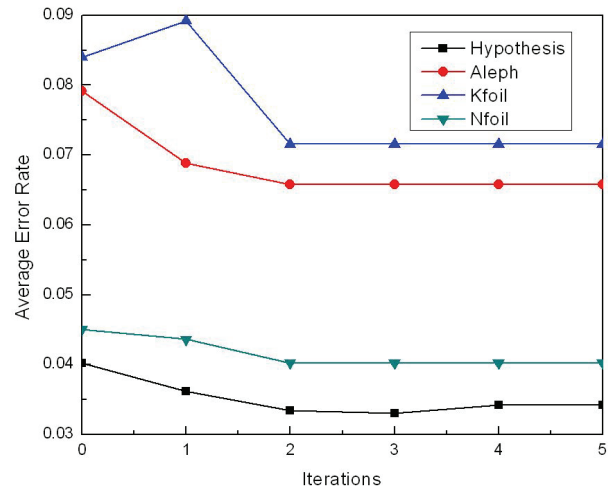
Figure 2. Iterative change of the error rates under 60 percent unlabeled data
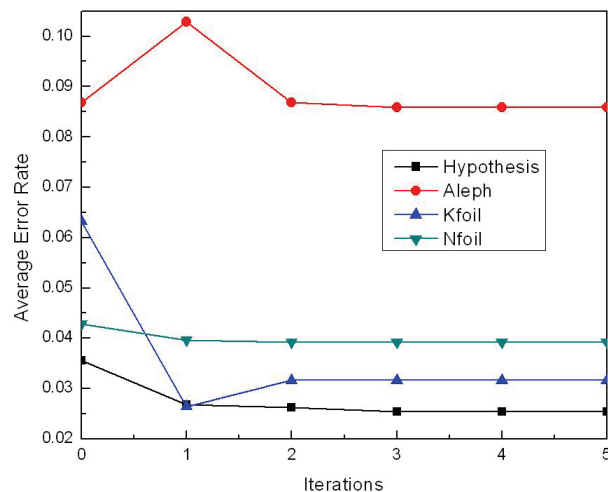
Figure 3. Iterative change of the error rates under 40 percent unlabeled data
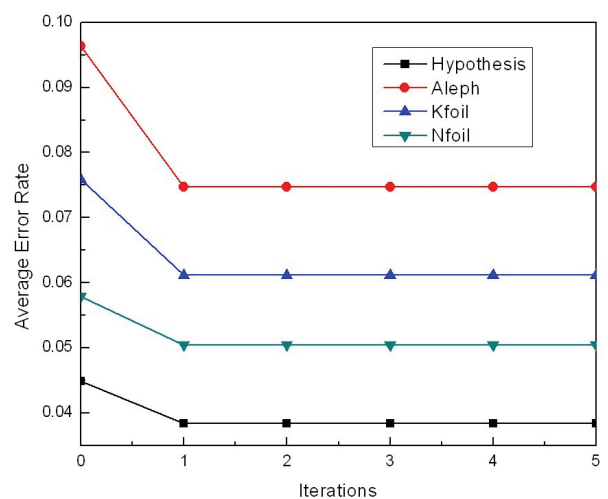
Figure 4. Iterative change of the error rates under 20 percent unlabeled data

Figure 1, 2, 3 and 4 reveal that on all percent unlabeled data, the final hypotheses generated by relational tri-training are better than the initial hypotheses, which

confirms that relational tri-training can effectively exploit unlabeled examples to enhance the learning performance.

The time complexity of relational tri-training can be analyzed from Figure 1, 2, 3 and 4. It shows that the maximum number of iterations under different unlabeled rate does not exceed five times, and after four iterations, the error rate of hypothesis is no longer obviously changed. Therefore, the time complexity of relational tri-training is very low based on the iteration number when the training is over.

C. Comparison with other systems

1) Algorithms used for comparison
Naive Bayes classifier

Naive Bayes algorithm is often applied to text learning problems. Nuanwan and Boonserm [20] test the performance of naive Bayes using two feature sets which are heading and content feature on Web-Kb dataset.

Co-training algorithm

Co-training algorithm was proposed by Blum and Mitchell [21]. Two individual classifiers are trained separately on two different views, i.e. two independent sets of attributes, and then each individual classifier labels some unlabeled ones for the other to augment the training set for re-training. The standard co-training assumes that the attributes are naturally partitioned into two sets, each of which is sufficient for learning and conditionally independent to the other given the class label. When the assumption is satisfied, the cotrained classifiers could make fewer generalization errors by maxmizing their agreement over the unlabeled data [22].

ICT-ILP

ICT (Iterative-Cross Training) [20] is a semi-supervised learning algorithm. It consists of two learners and each learner gets a small amount of labeled data. The strong learner (classifier1) starts the learning process from the labeled data and classifies unlabeled data (TrainingData2). The weak learner (classifier2) uses these newly labeled data to learn and classifies TrainingData1. ICT-ILP is the ICT which combines the inductive logic programming system into one of the classifiers, that is, the Classifier1 of ICT-ILP is the Progol system.

2) Experiment setting and result

To exploit the experiment result in [20], the experimental configuration is the same as that used in it. We randomly selected 30% of all examples to be initial labeled data. The unlabeled training data consists of 30% of all examples and 40% of all examples were used as a test set.

The performance evaluation is done using the standard precision (P), recall (R) and F_1-measure (F_1). These measurements are defined as follows.

$$recall = \frac{\#correct\ positive\ predictions}{\#positive\ examples}$$

$$precision = \frac{\#correct\ positive\ predictions}{\#positive\ predictions}$$

$$F_1 = \frac{2PR}{P+R}$$

Table V shows the results of experiments constructed on the Web-Kb dataset, where the results of ICT-ILP, Co-training and Naïve Bayes derive from [20]. In table V, for relational tri-training, classifier1, classifier2 and classifier3 denote Nfoil algorithm, Kfoil algorithm and Aleph algorithm, respectively. For ICT-ILP, the classifier1 is Progol system and the classifier2 is Naïve Bayes. For co-training and Naïve Bayes, classifier1 denotes the heading-based classifier and classifier2 denotes the content-based classifier. Note that, all algorithms except relational tri-training algorithm have not classifier3.

TABLE V. PERFORMANCE OF CLASSIFIERS ON THE WEB-KB DATASET

Algorithm	Classifier 1			Classifier 2			Classifier 3		
	P	R	F_1	P	R	F1	P	R	F1
R-tri-training	91.38	95.06	93.18	94.61	95.48	95.05	74.81	91.93	82.49
ICT-ILP	80.00	81.82	80.90	82.61	86.36	84.44	--	--	--
Co-training	73.95	84.69	75.64	79.69	60.72	66.14	--	--	--
Naïve Bayes	74.99	87.24	79.91	76.95	84.18	79.60	--	--	--

Comparing the recall results first, we see that R-tri-training outperform other algorithms on all classifiers. Comparing the precision results and F_1 scores, we see that two classifiers (Nfoil, Kfoil) of R-tri-training outperform all classifiers of other algorithms, and one classifier (aleph) of R-tri-training is comparative with the classifiers of other algorithms. The reason that R-tri-training got the highest performance came from the contribution of relational learning algorithm and unlabeled data.

IV. CONCLUSION

Classification of web page content is essential to many tasks in web information retrieval such as maintaining web directories and focused crawling. The uncontrolled nature of web content presents additional challenges to web page classification compared with traditional text classification, but the interconnected nature of hypertext also provides features that can assist the process. An interesting aspect of the Web is that it can be thought of as a graph in which pages are the nodes of the graph and hyperlinks are the edges.

We apply relational semi-supervised learning algorithm named R-tri-training to web page classification. The relational component is able to describe the graph structure of hyperlinked pages and the internal structure of HTML pages. Furthermore, the semi-supervised learning component is able to exploit unlabeled data to enhance the learning performance. Experiment on Web-Kb is given and analyzed. We compare the result with the supervised Naive Bayes, which is the well-known algorithm for the text classification problem. The performance of R-tri-training is also compared with other semi-supervised learning algorithms which are Co-Training and ICT-ILP. The experimental results show that R-tri-training algorithm outperforms those algorithms.

In the analysis of the error rates change of classifiers (see figure 1, 2, 3, 4), it is noted that the performance of R-tri-training algorithm are usually not stable because the unlabeled examples may often be wrongly labeled during the learning process. A promising solution to this problem may be using data editing mechanisms to help identify the wrongly labeled examples. Incorporating data editing mechanisms into R-tri-training is an interesting issue to be investigated in future work.

REFERENCES

[1] X. Qi, and B.D.Davison, "Web page classification: features and algorithms," *ACM computing surveys*, 2009,41(2):12–42.

[2] M. Craven, and S.Slattery, "Relational learning with statistical predicate invention: Better models for hypertext," *Machine Learning*, 2001,43: 97–119.

[3] M. Craven, S. Slattery, and K. Nigam, "First-order learning for Web mining," *Machine Learning: ECML-98*, 1998: 250–255.

[4] A. Rosenfeld, R.A.Hummel, and S.W.Zucker, "Scene labeling by relaxation operations," *IEEE Transactions on Systems, Man and Cybernetics*, 1976, 6: 420–433.

[5] S.Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*. San Francisco, CA: Morgan Kaufmann, 2003.

[6] Q.Lu and L.Getoor, "Link-based classification," *Proceedings of the 20th International Conference on Machine Learning (ICML)*, Menlo Park, 2003.

[7] R. Angelova, and G.Weikum, "Graph-based text classification: Learn from your neighbors," *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, 2006: 485-492.

[8] P.Sen and L. Getoor, *Link-based classification*. Technical Report CS-TR-4858, University of Maryland, 2006.

[9] S.A.Macskassy and F.Provost, "Classification in networked data: A toolkit and a univariate case study," *Journal of Machine Learning Research*, 2007,8(5): 935-983.

[10] D. Jensen, J. Neville and B. Gallagher, "Why collective inference improves relational classification," *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, 2004: 593–598.

[11] Y.J.Li and M.Z.Guo, "Relational tri-training: learning first-order rules exploiting unlabeled data," *Proceedings of the 12th china conference on machine learning*, 2010:28-38.

[12] L. De Raedt, *Logical and Relational Learning*. Springer, 2008.

[13] N.Landwehr, K.Kersting, and L.D.Raedt, "Integrating Naïve Bayes and FOIL," *Journal of Machine Learning Research*, 2007, 8: 481–507.

[14] N.Landwehr, A.Passerini, L. De Raedt, and P.Frasconi, "kFOIL: Learning simple relational kernels," *Proceedings of the twenty-first national conference on artificial intelligence and the eighteenth innovative applications of artificial intelligence conference*, Boston, 2006, 7: 16–20.

[15] http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/aleph

[16] Z.H.Zhou and M.Li, "Tri-training: Exploiting unlabeled data using three classifiers," *IEEE Transactions on knowledge and Data Engineering*, 2005,17(11):1529–1541.

[17] S.Muggleton, "Inverse entailment and Progol," *New Generation Computing Journal*, 1995, 13: 245–286.

[18] T.Horváth, R.Sloan, and G.Turán, "Learning Logic Programs with Random Classification Noise," *Proceeding of the 6th workshop on inductive logic programming(ILP-96)*, 1996: 97–118.

[19] S.R.Singh, H.A. Murthy and T.A.Gonsalves, "Feature selection for text classification based on Gini coefficient of inequality," *Proceedings of the Fourth International Workshop on Feature Selection in Data Mining*, 2010: 76–85.

[20] N. Soonthornphisaj and B. Kijsirikul, "Combining ILP with Semi-supervised Learning for Web Page Categorization," *International journal of Computational Intelligence*, 2005: 271–274.

[21] A. Blum, and T.Mitchell, "Combining labeled and unlabeled data with co-training," *Proceedings of the 11th annual conference on computational learning theory*, Wisconsin, 1998: 92–100.

[22] S. Golman and Y. Zhou, "Enhancing supervised learning with labeled data," *Proc. Of the 17th Int'1 Conf. on Machine Learning(ICML2000)*, San Francisco, 2000: 327–334.

Yanjuan Li is a Ph.D. candidate of Harbin Institute of Technology of China. She received B.S. degree in Computer Science and Technology College from Harbin Normal University in 2002, and received M.S. degree in Department of Computer Software and Theories from Harbin University of Science and Technology in 2005. His research interests include machine learning, data mining, pattern recognition and knowledge-based system. His current work is focused on learning with unlabeled examples, relational learning and its application to web page classification.

Maozu Guo is a professor at the Department of Computer Science & Engineering of Harbin Institute of Technology. He received the B.S. and M.S. degrees in computer science from Harbin Engineering University, China, in 1988 and 1991 respectively, and PhD degree in computer science from Harbin Institute of Technology, China, in 1997. As a visiting scholar, he once worked at AI group of the Computer Science Department at Lund University in Sweden for the academic year 2002-2003. His research interests are in machine learning, data mining, computational biology and bioinformatics. In these areas he has published over 90 technical papers in international journals or conference proceedings. He is on the editorial boards of Journal of Test and Measurement Technology and Journal of Optics and Precision Engineering. He served as program committee member for various international conferences and chaired many native conferences. He is a committee member of Chinese Association for Artificial Intelligence (CAAI), and the councilor of CAAI Machine Learning Society.

Workforce Planning of Navigation Software Project Based on Competence Analysis

Xie Shangfei

School of Mechanical Engineering and Automation, Northeastern University, Shenyang, China
Email: xieshangfei@126.com

Sun Zhili

School of Mechanical Engineering and Automation, Northeastern University, Shenyang, China
Email: zhlsun@mail.neu.edu.cn

Abstract—**This paper introduces the quantitative research method for the personnel configuration of the software project by studying the effects of the overall competence of the developers in the vehicle navigation software project on the factors like project quality. The study shows the overall competence of the developers is related to the after-submission defect density, productivity and the average delay of software Version 0.99. Further, a quantitative formula of competence and competence is drawn on the basis of statistics; meanwhile, according to the research result an integer programming configuration method for the navigation software project personnel based on competence analysis is concluded.**

Index Terms—**software project quality, competence analysis, defect density, integer programming, workforce planning**

I. INTRODUCTION

Vehicle navigation system, a kind of vehicle electronic device, not only can show electronic maps and locate the position of the vehicle, but also can search information and choose the optimal route through calculation to guide the drivers to the destination, thus the traffic efficiency will be improved. The fundamental functions of vehicle navigation system include human-computer interfaces, vehicle location, map display, route calculation, real-time guidance and information search [1, 2].

Any software development process is a strict logical thinking process, and the visibility and the controllability of the process is poor. The reliability and quality of the system depends on the developers' individual behavior features. The faults may be generated by software coding mode, development process and user's usage mode. The software coding mode includes code size, programming technology, algorithm, statement structure and so on. The longer software code size is the more complex software structure is. Development process includes the engineering technology, tools and developer's competency. Bailey indicates us the fault ratio in program and data and the result is shown as Table I [3, 4, 5].

TABLE Ⅰ.
FAULT RATIO IN PROGRAM AND DATA BY DEVELOPER

Fault Type	Ratio
System Level Faults	10.0% - 25.0%
Data Writing Faults	10.0%
Transcription Faults	2.5%
Data Input Faults	0.5%

Since the logic structure of the vehicle navigation software is complicated and its development cycle is long, the competence of developers greatly influences the quality of the vehicle navigation software products. How to combine the competence of the developers and the quality target of projects to appropriately configure the personnel is the problem the software developing enterprises are considering.

The famous counterintuitive principle in software testing by Myers basically states that the more defects found during testing, the more that remained to be found later. The reason is that at the late stage of formal testing, error injection of the development process is basically determined. High testing defect rates indicate that the error injection is high; if no extra effort is exerted, more defects will escape to the field. Low defect density implies the high quality of the product [6]. Defect Density will be adopted here to describe the quality of the software system.

This paper concludes the linear programming configuration method for navigation software project personnel based on competence analysis by studying the effects of the overall competence of the developers in the vehicle navigation software project on the factors like project quality (after-submission defect density), productivity and the average delay of software Version 0.99 and the interrelation through a quantitative statistic research.

II. RESEARCH SCOPE AND METHODS

A. Research Scope and Data

This paper studies a project with the same business direction (vehicle navigation software system) and the

similar team scale without taking the difficulty into account, and analyzes the correlation developer competence, after-submission defect density, Productivity, the average delay of software Version 0.99. The key factors that are closely related to the project performance are taken as the research object. Therefore, the method to forecast the overall competence of the developers on the basis of after-submission defect density(X1), Productivity(X2), the average delay of Ver0.99(X3).

According to the analysis of the project data last three years, twenty projects with the same business direction and similar personnel scale are selected as the specific research object. Detailed data are shown in Table I. Y is the total competence of the project personnel, X1 is the after-submission defect density, X2 is Productivity, and X3 is the average delay.

TABLE II.
ANALYSIS DATA SHEET

Project	Y	X1	X2	X3
A	58	3	6	2
B	68	3	6	2
C	70	4	4	1
D	63	2	6	6
E	78	1	4	4
F	73	2	5	5
...	...	...	...	...

.

B. Research Method and Procedure

Forecast the competence of the personnel according to QCD(after-submission defect density, Productivity, Ver0.99), and then conclude the correlation of Y, X1, X2 and X3 through regression analysis.

The research procedure is as follows:

1) Select the research samples: Twenty navigation software development project samples with stable business modes and similar scale are selected after screening the data.

2) Collect the project information: Collect the detailed information including project name, project number, starting and ending time, project scale, development time, product model number, the software model number, development module number, scale of per model(software model), % Reuse, demand total, after-submission defect number, Productivity, average delay of Ver0.99, etc.

3) Collect the detailed competence information: Collect the detailed competence information of the project personnel, including more than sixty critical competence items in the project development process such as process capability (project planning, project monitoring, quantitative analysis, cause analysis, project conclusion, etc.), professional capability(fundamental knowledge, system architecture, C/C++ development and testing, vehicle bus, imbedded developing environment and tools, etc), and soft capability (communication capability, problem analysis and solving capability, learning capability, initiative responsibility, document writing capability, etc). The competence of each item of

each sample is equal to the total sum of each member's scores of this item.

4) Find the critical competencies: Find the critical competencies influencing X1, X2, X3 through correlation analysis. Carry out the correlation calculation of the 66 capability items and X1, X2, X3 respectively. The capability items whose relative coefficient $r \geq 0.3$ is confirmed as critical items, and the score of each critical item in the samples are kept as the basis of the data for further study next stage.

5) Sum up the score: Sum up the score of each critical item obtained from 4) to get the overall personnel competence of each project Y. And a regression equation is concluded by regression analysis and check the equation.

After the research result verification, a peer review is required. The review result will be applied in the construction and programming of vehicle navigation software project next stage.

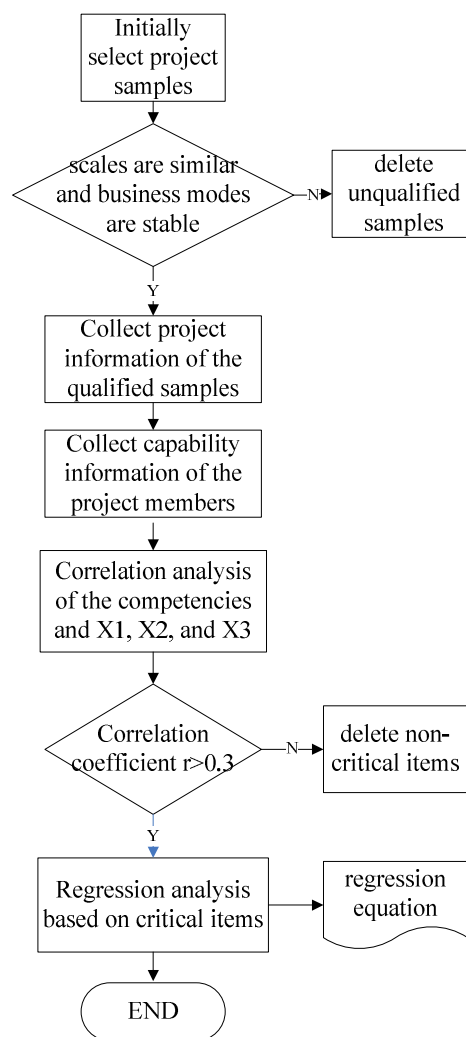The research procedure is shown as Figure 1.

Figure 1.   Quantitative analysis to the effect of critical competencies on project performance

Ⅲ. RESEARCH SCOPE AND METHODS

A. The Critical Competencies Analysis of Project Personnel

Workforce Competency refers to a series of skills and process capability possessed and developed by a person who works on the specific job type in an organization [7]. A human resource capability can be explained in an abstract level such as requirements understanding ability, coding ability, unit testing, project planning ability and communicative ability, etc in the software development.

During the quantitative study of the influences of project personnel competence on vehicle navigation software QCD, the identification of critical competencies is an essential step. The critical competencies of different project personnel through Spearman analysis are shown in Table Ⅲ:

TABLE Ⅲ. MATRIX OF THE CRITICAL COMPETENCIES

Role	CM1	CM2	CM3	CM4	CM5
Software development	A	B	C	D	E
	F	G	H	I	-
Software testing	J	K	L	H	M
Testing management	N	O	L	H	I
Project management	P	N	O	H	I
Quality management	N	O	H	I	M
Process improvement	H	I	M	-	-
Business consultation	P	O	J	H	I

A: Requirements understanding
B: Outline design
C: Specific design
D: Coding
E: Unit testing
F: Integration testing
G: Embedded OS
H: Communicative ability
I: Documentation writing
J: Requirements understanding
K: System testing
L: Testing planning
M: Initiative responsibility
N: Quantitative management
O: Reason analysis
P: Project planning

According to the above analysis, the critical competencies of members of software development, software testing, testing management, project management in the project. In addition to the total competence calculation through regression, the critical competencies can be improved by training, tutoring, self-learning, conference, project practicing. The regression analysis shows that the more competent the members are,

the higher the project product quality will be, and fewer defects will be.

B. The Regression Equation Based on the Critical Competencies

According to the critical competencies score of members of the project samples, we can obtain a regression equation including after-submission defect density, Productivity and the average delay of software Version 0.99. The regression equation is project competencies score=92.43.-1.528* after-submission defect density-3.345 *Productivity-0.338* the average delay of software Version 0.99.

We can know the higher project competencies score is the lower after-submission defect density is, and the project quality is better according to the regression equation.

C. The Methods of Workforce Planning based on the Integer Programming and Competencies Analysis

According to the above result of the regression analysis we find the quantitative relationship between the software project quality and the project members' competencies level, and based on the relationship and the regression equation the method of workforce planning in the navigation software project can be studied further.

In this part two ways of workforce planning in the navigation software project can be discussed. The first one is the way of workforce planning with the least project member number to ensure the project's quality, and the second one is the way of workforce planning with the least project cost to ensure the project's quality. More methods of the workforce planning can be designed in this way and they can be applied to all kinds of project situations.

1)   The method of workforce planning with the least project member number:

Precondition : The goals of after-submission defect density, productivity and the average delay of software Version 0.99 have been figured out, and the project competencies score N has been computed based on the regression equation in part 2.2.

Assumption:

The number of the project management is X1 and the persons' competencies average we can choose is V1.

The number of Quality management is X2 and the persons' competencies average we can choose is V2.

The number of Testing management is X3 and the persons' competencies average we can choose is V3.

The number of Software development is X4 and the persons' competencies average we can choose is V4.

The number of Software testing is X5 and the persons' competencies average we can choose is V5.

The upper limit of the project number is n.

In a common project X1=1, X2=1, X3=1

The IP equation is:

$$\text{Min } Z = X1+X2+X3+X4+X5 \tag{1}$$
$$\text{s.t. } X1*V1+ X2*V2+ X3*V3+X4*V4+X5*V5 \geq N \tag{2}$$
$$X1+X2+X3+X4+X5 \leq n \tag{3}$$
$$X1=1 \tag{4}$$
$$X2=1 \tag{5}$$
$$X3=1 \tag{6}$$
$$X4, X5 \geq 1 \text{ (Must be a integer)} \tag{7}$$

2)   The method of workforce planning with the least project cost

Precondition : The goals of after-submission defect density, Productivity and the average delay of software Version 0.99 have been clear, and the project competencies score N has been computed based on the regression equation in part 2.2.

Assumption:

The number of the project management is X1 and the persons' competencies average we can choose is V1. The persons' salary average we can choose is S1.

The number of Quality management is X2 and the persons' competencies average we can choose is V2. The persons' salary average we can choose is S2.

The number of Testing management is X3 and the persons' competencies average we can choose is V3. The persons' salary average we can choose is S3.

The number of Software development is X4 and the persons' competencies average we can choose is V4. The persons' salary average we can choose is S4.

The number of Software testing is X5 and the persons' competencies average we can choose is V5. The persons' salary average we can choose is S5.

The upper limit of the project number is n.

In a common project X1=1, X2=1, X3=1

The IP equation is:

$$\text{Min } Z = X1*S1+X2*S2+X3*S3+X4*S4+X5*S5 \tag{8}$$
$$\text{s.t. } X1*V1+ X2*V2+ X3*V3+X4*V4+X5*V5 \geq N \tag{9}$$
$$X1+X2+X3+X4+X5 \leq n \tag{10}$$
$$X1=1 \tag{11}$$
$$X2=1 \tag{12}$$
$$X3=1 \tag{13}$$
$$X4, X5 \geq 1 \text{ (Must be a integer)} \tag{14}$$

Many other ways can be studied further except above ways. For example, we can discuss the way when the number of software development or software testing is limited, and the method is similar.

F. RESULT AND APPLICATION

After-submission defect density = (92.434-project competencies score-3.345*Productivity-0.338* the average delay of software Version 0.99)/1.528 (15)

In the project planning phase we can use the regression equation in part 2.2 and data including productivity, project competencies score and the average delay of software Version 0.99 to predict the project quality (after-submission defect density). If the project quality level prediction does not reach the goal, the project member can be adjusted and more experienced persons can join the project to make the project competencies score higher to increase the project quality.

The matrix of the critical competencies can be used in employee competency development, all kinds of development activities can planed to improve the project member's competency level. At last the competency development can increase the performance of the project member to make the whole project QCD better.

The method of workforce planning in the navigation software project based on the integer programming and competencies analysis enhances the capability and accuracy of the current ways of the software project planning. It can be used in the practical software development, and we can find the appropriate ways to make a workforce planning based on competencies in a specifically project in the same idea.

REFERENCES

[1] C. Nwagboso, P. Georgakis, D. Dyke, "Time Compression Design with Decision Support for Intelligent Transport Systems Deployment," Computers in Industry, 2004, vol . 54, pp. 291-306.
[2] Wang Jin-guo, Wang Wei-ya, Tian Feng, Wei Jun-bo. "Intelligent Vehicle Navigation System," Journal of Traffic and Transportation Engineering, 2005, No.2, pp. 106-109.
[3] Zheng Guo-xiong, Li Ying-nu. "Study on Reliability Strategies for Military Software," Journal of Spacecraft TT&C Technology, 2003, vol. 22, No.3, pp. 75-79.
[4] Zhang Xue-mei, Hoang Pham. "An analysis of factors afecting software reliability", The Journal of Systems and Software, 2000, vol. 50, pp. 43-56.
[5] A. L. Goel, K. Okumoto, "An analysis of recurrent software failures in a real-time control system". Proc. ACMAnnu Tech Conf, ACM, Washington, DC, 1987, pp. 496-500.
[6] Yu Bo, Jiang Yan. "Software Quality Management Practice: Practical Methods about Software Defect Prevention, Modification and Management," Beijing: Publishing House of Electronics Industry，2008, pp. 274-277.
[7] B. Curtis, W. Hefley, S. Miller, The People Capability Maturity Model: Guidelines for Improving the Workforce. India: Dorling Kindersley, 2002, pp. 517-518.

Does the Valence of Online Consumer Reviews matter for Consumer Decision Making? The Moderating Role of Consumer Expertise

Peng Zou
School of Management, Harbin Institute of Technology, Harbin, China
Email: zoupeng1975@gmail.com

Bo Yu
School of Management, Harbin Institute of Technology, Harbin, China
Email: yub@hit.edu.cn

Yuanyuan Hao
School of Management, Harbin Institute of Technology, Harbin, China
Email: hyy@163.com

Abstract— **The previous studies have shown inconsistent relationship between the valence (positive or negative) of online consumer reviews and consumer decision making. With accessibility/diagnosticity theory, this study attempts to explain this discrepancy through exploring consumer expertise as a moderator. Our results from a 2 * 2 experiment design indicate that the impact of online reviews valence is moderated by consumer expertise: The impact difference between negative reviews and positive reviews is greater for consumers with low expertise than for those with high expertise. Our study adds to the literature relevant with e-WOM effect. And we also provide managerial implications for e-marketers.**

Index Terms—**online consumer reviews, valence, consumer expertise, accessibility/diagnosticity theory**

I. INTRODUCTION

With the rapid development of Internet and e-commerce, online product reviews generated by users, one form of electronic word-of-mouths (e-WOMs), have become an important information source for consumer purchase decision making [1][2][3]. Communication direction (also named as Valence, positive or negative) [4] is one of the most focused dimensions of online reviews, because communication effects and WOM marketing strategies are significantly distinguished for different valences reviews. Thus, disclosing the relationship between online reviews valence and consumer decision is valuable for marketers to make WOM management strategies.

Previous researches show an inconsistent relationship between the valence of online reviews and consumers' purchase intention (or behavior). Some studies indicate the positive relationship between them [2]; while some other researches find the negative relationship between them [6]. There are also other else studies [7][8] arguing that there is no significant relationship between these two factors. Which conclusion actually reveals the fact?

From the end of WOM reception, some few researches (such as [9]) showed that the communication effect of positive e-WOMs may vary greatly due to WOM receivers' different information processing levels, which depend on consumers' acquired product knowledge (also named consumer expertise). While according to accessibility/diagnosticity theory [10], there generally exists asymmetry between the effect of positive WOMs and that of negative ones, and negative ones are usually more diagnostic than positive ones. So there comes a new question: whether the effect of negative reviews is equally impacted by consumer expertise as positive reviews? As another word, whether does the effect of review valence (positive or negative) is moderated by consumer expertise? As far as we know, there is lack of researches exploring whether the effect of online reviews valence varies depending on consumer expertise.

Therefore, in this study, based on the accessibility/diagnosticity theory, we attempt to explain the inconsistent results about the effect of online reviews valence by exploring consumer expertise as a moderator.

II. HYPOTHESIS

According to accessibility/diagnosticity theory, negative information is usually more diagnostic than positive one [10]. Because the product rated to be positive can be in high, average and low quality, prone to be more ambiguous; On the contrary, negative information strongly suggests inferior performance. In addition, this theory also points out that the diagnosticity of information may be situation dependent. When it is difficult for the consumer to judge or it is ambiguous in some situation, the diagnosticity of available information will increase. When a consumer perceives he/she lacks relevant product knowledge, perceived decision risk

increases because it is difficult for him/her to evaluate whether reviews truly indicate product quality. Since negative WOMs are relevant with potential loss or risk, consumers would rather trust negative WOMs to reduce potential purchase risk. Thus, the diagnosticity and impact of negative WOMs become much stronger when consumer expertise is low. Whereas, when the consumer with low expertise reads positive WOMs, the intrinsic vagueness of these WOMs enhances, which weakens their diagnosticity. Thus the effect of positive WOMs decreases for consumers with low expertise. On the contrary, when a consumer perceived product knowledge or expertise is high, his/her perceived decision making risk is not as high as the consumer with low expertise. Because he/she believes that the truthfulness of others' reviews can be assessed properly according to his/her product knowledge and information processing ability. Therefore, the diagnosticity and impact of negative WOMs become weaker for consumers with high expertise. In contrast, the ambiguity of positive WOMs reduces because of consumers' rich product knowledge and powerful information processing ability. So the effect of positive WOMs becomes greater for consumers with high expertise. Based on the above arguments, we propose the following hypothesis.

Hypothesis: Consumer expertise moderates the effect of online consumer review valence. In detail, the impact of review valence is stronger for low expertise consumers than for high expertise consumers.

(As another word, the impact difference between positive reviews and negative reviews is greater for low expertise consumer than high expertise consumer)

III. RESEARCH METHOD

This study tests the above hypothesis through 2 (high expertise vs. low expertise) * 2 (positive online reviews vs. negative online reviews) experiment design. Between these two variables, review valence can be manipulated in advance, while consumer's expertise usually can't be manipulated before data collection. We first design two different experiment groups and then divide these groups into four groups according to the average scores of consumer expertise.

A. Subjects Selection

According to Nelson's product taxonomy of search goods and experience goods which is commonly adopted [11], we separately choose one most representative product from several products through a focus group interview, U Flash Drive for search goods and Face Lotion for experience goods. Then we separately collect approximately the same amount of online reviews of these two products for the experiment groups.

B. Manipulations of Review Valence

We first collected several online reviews relevant with these two products on some popular online shopping mall websites in China, such as Joyo website (www.amazon.cn). Then, we separately selected some of

1-star numerical rating reviews rating and 5-star numerical rating reviews with high helpfulness rating. To control the impact of variance between positive and negative contents for each review, we only extracted completely positive (completely negative) sentences from these reviews for positive (negative) reviews. To control the effect caused by review type, according to the definition of these two types of reviews [12], three attribute-based reviews and three benefit-centric reviews are chosen from these reviews for each group of reviews. To control the effect of information volume, we further select review sentences with about the same words.

C. Dependent Variable

This study examines how some factors moderate the effect of online reviews on consumer decision making. Previous relevant studies ever adopted WOM effect, WOM helpfulness, and consumer purchase intention as dependent variable. We choose WOM effect as the dependent variable in our study, which may be suitable for experimental situation. We measure WOM effect by the three-item scale from Jeon&Park [13] on a five-point rating of agreement. (See the appendix)

D. Control of Other Variables

Several other factors may affect WOM effect of online reviews, which are not focused in this study, such as individual differences including personal cognitive style, online purchase experience, prior brand attitude, general attitude on online reviews, product type, democracy (such as gender) and so on. According to Park&Lee [12], we try to control the effect of individual cognitive style through randomly assigning the participants to the experimental groups. We control the brand effect through hiding brand names of products. The individual differences in online purchase experience, general attitude on online reviews, product type and gender may also have effect on the dependent variable. If differences exist among experimental groups in these factors, we will include them into our model as covariate variables. Among them, one item about whether the respondent has online purchasing experience is asked to investigate personal online purchase experience. Besides, consumer general attitude on reviews are measured by four items with a 5 points rating of agreement according to Park&Lee's scale [12] (see the appendix).

E. The Experimental Procedure

We mainly sample university students as respondents for the following reasons: University students are a group with high Internet usage rate. And great numbers of these young surfers make their online purchasing depending on e-WOMs. Besides, partial impact of democracy can be removed because there is less heterogeneity in demographic features for this group of people. In the experiment, we invite 320 university students to participate in this experiment. After reading one group of online reviews, the respondents are asked to answer the following questions in the questionnaire. Finally, we got 290 validated samples.

IV.   ANALYSIS RESULTS

A. Manipulation and Control Checks

One item is designed to check whether review valence is manipulated successfully (see the appendix). T test for this item shows significant difference in attitude towards these reviews between positive reviews group and negative reviews group, $(T(285.37))=50.68$, $P<0.001$, Mean(positive)=4.56, and Mean(Negative)=1.65). Thus, review valence has been manipulated as intended.

The responses on the four items measuring consumer general attitude on online reviews (Cronbach's Alpha=0.802) are averaged to indicate the overall consumer general attitude on reviews. F test shows no significant difference in consumer general attitude towards online reviews among the groups $(F(3)=2.14$, $P<0.10)$. Whereas, different gender groups vary in WOM

effect $(T(285)=2.39$, $P<0.01$, Mean(Male)=3.49, Mean(female)=3.69). In addition, T test shows significant difference in WOM effect between groups with online purchase experience and those with no online purchase experience. $(T(285)=-3.819$, $P<0.001$, Mean(no online purchase experience)=3.39, Mean(ever purchased online)=3.70)). Thus, product type, gender and online purchase experience are included as covariates in the following ANOVA analysis.

B. Hypotheses Testing

Descriptive statistics of WOM effect among these groups are shown in Table I. Taking product type, gender and online purchase experience as covariates, we conduct an all two-way ANCOVA analysis to test the interaction effect between consumer expertise and the valence of reviews (See Table II for results).

TABLE I
DESCRIPTIVE STATISTICS OF WOM EFFECT AMONG THE EXPERIMENTAL GROUPS

Review Valence / Expertise	Positive reviews	Negative reviews	Total	T-value (P) for pairwise comparison
Low Expertise	3.239 (n=68)	3.676 (n=74)	3.493 (n=142)	-3.584 (0.000)
High Expertise	3.589 (n=73)	3.727 (n=75)	3.642 (n=148)	-1.286 (0.200)
Total	3.416	3.720	-	-

(Note: n denotes number of responses in corresponding groups; P denotes the p value for T test)

TABLE II
ANCOVA RESULTS FOR THE INTERACTION EFFECT BETWEEN
THE VALENCE OF ONLINE REVIEWS AND CONSUMER EXPERTISE

Effect	F-value	P-value	Sig.
Review Valence	15.052	0.000	***
Consumer Expertise	3.134	0.078	*
WOM Valence * Consumer Expertise	4.106	0.044	**
Gender (as covariate)	6.330	0.012	**
Online Purchase Experience (as covariate)	11.208	0.001	***
Product Type (as covariate)	0.629	0.428	

(Note: * denotes significance at 0.1 level; ** denotes significance at 0.05 level; *** denotes significance at 0.01 level)

According to Table II, the main effect of review valence is significant $(P<0.001)$. The effect of negative reviews (3.720) is stronger than that of positive reviews (3.416) if we don't consider the impact of consumer expertise. While the main effect of consumer expertise is significant at 0.1 level $(P<0.078)$, indicating slightly significant difference in WOM effect for consumers with low expertise and ones with high expertise if the impact of review valence is not considered. The interaction effect between review valence and consumer expertise is shown significant $(P<0.044)$. Among all the three covariates, the main effects of gender, consumer online purchase experience are significant (P values are 0.012 and 0.001 separately). Whereas, there is no significant difference between different types of products when we control other factors to be equal.

To confirm the detailed form of interaction effect between review valence and consumer expertise and test

the above hypothesis, we further compare the impact difference of positive reviews and negative reviews separately for low consumer expertise and high consumer expertise situations through T tests. (For T tests results, see Figure 1 and the last column in Table 1) For consumers with low expertise, the effect of negative reviews is significantly greater than that of positive reviews $(T=3.584$, $P<0.001$, Mean (negative reviews) =3.676, Mean (positive reviews) =3.239). For consumers with high expertise, though the effect of negative reviews (3.727) is slightly greater than that of positive reviews (3.589), this difference is not statistically significant $(T=1.286$, $P<0.2)$. Thus, the impact difference of negative reviews and positive reviews for consumers with low expertise (3.676-3.239=0.437) is significantly greater than that for consumers with high expertise (3.727-3.589=0.138). It is mainly because consumers with high expertise have more knowledge about the

product than ones with low expertise so that they can make purchase decision more independently. Therefore, the hypothesis in our study is supported.
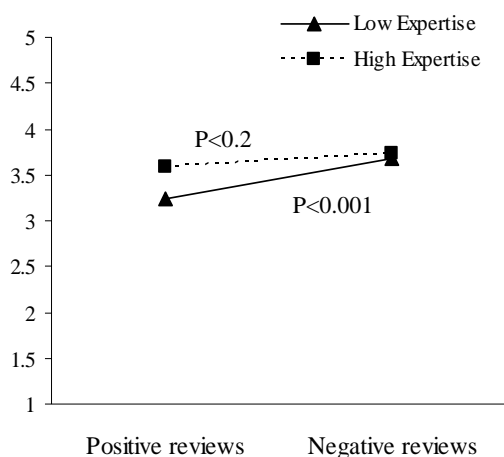
(Note: Doted line indicate non-significant difference at 0.05)

Figure 1 T test results for the form of interaction effect between consumer expertise and WOM valence

V. CONCLUSIONS

To explain the inconsistent relationship between the online reviews valence and consumers' decision in previous studies, we explore the moderating effect of consumer expertise on the impact of review valence.

From the theoretical perspective, we confirm the effect of online reviews valence is situation dependent and may change under the moderation of consumer expertise for the first time. Inferring from accessibility/diagnosticity theory, we suppose and validate that the effect of review valence is greater for consumers with low expertise than that for consumers with high expertise. Our findings may add to existing literature relevant with impact factors of e-WOM communication effect.

From the managerial perspective, our findings have implications for marketers on how to develop e-WOM marketing more effectively and economically for consumers with different levels of product expertise. Our study can help in telling e-WOM marketers whether e-WOM valence has impact on consumer decision making for consumers with different expertise, for which type of consumers the effect of e-WOM valence is greater and which type of consumers should be first focused when developing positive or negative WOM marketing. According to the different directions and strengths of WOM valence under different situations, marketers can correspondingly adjust the e-WOM marketing strategies more economically and effectively, for selling the goods such as wearing, which need less knowledge of product should design more standard WOM format so that the consumer referral influence potential buyer. For selling the goods such as digital camera, which needs more specific attribute information should add more descriptive content rather than simple positive or negative option choice.

REFERENCES

[1] Dellarocas, C., "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," *Management Science*, Vol. 49, No. 10, pp. 1407-1424, 2003.

[2] Chwen-Yea Lin, Kwoting Fang and Chien-Chung Tu, "Predicting consumer repurchase intentions to shop online," *Journal of Computers*, VOL. 5, NO. 10, pp. 1527-1533, 2010.

[3] Shouming Chen and Jie Li, "Examining Consumers' Willingness to Buy in Chinese Online Market," *Journal of Computers*, VOL. 5, NO. 5, pp. 815-824, 2010.

[4] Ray L. Benedicktus, Melinda L. Andrews, "Building Trust with Consensus Information: the Effects of Valence and Sequence Direction," *Journal of Interactive Advertising*, VOL 6, NO. 2, pp. 15-25, 2006

[5] Sen and D. Lerman, "Why are you telling me this? An examination into negative consumer reviews on the Web," Journal of Interactive Marketing, Vol. 21, p. 76-94, 2007.

[6] J.A. Chevalier and D. Mayzlin, "The Effect of Word of Mouth on Sales: Online Book Reviews," Journal of Marketing Research, Vol. 43, pp. 345-354, 2006.

[7] Yong Liu, "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," Journal of Marketing, Vol. 70, pp. 74-89, 2006.

[8] Duan, W., B. Gu, and A. B. Whinston, "The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry," Journal of Retailing, Vol. 84: pp. 233-242, 2008.

[9] D.H. Park and S. Kim, "The effects of consumer knowledge on message processing of electronic word-of-mouth via online consumer reviews," Electronic Commerce Research and Applications 2008.

[10] Feldman, JM., Lynch, JG., "Self-generated Validity and Other Effects of Measurement on Belief, Attitude, Intention and Behavior," Journal of Applied Psychology, Vol. 73, pp. 421-435, 1988.

[11] Nelson, P., "Advertising as Information," *Journal of Political Economy*, Vol. 82, pp. 729, 1974.

[12] D.H. Park, J. Lee, and I. Han, "eWOM overload and its effect on consumer behavioral intention depending on consumer involvement," *Electronic Commerce Research and Applications*, 2008.

[13] Jeon SY, Park HJ., "The influence of information characteristics on word-of-mouth effect," *J. Consumer Study*, Vol. 14, pp. 21-44, 2003.

Peng Zou is Lecturer of marketing in the school of management, Harbin Institute of Technology. He eared his PhD in Management Science and Engineering, Harbin Institute of Technology. He teaches undergraduates and graduates courses in Marketing and Customer relationship management His research areas of interest include customer relationship management.

APPENDIX

TABLE A. MEASURES AND MANIPULATION CHECK ITEMS OF VARIABLES

	Measures
WOM Effect (Cronbach's Alpha=0.827)	I will refer to these reviews in my purchase decision.
	Overall, I think these reviews are credible.
	These reviews will crucially affect my purchase decision.
General Attitude on Online Reviews (Cronbach's Alpha=0.802) "When I buy a product online, "	I always read online product reviews posted by other users.
	The online product reviews are helpful for my decision making.
	The online product reviews make me confident in purchasing the relevant product.
	If I don't read the online product reviews, I will worry about my decision.
Consumer Expertise (Cronbach's Alpha=0.939)	I am knowledgeable about USB Flash Drive/ Face Lotion.
	I have rich purchasing experiences on USB Flash Drive/Face Lotion.
	I learn well about USB Flash Drive/Face Lotion.
	I am an expert on USB Flash Drive/ Face Lotion.
WOM Valence	What is the overall attitude of these user reviews towards this product? A. completely negative … E. completely positive

On-line NNAC for a Balancing Two-Wheeled Robot Using Feedback-Error-Learning on the Neurophysiological Mechanism

Xiaogang Ruan

School of Electronic Information and Control Engineering, Beijing, China
Email: adrxg@bjut.edu.cn

Jing Chen (corresponding author)
School of Electronic Information and Control Engineering, Beijing, China
Email: chenjing0828@139.com

Abstract—**A biomimic control structure is designed based on the sensorimotor system and the central nervous system of human brain. The main control center is focus on the cerebellar cortex functioning in supervised training. And then a method based on cerebellar model with feedback-error-learning on the neurophysiological mechanism was proposed to control a balancing two-wheeled robot which is a bionic self-balancing plant imitating human body. The cerebellar cortex as a forward controller was an on-line neural network adaptive controller and reflected the cerebellar function in balancing control problem. From the simulation including balancing control experiment, disturb experiment and mass changed experiment, we can see that the robot can be balanced in fixed position well by the biomimic control structure, and get a better result comparing the traditional feedback control.**

Index Terms—**Balancing Two-Wheeled Robot; balancing control; sensorimotor system; feedback-error-learning; cerebellar internal model**

I. INTRODUCTION

There are many kinds of two-wheeled self-balancing robot, such as wheeled-inverted pendulum, self-balancing wheelchair, JOE, nBot[1][2][3][4], and so on. All of them have ability of keeping balanced themselves and running forward. Many researches [3][5][6] on the balancing control are done after decoupling. Essentially, it is a kind of control method based on linear inverted pendulum. A lot of work has been done for such an inverted pendulum system. An adaptive output recurrent cerebellar model articulation controller is utilized to control wheeled inverted pendulums (WIPs)[7]. A mobile robot in [8] was controlled using a trajectory tracking algorithm based on a linear statespace model. Grasser et al. [3] developed a dynamic model for designing a mobile inverted pendulum controller using a Newtonian approach and linearization method. In [1], a dynamic model of aWIP was created with wheel motor torques as input, accounting for nonholonomic no-slip constraints. Vehicle pitch and position were stabilized using two controllers. Jung and Kim [9] created a mobile inverted

pendulum using neural network (NN) control combined with a proportional-integral-derivative controller. However, the balancing two-wheeled robot is a kind of MIMO nonlinear system, so the research on the control of MIMO nonlinear system is very important to robot control field.

In 1987, Kawato [10] proposed the inverse model and then proposed feedback error learning model [11][12], which was designed from a biological perspective to establish a computational model of the cerebellum for learning motor control with internal models in the central nervous system. The error between expected and actual track was taken as the training signal, at the same time it solved the problem of normalization with different error information. As Kawato's opinion, forward internal model predicts the action sequence and correspond with feedback control, which can overcome the time delay. Reference [13] present some theoretical investigations of the technique of feedback error learning (FEL) from the viewpoint of adaptive control, and prove its stability for SISO control system. Based on the balancing control of nonlinear system, in the year 2007, an on-line adaptive control for balancing based on feedback-error-learning scheme was proposed successfully into the inverted pendulum balancing control. To some extent, this learning scheme shows the function of cerebellum in keeping body balancing.

As we all know, in the nerve system, motor nervous system is for the motion and posture control. Our robot is designed imitating human, while cerebellum takes part in the balancing control. Therefore, it is very important to research the balancing control of robot according to cerebellar balancing control principle in the sensorimotor system, and there is important theoretical and practical value, especially for the human-simulated intelligent control (SHIC).

In Section Ⅱ, the sensorimotor system is introduced, including Central Nerve System and Cerebellum Internal Model. The sensor motor principle is discussed on the neurophysiology concept. In Section Ⅲ, the model of the balancing two-wheeled robot has been built and the

bionic control structure simulating the sensorimotor system is introduced. In Section Ⅳ, the feedback-error-learning algorithm is proposed with the cerebellar cortex as the adaptive neural network, and the on-line learning process is proposed for the continuous MIMO two-wheeled robot. The simulation experiment has been done for the balancing control with fixed-point using the method proposed in this section, and the comparative experiments with traditional feedback control method have been done. At the same time, experiment with step disturb has been made. Finally, the article comes to a conclusion in Section V.

II.　Sensorimotor System

A.　Central Nervous System

An important feature of the Central Nervous System (CNS) is its remarkable ability to adapt to changes both in the environment and inside the body. Motor nervous system includes motor center (MC) and PNS (Peripheral Nervous System). The whole nervous system can also be called sensorimotor nervous system. Motor nervous system is composed of three-level hierarchical structure and two auxiliary monitor systems. From low to high the three-level hierarchical structure is respectively the spinal cord (SC), brainstem (BS) and cerebral cortex (CC), while two auxiliary monitor systems mainly are the cerebellum and ganglion nuclear group in the forebrain area as the core. These brain areas related to motor control form interconnected circuit, and deal with all kinds of information processing of motion and postures. Figure 1(a) shows the regulation and control mode of behavior central; Figure 1(b) shows the global network of central nervous system linking the cerebellum, the basal ganglia and the cerebral cortex. the CB, the BG, and the CC exist for different types of learning.
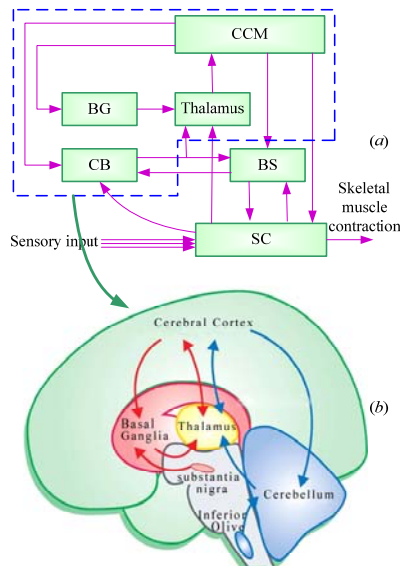
Figure 1　(a) regulation and control mode of behavior central; (b) The global network of central nervous system linking the cerebellum, the basal ganglia and the cerebral cortex. CCM, cerebral cortex model; CB, cerebellum; BS, brainstem; SCM, spinal cord model

During the course of sensorimotor learning, the cerebral cortex(CC), basal ganglia(BG) and cerebellum(CB) work in parallel but particular ways [14][15][16]. The loop through the basal ganglia learns to discover ballpark actions that are appropriate in a given context [17], utilizing reinforcement learning. The loop through the cerebellum learns to refine the action through supervised learning. The cerebral cortex, driven by input from basal ganglia and cerebellum, learns through practice to perform these operations faster and more accurately, utilizing unsupervised Hebbian learning. The coordination of these diverse forms of learning has been simulated by Doya[15].

B.　Cerebellum Internal Model

Along with the development of medical science, the research about brain structure is very mature to some extent. Figure 2(a) shows that the function of brain contains thinking, feeling, hearing, speech, motion, sensation, vision, balance and so on. Brain is the place of advanced nerve centre and controls these actions of human being. Thereinto, the cerebellar part mainly realizes the balance control of body. So that it is inevitable to design the balance system of robot using the cerebellar model mechanism in the research of bionic intelligent control. Figure 2 (b) exhibits the structure of cerebellum.
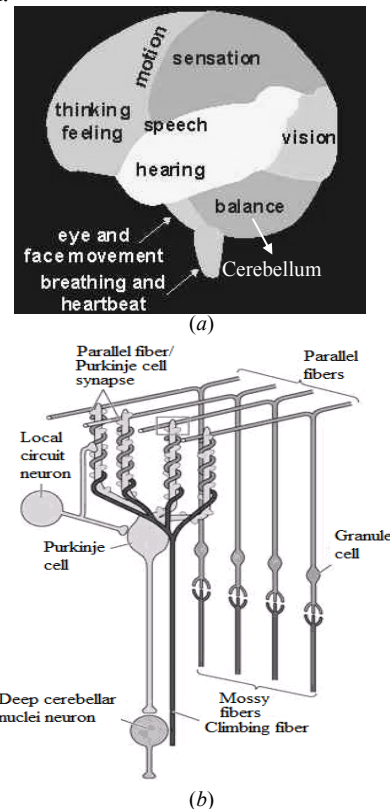
Figure 2　(a) structure of brain (b) structure of cerebellum

As above discussed about the CNS, basal ganglia and cerebellum have different learning style. Cerebellum play an important part in the supervised autonomic learning, and it is shown in Figure 3 that cerebellum work with

other part of brain for regulating and controlling the behavior.

In 1960s, many researchers (such as Brindley) in the field of neurophysiology proposed that if synapse connection between parallel fibre and Purkinje cell could be amended, memory would be shaped. In the year 1968, Flourens became aware of the dynamic function of cerebellum. He pointed out that the function of cerebellum is to coordinate motion. For motion cerebellum is not necessary, but if there was no cerebellum, limb motion would became jerky, trembling, loose and imprecise. Ito [18] brought the concept of internal model into neurophysiology and proposed that internal model is the result of learning or conditioning. Basal ganglia used internal model to carry out motion control. And he pointed out the hypothesis of forward model. The disadvantage is that the feedback gain of forward model is small but the feedback delay is great, so it is difficult to realize the motion control depending only on forward model for basal ganglia organization. It makes against the rapid and steady of the motion control. The cerebellum is known to be critical for accurate adaptive control and motor learning.
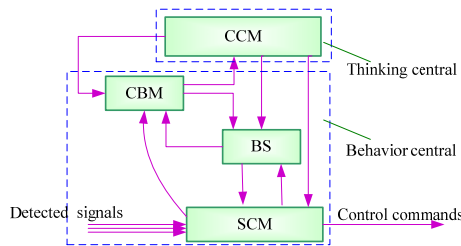
Figure 3 Regulate and control mode of behavior central with cerebellar model. CCM, cerebral cortex model; CBM, cerebellar model; BS, brainstem; SCM, spinal cord model

III. TWO-WHEELED ROBOT SELF-BALANCING PROBLEM

A. B-TWR System Description

Here we study the B-TWR with two control inputs, which are the torques of the two wheels. And the system degree of the freedom is three, which is more than the control inputs. So the system belongs to under-actuated system.

B-2WMR is a nonholonomic system, and here the rigid model is discussed, which has two coaxial driving wheels. And there is an internal body which can install some subsystems, such as controllers, sensors, etc. The internal body is said to be Intermediate Body (i.e. IB). The holistic physical framework is shown in Figure 4(left).
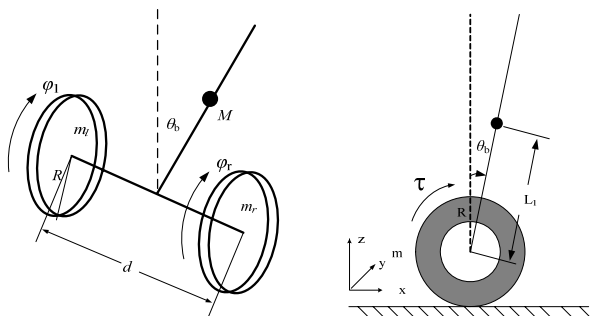
Figure 4 left: system whole structure; right: side elevation of robot after simplification

Definition of parameters and variables are all shown in Table I.

TABLE I
DEFINITION OF PARAMETERS

Parameter	Unit	Value
m_l, m_r : the mass of the left and the right wheel	kg	1
M: the mass of the Intermediate Body	kg	10
R: radius of wheel	m	0.15
d: the distance between the center point of two wheels	m	0.44
L_1 : the distance between the barycentre of IB and O	m	0.4
$J_l = J_r = J_w$: the moment of inertia of (left and right wheel) about its axis	kg·m²	1.125e-2
J_y : the moment of inertia of the robot about the y –axis	kg·m²	1.5417
J_z : the moment of inertia of the IB about the z –axis	kg·m²	0.5893
g: acceleration of gravity	m/s²	9.8
μ_l, μ_r :frictional coefficient of two wheel with ground	N·m/(rad/s)	10,10
μ_w : frictional coefficient of the wheel axis	N·m/(rad/s)	1
x v position and velocity	m and m/s	var
$\theta_b, \dot{\theta}_b$: the inclination angle and angle velocity of the Intermediate Body	rad and rad/s	var
$\theta, \dot{\theta}$: yaw angle and angle velocity of robot.	rad and rad/s	var
φ_r, φ_l : rotary angle of the right and left wheel.	rad	var
ω_l, ω_r : rotary angle velocity of the right and left wheel	rad/s	var
τ_d : the disturb put on the body of robot	N.m	var
τ_l : torque provided by left motor	N.m	[-5,5]
τ_r : torque provided by right motor	N.m	[-5,5]

B. B-TWR Mathematical Model

The math model of balancing two-wheeled robot has been built, and the correctness has been validated.

The dynamic model can be described as

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{D}(\mathbf{q}, \dot{\mathbf{q}}) = \mathbf{E}\boldsymbol{\tau} + \mathbf{u_d} \qquad (1)$$

In this equation, $\mathbf{M}(\mathbf{q}) \in \mathbb{R}^{n \times n}$ is the system inertial matrix, $\mathbf{D}(\mathbf{q}, \dot{\mathbf{q}}) \in \mathbb{R}^{n \times 1}$ is the Coriolis force and gravity matrix, $\boldsymbol{\tau} \in \mathbb{R}^r$ is the input vector, $\mathbf{E} \in \mathbb{R}^{n \times r}$ is the transition matrix of the input vector. $\boldsymbol{\omega} = \mathbf{u_d} \in \mathbb{R}^{r \times 1}$ is the disturb put on the robot.

where

$$\mathbf{M}(\mathbf{q}) = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

$$\mathbf{D}(\mathbf{q}, \dot{\mathbf{q}}) = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}, \mathbf{E} = \begin{bmatrix} -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{u_d} = \begin{bmatrix} \tau_d \\ \tau_{dl} \\ \tau_{dr} \end{bmatrix}, \boldsymbol{\tau} = \begin{bmatrix} \tau_l \\ \tau_r \end{bmatrix}$$

$$a_{11} = (ML_1^2 + J_y); a_{12} = a_{21} = a_{13} = a_{31} = \frac{1}{2} ML_1 R \cos\theta_b$$

$$a_{22} = (m_l R^2 + J_w) + \frac{1}{4} MR^2 + \frac{R^2}{d^2} (ML_1^2 \sin^2\theta_b + J_z)$$

$$a_{23} = a_{32} = \frac{1}{4} MR^2 - \frac{R^2}{d^2} (ML_1^2 \sin^2\theta_b + J_z)$$

$$a_{33} = (m_r R^2 + J_w) + \frac{1}{4} MR^2 + \frac{R^2}{d^2} (ML_1^2 \sin^2\theta_b + J_z)$$

$$d_1 = -MgL_1 \sin\theta_b - M\dot\theta^2 L_1^2 \sin\theta_b \cos\theta_b + 2\mu_w \dot\theta_b - $$
$$\mu_w (\dot\varphi_l + \dot\varphi_r)$$

$$d_2 = ML_1^2 (\frac{R}{d})^2 (\dot\varphi_l - \dot\varphi_r) \dot\theta_b \sin 2\theta_b - \frac{1}{2} ML_1 R\dot\theta_b^2 \sin\theta_b + $$
$$(\mu_l + \mu_w)\dot\varphi_l - \mu_w \dot\theta_b$$

$$d_3 = -ML_1^2 (\frac{R}{d})^2 (\dot\varphi_l - \dot\varphi_r) \dot\theta_b \sin 2\theta_b - \frac{1}{2} ML_1 R\dot\theta_b^2 \sin\theta_b + $$
$$(\mu_r + \mu_w)\dot\varphi_r - \mu_w \dot\theta_b$$

The model is different from others. The frictions of two wheels with ground and of the wheel axis are considered. It's supposed that the two wheels are restricted by the restriction of the pure rolling.

$\mathbf{q}$ is the generalized coordinate of system, and can be defined as $\mathbf{q} = (\theta_b, \varphi_l, \varphi_r)^T$, τ_d is the disturb put on the body of robot. τ_l and τ_r are respectively the torque provided by the left and right motor. We choose the Maxon DC motor, and the torque is limited in the bound of $\pm 5Nm$ owing to the performance of motor.

The kinematics model of robot can be described as

$$\begin{cases} \dot{x} = \frac{1}{2}[\dot\varphi_l + \dot\varphi_r]R\cos\theta \\ \dot{y} = \frac{1}{2}[\dot\varphi_l + \dot\varphi_r]R\sin\theta \\ \dot\theta = \frac{R}{d}(\dot\varphi_r - \dot\varphi_l) \end{cases} \quad (2)$$

Where (x,y) denotes the position of robot in the Cartesian coordinate, and $\dot\theta = R(\dot\varphi_l - \dot\varphi_r)/d$ is the yaw angle velocity of robot.

C. The bionic control structure

Simulating human brain structure, we form the bionic control structure as in Figure 5. For the sensorimotor, there are mainly five part including receptor, efferent nerve; central nervous; afferent nerve and effector. The on-line NNAC concentrates on the central nervous.

IV. FEEDBACK ERROR LEARNING MODEL

A. Theoretical model with FEL

Supervised learning depends strongly on the availability of an external teacher. For a given set of inputs, neural network uses the error between the desired response from the teacher and the network's actual output to adjust the interconnection weights between each

neuron. The whole control scheme is proposed in Figure 6.
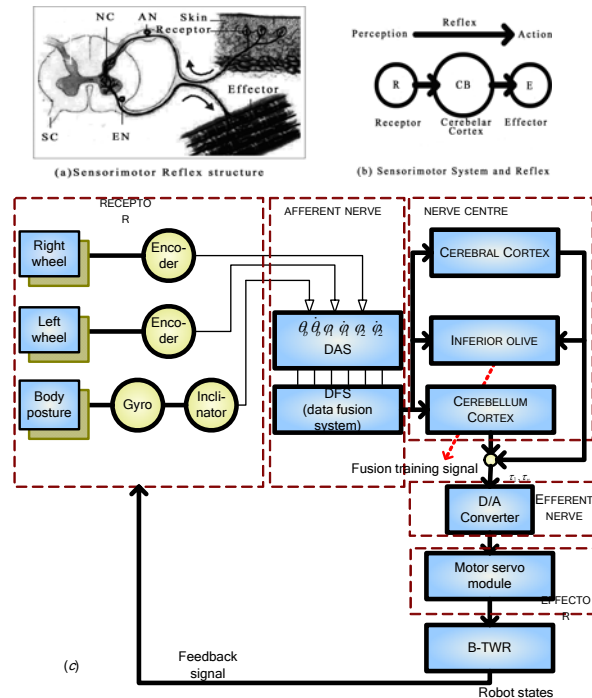
Figure 5 Physiological principles of sensorimotor and the bionic control structure for B-TWR. SC, spinal cord; EN, efferent nerve; NC, central nervous; AN, afferent nerve; DAS, data acquisition system.
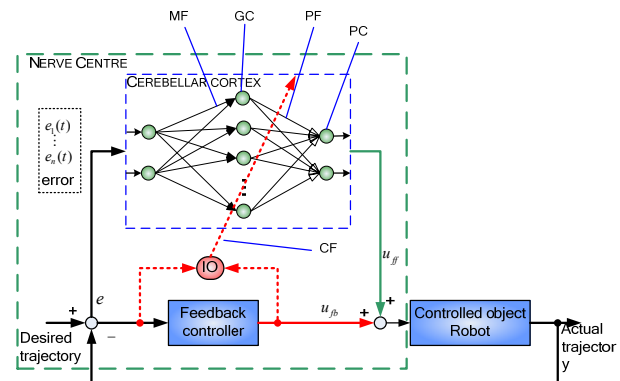
Figure 6 NNAC based on feedback-error-learning scheme. MF, mossy fibres; GC, granule cells; PF, parallel fibres; PC, purkinje cells; CF, climbing fibres; IO, inferior olive;

Each part of the 3-layers network is corresponded with cerebellum. The weight connections from input layer to hidden layer denote the Mossy fibres (MF). The weight connections from hidden layer to output layer denote Parallel fibres (PF). The hidden layer cells denote Granule cells (GC). The output layer cells denote Purkinje cells (PC). The error signals are delivered to cerebellar cortex (CBC) through Climbing fibres (CF)(red dashed line). In the cerebellum, climbing fibres originating from the inferior olive convey error signals that drive a long-term depression (LTD)-based learning process at parallel-fibre–Purkinje-cell synapses. Then the network can be modified and realize the on-line adaptive control process.

As in literature [19], the NNAC is updated so as to realize the inverse model of controlled plant after learning. The parameter update law is

$$\frac{dw}{dt} = -\eta \left(\frac{\partial u_{ff}}{\partial w} \right)^{\mathrm{T}} \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right) \tag{3}$$

where η is the learning rate of neural network and $0 < \eta < 1$, $0 < \mu < 1$, $k^{\mathrm{T}} \geq 0$. For the two wheeled robot system, we have $\mathbf{e} \in \mathbb{R}^{n \times 1}, u_{ff} \in \mathbb{R}^{r \times 1}, u_{fb} \in \mathbb{R}^{r \times 1}, k \in \mathbb{R}^{n \times r}$.

Therefore the aim of the training algorithm is to adjust the weights of the network so as to minimize the learning error:

$$\Xi = \frac{1}{2}(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e})^{\mathrm{T}}(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e}) \tag{4}$$

For our robot, the cerebellar cortex network is designed with three layers. Input layer with six nerve cells, hidden layer with six nerve cells, and the output layer with two nerve cells which are the torques of two-wheels make up of the neural network.

B. On-line training algorithm for B-TWR

The proposed on-line training algorithm is given in Step 1 to Step 5:

ON-LINE TRAINING ALGORITHM

Step 1: Initialize the parameters: choose weights w0, bias weights b0 randomly, $\mu=0.56$, $\eta=0.5$, $k=[1,1,1,1,1,1;1,1,1,1,1,1]'$.

Step 2: Repeat for each new input training pattern $\mathbf{e}$ and u_{fb}.

Step 3: Calculate the parameter update law shown in formula (3).

Step 4: Update the weights according to the parameter update law.

Step 5: Calculate the output of the NNAC to control the robot.

The neural network can be expressed as follows:

$$\begin{cases} N_1^{6 \times 1} = W_1^{6 \times 6}\mathbf{e} + b_1^{6 \times 1} \\ A_1^{6 \times 1} = \tanh(N_1^{6 \times 1}) \\ N_2^{2 \times 1} = W_2^{2 \times 6}A_1^{6 \times 1} + b_2^{2 \times 1} \\ A_2^{2 \times 1} = N_2^{2 \times 1} \end{cases}$$

N_1: the input of hidden layer
A_1: the output of hidden layer
N_2: the input of output layer
A_2: the output of output layer
W_1: the weights from input to hidden layer
b_1: the bias weights of hidden layer
W_2: the weights from hidden to output layer
b_1: the bias weights of output layer
We can conclude formula (5) as in Appendix A.

$$\begin{cases} \dfrac{dW_2}{dt} = -\eta \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right) A_1^{\mathrm{T}} \\ \dfrac{d\mathbf{b}_2}{dt} = \eta \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right) \\ \dfrac{dW_1(i,j)}{dt} = -\eta \sum_{j=1}^{6} (1-A_1^2(j))e(i)D(j) \\ \dfrac{db_1(j)}{dt} = \eta \sum_{j=1}^{6} (1-A_1^2(j))D(j) \end{cases} \tag{5}$$

$$(i=1\cdots6; j=1\cdots6)$$

$$D = W_2^{\mathrm{T}} \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right)$$

For our two-wheeled self-balancing robot, the varieties are following.

$$\mathbf{e} = \begin{bmatrix} 0-\theta_b & 0-\dot{\theta}_b & 0-\omega_l & 0-\omega_r & 0-\varphi_l & 0-\varphi_r \end{bmatrix}^{\mathrm{T}}$$

$$\omega_l = \dot{\varphi}_l; \omega_r = \dot{\varphi}_r$$

The initial incline angle is 9 degree($\pi/20$ rad), and other variables are all zero.

Feedback controller is a feedback gain matrix:

$$K\_back = \begin{bmatrix} 100 & 40 & 1 & 4 & 0 & 3 \\ 100 & 40 & 4 & 1 & 3 & 0 \end{bmatrix}$$

C. Simulation Experiment

If we carry out the control only using feedback controller, we can see that the result is convergent. Nevertheless, if we use the same feedback controller and add a cerebellar cortex as a feedforward controller, we would get a better result. The simulation result is shown in Figure 7.

From Figure 7 we can see that the result with NNAC control is better than the feedback control result. The vibration amplitude is smaller.
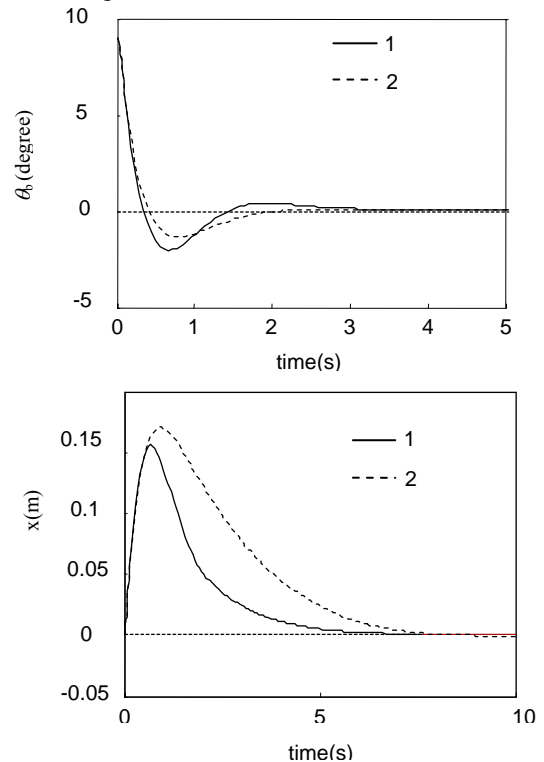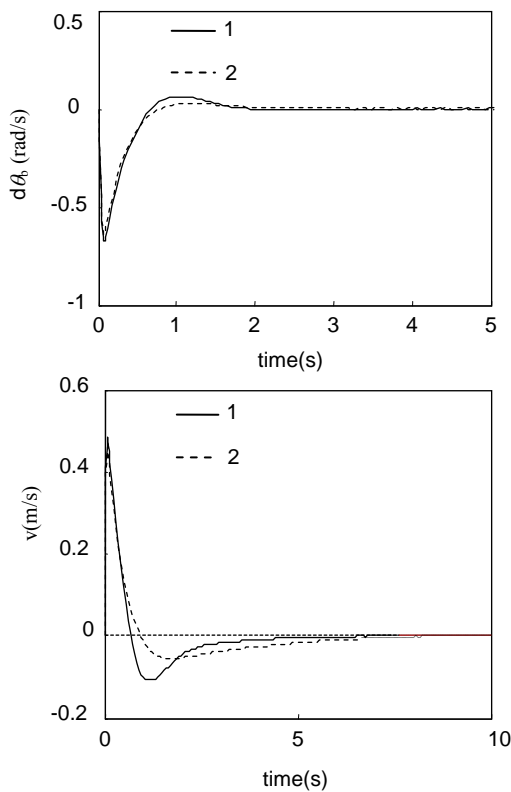
Figure 7   Simulation results with On-line NNAC and feedback controller. The real line (1) denotes the control result with only the feedback controller, while the dashed line (2) denotes the result with on-line NNAC.

Now let's detect its anti-interference performance. A step signal with amplitude 1 was given to the body of the robot. The experiment results are shown in Figure 8. We can find that the robot can generate a little angle to balance the step disturb and become stable at a certain point within a period of time.
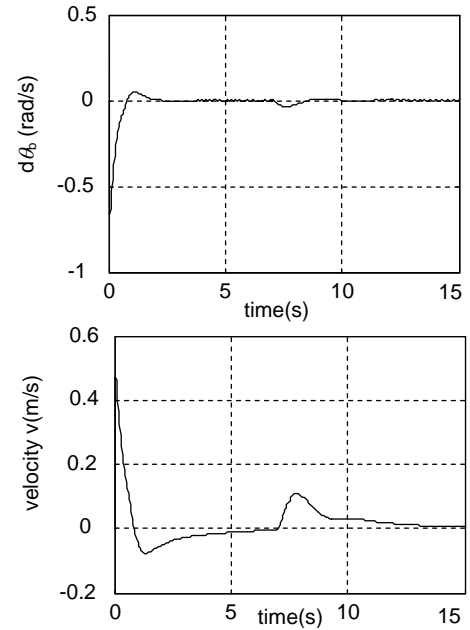
Figure 8   Simulation results using On-line NNAC with a step disturb.

Figure 9 shows the different result between the On-line NNAC and traditional feedback control with step disturb.
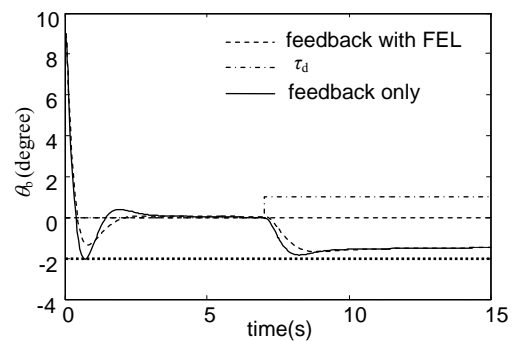
Figure 9   Simulation results respectively using On-line NNAC. and feedback controller with a step disturb.
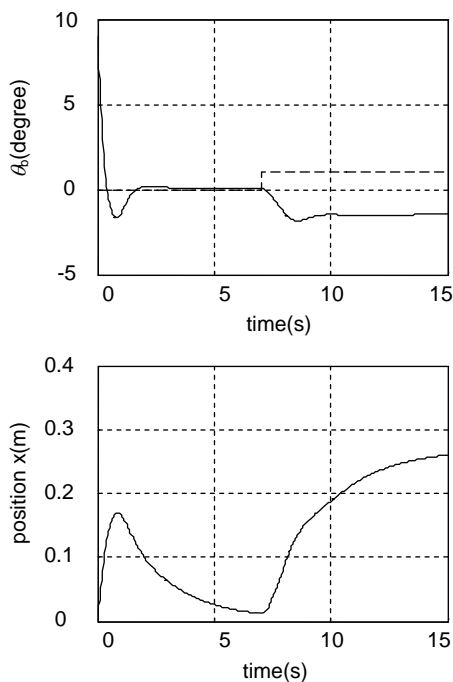
When the mass of robot body have some change. Through the simulation experiments, the mass M changes from 10kg to 30kg. Result analysis of comparative experiment can be made from Figure 10. In contrast to Figure 7, the traditional feedback control appears obvious vibrating when the mass of body became to 30kg, while there is not obvious change with On-line NNAC method (in Figure 10). These phenomenons illustrate that cerebellar cortex play an important part in the accurate balancing control and can effectively eliminate the shaking due to the system changing.
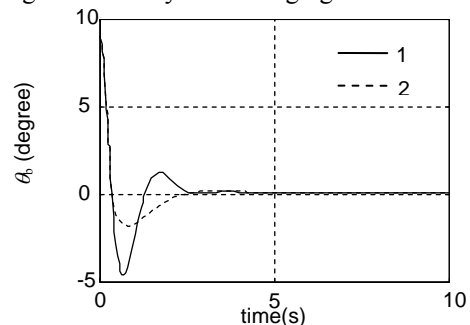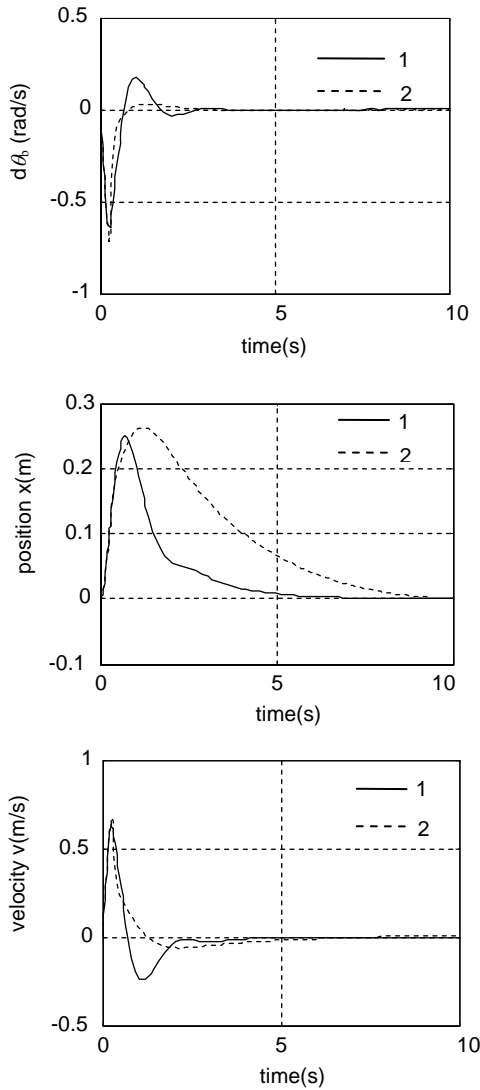
Figure 10 Simulation results with feedforward controller and feedback controller when M was changed to 30kg.

TABLE II
CONTRASTIVE ANALYSIS

M	Method	Overshoot σ	Response time t_s
10kg	On-line NNAC	14.82%	1.4754s
	feedback	22.60%	1.9477s
30kg	On-line NNAC	20.38%	1.8883s
	feedback	51.80%	2.2182s

From Table II, we can find that, no matter when mass is 10kg or 30kg, the overshoot and the response time with on-line NNAC method is little than with traditional feedback control. Especially, when the mass changes from 10kg to 30kg, the response all became slow to a certain extent. However, the overshoot with our proposed method changed from 14.82% to 20.38%, while the overshoot with traditional method had a larger change from 22.60% to 51.80%. From this point, we can conclude that, the on-line NNAC imitating the sensorimotor system and CNS can adap environment change and have an stable control effect.

V. CONCLUSION

On the Neurophysiological concept, a biomimic control structure was designed based on the sensorimotor system and the central nervous system of human brain. The Research focused on the cerebellar cortex functioning in supervised training. An on-line NNAC method based on cerebellar model with feedback-error-learning on the neurophysiological mechanism was proposed to control the balancing two-wheeled robot which is a bionic self-balancing plant imitating human being. The cerebellar cortex as a forward controller was an on-line neural network adaptive controller and reflected the cerebellar function in balancing control problem. Through three representative simulation experiments including balancing control experiment, disturb experiment and mass changed experiment, we can see that the robot can be balanced in fixed position well by the biomimic control structure, and can get a better result comparing the traditional feedback control through the contrast of overshoot and response time.

APPENDIX A. PARAMETER UPDATE DEDUCING

$$\frac{\partial u_{ff}(1)}{\partial W_2(1,j)} = \frac{\partial N_2(1)}{\partial W_2(1,j)} = \frac{\partial u_{ff}(2)}{\partial W_2(2,j)} = \frac{\partial N_2(2)}{\partial W_2(2,j)} = A_1(j)$$

$$\frac{\partial u_{ff}(2)}{\partial W_2(1,j)} = \frac{\partial N_2(2)}{\partial W_2(1,j)} = 0; \frac{\partial u_{ff}(1)}{\partial W_2(2,j)} = \frac{\partial N_2(1)}{\partial W_2(2,j)} = 0$$

$$\frac{dW_2(1,j)}{dt} = -\eta \left(\frac{\partial u_{ff}(1)}{\partial W_2(1,j)} \quad 0 \right)^{\mathrm{T}} \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right)$$

$$= -\eta \left(A_1(j) \quad 0 \right)^{\mathrm{T}} \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right)$$

$$\frac{dW_2(2,j)}{dt} = -\eta \left(0 \quad \frac{\partial u_{ff}(2)}{\partial W_2(2,j)} \right)^{\mathrm{T}} \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right)$$

$$= -\eta \left(0 \quad A_1(j) \right)^{\mathrm{T}} \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right)$$

$$(j = 1 \cdots 6)$$

$$\frac{d\mathbf{b}_2(1)}{dt} = -\eta \left(-1 \quad 0 \right)^{\mathrm{T}} \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right)$$

$$\frac{d\mathbf{b}_2(2)}{dt} = -\eta \left(0 \quad -1 \right)^{\mathrm{T}} \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right)$$

$$\frac{dW_1}{dt} = -\eta \left(\frac{\partial u_{ff}(1)}{\partial W_1} \quad \frac{\partial u_{ff}(2)}{\partial W_1} \right) \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right)$$

$$\frac{\partial u_{ff}(1)}{\partial W_1(i,j)} = \frac{\partial u_{ff}(1)}{\partial A_1} \frac{\partial A_1}{\partial N_1} \frac{\partial N_1}{\partial W_1(i,j)}$$

$$= \sum_{j=1}^{6} (1 - A_1^2(j)) W_2(1,j) \frac{\partial N_1(j)}{\partial W_1(i,j)}$$

$$= \sum_{j=1}^{6} (1 - A_1^2(j)) W_2(1,j) e(i)$$

In the same way,

$$\frac{\partial u_{ff}(1)}{\partial W_1(i,j)} = \sum_{j=1}^{6} (1 - A_1^2(j)) W_2(2,j) e(i)$$

$$D = W_2^{\mathrm{T}} \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right)$$

So that

$$\frac{dW_1(i,j)}{dt} = -\eta \sum_{j=1}^{6} (1 - A_1^2(j)) e(i) D(j)$$

$$\frac{d\mathbf{b}_1}{dt} = -\eta \left(\frac{\partial u_{ff}(1)}{\partial \mathbf{b}_1} \quad \frac{\partial u_{ff}(2)}{\partial \mathbf{b}_1} \right) \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right)$$

$$\frac{\partial u_{ff}(1)}{\partial \mathbf{b}_1(j)} = \frac{\partial u_{ff}(1)}{\partial A_1} \frac{\partial A_1}{\partial N_1} \frac{\partial N_1}{\partial \mathbf{b}_1(j)} = -\sum_{j=1}^{6} (1 - A_1^2(j)) W_2(1,j)$$

In the same way,

$$\frac{\partial u_{ff}(1)}{\partial W_1(i,j)} = -\sum_{j=1}^{6} (1 - A_1^2(j)) W_2(2,j)$$

$$D = W_2^{\mathrm{T}} \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right)$$

So that

$$\frac{d\mathbf{b}_1(j)}{dt} = \eta \sum_{j=1}^{6} (1 - A_1^2(j)) D(j)$$

$$\begin{cases} \dfrac{dW_2}{dt} = -\eta \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right) A_1^{\mathrm{T}} \\[2mm] \dfrac{d\mathbf{b}_2}{dt} = \eta \left(\mu u_{fb} + (1-\mu)k^{\mathrm{T}}\mathbf{e} \right) \\[2mm] \dfrac{dW_1(i,j)}{dt} = -\eta \sum_{j=1}^{6} (1 - A_1^2(j)) e(i) D(j) \\[2mm] \dfrac{d\mathbf{b}_1(j)}{dt} = \eta \sum_{j=1}^{6} (1 - A_1^2(j)) D(j) \end{cases} \quad (5)$$

$$(i = 1 \cdots 6; \ j = 1 \cdots 6)$$

ACKNOWLEDGMENT

REFERENCES

[1] K Pathak, J Franch, S K Agrawal. Velocity and Position Control of a Wheeled Inverted Pendulum by Partial Feedback Linearization [J]. IEEE Trans. on Robotics (1552-3098), 2005, 21(3): 505-513.

[2] A Blankespoor, R Roemer. Experimental verification of the dynamic model for a quarter size self-balancing wheelchair. [C]// Proc. ACC. Boston: ACC, 2004: 488-492.

[3] F Grasser, A D'Arrigo, S Colombi, A Rufer. Joe: A mobile, inverted pendulum [J]. IEEE Transactions on Industrial Electronics (S0278-0046), 2002, 9(1):107-114.

[4] NBot Balancing Robot, a two wheel balancing robot. (2003). [Online]. Available:http://www.geology.smu.edu/~dpa-ww/robo/nbot/index.html.

[5] DING Xue-ming; ZHANG Pei-ren; YANG Xing-ming; XU Yong-ming. Motion Control of Two-wheel Mobile Inverted Pendulum Based on SIRMs [J]. Journal of System Simulation (in Chinese).2004,16(11):2618-2621.

[6] CHEN Xing; WEI Heng-hua; ZHANG Yu-bin. Modeling of Dual-wheel Cart-Inverted Pendulum and Robust Variance Control[J]. Computer Simulation(in Chinese). 2006,23(3) : 263-266.

[7] Chih-Hui Chiu. The Design and Implementation of a Wheeled Inverted Pendulum Using an Adaptive Output

[8] Recurrent Cerebellar Model Articulation Controller [J].IEEE Transactions on Industrial Electronics. 2010, 57(5):1814-1822.

[8] Y. S. Ha and S. Yuta, "Trajectory tracking control for navigation of the inverse pendulum type self-contained mobile robot," Robot. Autonom. Syst., vol. 17, no. 1/2, pp. 65–80, Apr. 1996.

[9] S. Jung and S. S. Kim, Control experiment of a wheel-driven mobile inverted pendulum using neural network. IEEE Transactions on Control Systems Technology. 2008, 16(2): 297–303.

[10] Kawato, M., Furawaka, K., Suzuki, R. A hierarchical neural network model for the control and learning of voluntary movements. Biol Cybern, 1987,56, pp. 1-17.

[11] Kawato, M. Feedback-error-learning neural network for supervised motor learning. In Advanced neural computers (ed. R. Eckmiller) 1990, pp. 365–372.

[12] M. Kawato, H. Gomi, A computational model of four regions of the cerebellum based on feedback error learning. Biol. Cybern. 69 (1992)95–103.

[13] Jun Nakanishi, Stefan Schaal. Feedback error learning and nonlinear adaptive control. Neural Networks. 2004,17:1453-1465.

[14] James C.Houk and Steven P.Wise. Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: their role in planning and controlling action. Cereb.Cortex. 1995,5:95–110.

[15] Doya, K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Networks. 1999, 12: 961–974. (doi:10.1016/S0893-6080(99)00046-5)

[16] J.C Houk, C Bastianen,etc. Action selection and refinement in subcortical loops through basal ganglia and cerebellum. Phil. Trans. R. Soc. B. 2007,362:1573-1583(doi: 10.1098/rstb.2007.2063).

[17] Houk, J. C. Agents of the mind. Biol. Cybern. 2005,92:427–437. (doi:10.1007/s00422-005-0569-8)

[18] Ito, M. Neurophysiological aspects of the cerebellar motor control system[J] International journal of neurology, 1970.7 (2), pp. 162-176.

[19] Ruan, Xiaogang ; Ding, Mingxiao; Gong, Daoxiong; Qiao, Junfei On-line adaptive control for inverted pendulum balancing based on feedback-error-learning[J]. Neurocomputing, v 70, n 4-6, January, 2007, p 770-776.

Xiaogang Ruan, born in Sichuan Province, China in 1958, and received Ph.D. degree from Zhejiang University in 1992, Hangzhou, China. Now he is a professor of the Beijing University of Technology, and he is also as a director of IAIR(Institute of Artificial Intelligent and Robotics). His research interests include Automatic Control, Artificial Intelligence, and Intelligent Robot.

Jing Chen, born in Hebei Province, China in 1984, and received Master's degree from BJUT in 2008, Beijing, China. Now she is a Doctor by research student in the School of Electronic Information and Control Engineering, Beijing University of Technology (BJUT). Her main research interest is Artificial Intelligence, Neuroscience, and Intelligent Robot.

Evaluation of Influence of Motorized Wheels on Contact Force and Comfort for Electric Vehicle

JIN Li-qiang,
State key laboratory of automotive simulation, Jilin University, Chanchun, China
e-mail: jinlq@jlu.edu.cn

SONG Chuan-xue
State key laboratory of automotive simulation, Jilin University, Chanchun, China
e-mail: songchx@126.com

WANG Qing-nian,
State key laboratory of automotive simulation, Jilin University, Chanchun, China
e-mail: wqn@jlu.edu.cn

Abstract—**Volatile oil prices and increased environmental sensitivity together with political concerns have moved the attention of governments, automobile manufacturers and customers to alternative power trains. From the actual point of view the most promising concepts for future passenger cars are based on the conversion of electrical into mechanical energy. In-wheel motors are an interesting concept towards vehicle electrification that provides also high potentials to improve vehicle dynamics and handling. Nevertheless in-wheel motors increase the unsprung mass worsening vehicle comfort and safety. All kinds of motorized-wheel structures are analyzed. For the unitary motorized wheel, since increasing the unsprung mass of vehicle, it adds adverse effects on the vehicle road holding performance and vehicle comfortableness. To decrease the adverse effect for unsprung mass increasing, a method by attaching the motor in the wheel through springs and dampers of exclusive use is brought out. It is verified that the vibration of EV with suspended-motor motorized-wheel is better than the EV with detached-motorized wheel. For high cost of suspended in-wheel motor, The effect of increasing of unsprung mass by using integral motor is analyzed by simulation with whole vehicle model. The more unsprung mass of integral motor deteriorates the vehicle comfort and vibration performance at low excitation frequency of road. But it is improved by using integral motor at higher excitation frequency of road, while unsprung mass is increased by using integral motor. The fluctuation amplitude of vibration at high frequency is much more big than at low frequency. In this view, the conclusion that in-wheel motor integrated into wheels is acceptable structural concept for electric vehicle with motorized wheels with lower cost than suspended in-wheel motor is achieved**

Index Terms— **Electric Vehicle; motorized wheels; road holding performance**

I. INTRODUCTION

Based on the increasing environment and air pollution, Electric Vehicle has become the new technique trend for its low emission and energy saving characteristics. However, recent EV is still uncompetitive in performance and cost compared with conventional vehicle. Application of motorized wheel in EV is a good way to downsize the cost and improve the vehicle performance[1,2]. In motorized wheels driving system, mechanical transmission is eliminated through integrated design of motor and wheel. Vehicles with this system could achieve the advantages that: (1) Vehicle mass would be downsized. Conventional transmission system takes a big part of vehicle mass, while the motor and controller are comparatively much smaller. (2) Cost would be downsized. The R&D cost of high performance vehicle is considerable. Motorized wheel driving system takes the utility of power electronics for the power transmission which is based on the wire transmit and control structure. This kind of transmission, compared with the conventional mechanical transmission, is much easier to be modulized, cheaper for the R&D cost and simpler in vehicle design. (3) Vehicle comfortableness would be improved. Because the mechanical transmission is eliminated, there is no noise and vibration generated. Chassis height could be reduced to provide more passenger room. (4) Vehicle safety would be improved. The driving torque could be controlled independently in motorized wheel system, so it is much flexible and comfortable to realize the vehicle dynamic control including ABS, TCS and VSC solely by controlling the driving motor[2,3]. Because the motor has a faster response than mechanical transmission, vehicle with motorized wheel has better brake, dynamic and safety[4,5,6].

In this paper, based on the vertical vibration model built, the effects of structure of motorized wheels were analyzed on the comfort and road holding performance of Electric Vehicles, and then vehicle comfort and vibration performance is analyzed by simulation using whole vehicle model comparing integral motor structure with split motor structure. The solutions is brought out to enhance the comfort and road holding performance..

I.   STRUCTURE ANALYSIS

Fig.1~3 show the typical structure of motorized wheels. Fig.1 is the structure of dualmotor. It is modified from an induction motor or a synchronous motor. When vehicle running straight forward, the flux generated lies along the axial direction. Because the stator flux would be divided into two equal parts, which goes through the rotors on each side, torque and speed of one rotor would be the same as that of the other. When vehicle turning, under the same voltage frequency, the speed of outer wheel connected with its rotor (defined as rotor 1) will increase. As it gets closer to the synchronous speed, the current in rotor1 will decrease for the reactance; the slip frequency of rotor2 will increase, which will lead to the increasing of rotor winding's current and reactance. Thus, more flux flow into rotor1, making torque in rotor2 decrease. However, the torque delivery will cause a oversteering under high speed turning. This motor structure also takes much place in the middle, which will contribute to a bad comfort and trafficability[6,7].
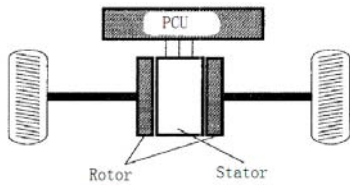
Figure 1.   Motorized wheel with dualmotor

(a)   (b)
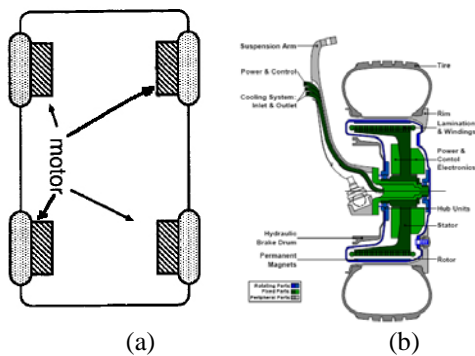
Figure 2.   Motorized Wheel with Integrated Motor
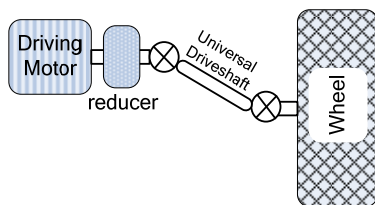
Figure 3.   Motorized Wheel Connected with Universal Driveshaft

Fig.2(b) shows a motorized wheel integrated with a motor in the wheel. Fig.2(a) is the 4WD vehicle powered by integrated-motorized wheels[8]. This configuration requires less mounting place, which is the ideal one for layout. The motor is designed into an in-wheel motor that can be easily placed in the wheel hub. However, this structure has some certain disadvantages: It is hard to

introduce a speed reducer, so the motor should provide adequate torque to drive the vehicle; Brake is hard to be installed in the wheel hub after motor is mounted; Meanwhile, the integrated structure will lead to a bad cooling, as a result of which, motor should have reliable performance allows to work under terrible conditions including high temperature and strong vibration. Moreover, adding in unsprung mass will deteriorate the comfortableness and road holding ability.

Fig.3 is the detached-motorized wheels, in which motor and speed reducer are connected with vehicle body. This kind of structure has smaller room for overall arrangement than that in Fig.2, but is much easier to realize the motor driving from little modification on the conventional vehicle. With this structure, motor could work under good conditions and provide more reliable performance without any amplification of un sprung mass.

III EFFECTS OF THE STRUCTURE OF MOTORIZED WHEEL ON VEHICLE PERFORMANCE

Structure in Fig.1 should be redesigned for the uncomfortable of independent control of wheel's torque. It is just a kind of tentative structure that still has a long way to the real application. In the coming part, comparison and analyze would be taken on the comfortableness and road holding performance between detached-motorized wheels and integrated one.

A. Evaluation of comfortableness and road holding performance

It is very important to keep a good road holding performance of the tire so that it could provide adequate forces for vehicle controlling.

When the vehicle runs on the rough road, the increase of unsprung mass would delay the response. The amplitude of the vertical deformation of the tire and the distance between tire and road would increase, as a result of which the corresponding amplitude of road holding force and the vertical vehicle body acceleration transmitted from sprung mass would increase as well making the comfortableness deteriorate.

Lateral force of tire is nonlinear with the road holding force. Thus, the increase of amplitude holding force will result the decrease of lateral force[9], as is shown in Fig.4.
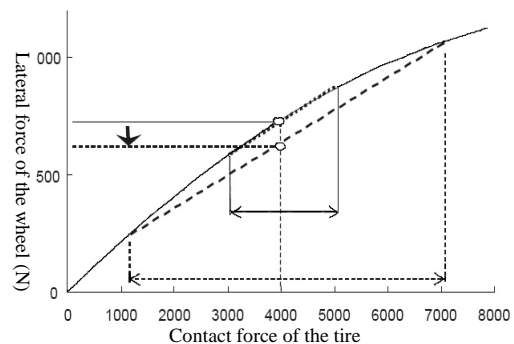

Figure 4.   Relationship between the road holding force and the lateral force of the tire

When the fluctuation of road holding force decreases, the lateral force will increase and lead to a great improvement of the road holding performance[10].

B. Vibration model of vehicle with motorized wheel

Independent suspension is usually used in the vehicle driven by motorized wheels. Thus, the 2-degree vibration model in Fig.5 can completely express the vertical vibration of vehicle.
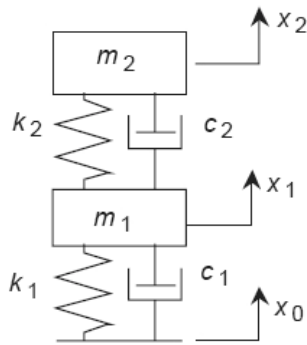

Figure 5.   1/4 vibration model of vehicle

In Figure5, x_0 is the road excitation; x_1 is the vibrating displacement of unsprung mass; x_2 is the vibrating displacement of sprung mass(vehicle body); k_1, c_1 are the vertical elastic constant and damping constant of the tire; k_2, c_2 are the elastic constant and damping constant of the suspension; m_1 is the unsprung mass of the vehicle. For the structure in Fig.2, unsprung mass includes the motor and the wheel, while for structure in Fig.3, it only includes the wheel; m_2 is the sprung mass of the vehicle. The vibration equations of the system are:

$$m_1 \ddot{x}_1 + F_n + F_t = 0$$

$$m_2 \ddot{x}_2 + c_2 \left(\dot{x}_2 - \dot{x}_1 \right) + k_2 \left(x_2 - x_1 \right) = 0$$

$$F_t = k_1 \left(x_1 - x_0 \right) + c_1 \left(\dot{x}_1 - \dot{x}_0 \right)$$

$$F_n = k_2 \left(x_1 - x_2 \right) + c_2 \left(\dot{x}_1 - \dot{x}_2 \right)$$

where： F_t －fluctuation of road holding force；

F_n －vehicle body's vibration inputs；

C. performance of road holding force and body comfort for vehicle driven by motorized wheels with high stiffness

Simulation model was built based on the vibration model mentioned before using Simulink software. Sine excitation signal was provided with the amplitude 5mm and frequency increasing in step. Simulation analysis of frequency of the road holding force, vehicle body acceleration and vehicle body vertical force fluctuation is

fulfilled based on the split motor and integral motor. Comparative Results are shown in Fig.5 Fig.6 and Fig.7. Parameters of mass and damping coefficient used in simulation are shown in Table I. the suspension stiffness used in simulation is shown in table II. It can be seen that the stiffness of suspension and tire is higher than that used in car normally.

Table I. Simulation Parameter

Parameter	Detached-motorized Wheel	Integrated-motorized Wheel
m_1 (kg)	30	60
m_2 (kg)	360	330
c_1 (N/(m/s))	50	50
c_2 (N/(m/s))	1496	1496

Table II The stiffness of the suspension and tire as follows:

Parameter	Detached-motorized Wheel	Integrated-motorized Wheel
k_1 (N/m)	360000	360000
k_2 (N/m)	32000	32000
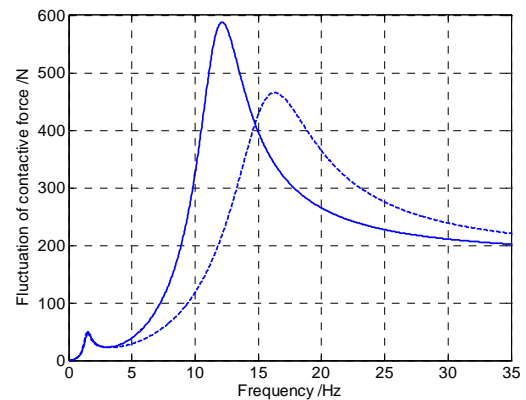
Figure 6.   Comparison of the Amplitude of the road holding Force

The comparison of the road holding performance of Electric vehicle with split motor or integral motor is shown in Fig.5. The EV with split motor has less unsprung mass and there is less fluctuation of the road holding force than integral motor. The differential of the fluctuation amplitude is about 120N.
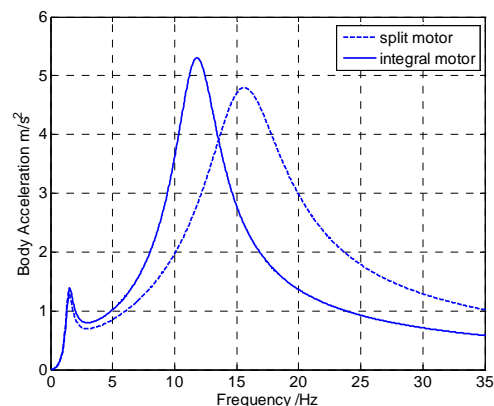

Figure 7.   Comparison of the vehicle body Acceleration

The comparison of the vehicle body acceleration is shown in Fig.7. the vehicle with integral motor has higher acceleration than the one with split motor because of more unsprung mass at the resonant frequency（10-15Hz）. The resonant frequency of integral motor is lower than split motor. But the acceleration of EV with integral motor would be lower than split motor when the resonant frequency is exceeded.
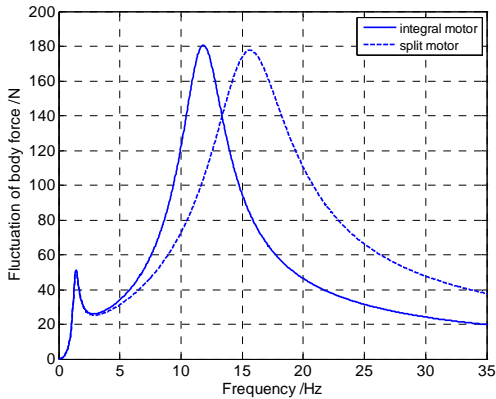

Figure 8.   Comparison of Fluctuation of body vertical force

The comparison of fluctuation of body vertical force is shown in Fig.7. There're only small differences between the integral motor and split motor. But the resonant frequency of EV with integral motor is lower than split motor.

From figures above, integrated-motorized wheel has a larger unprung mass, so in the resonant frequency（10-15Hz）, the road holding force fluctuates much more than using detached-motorized wheel. The lateral force of tire decreases accordingly. Therefore, conclusions can be drawn that high speed stability of vehicles driven by integrated-motorized wheel is worse than that of vehicles driven by detached-motorized wheel. Vehicle body acceleration is the important critical reference for comfortableness. From Figure 3, comfortableness of EVs with detached-motorized wheel is better than that with integrated-motorized wheel below resonant frequency.

D. performance of road holding force and body comfort for vehicle driven by motorized wheels with low stiffness

Simulation is carried out by same model and same simulation parameter but the k_1 and k_2, it is shown in Tab.III

Table III The stiffness of the suspension and tire

Parameter	Detached-motorized Wheel	Integrated-motorized Wheel
k_1 (N/m)	180000	180000
k_2 (N/m)	16000	16000

For low stiffness of suspension and tire, The fluctuation amplitude of road holding force of EV is decreased remarkably comparing with high stiffness. As is shown in Fig.8, the EV with split motor has less fluctuation amplitude than that with integral motor. The differential between them is about 40N.
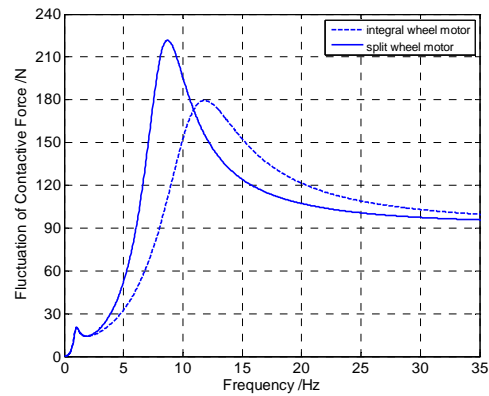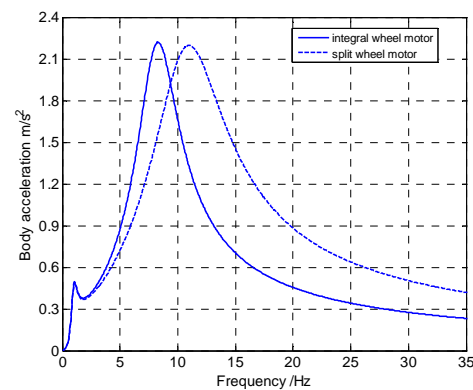

Figure 9.   Fluctuation of contactive force

Figure 10. Comparison of body acceleration

At low stiffness of suspension and tire, the body acceleration is decreased dramatically too. As is shown in Fig.9, the max acceleration of body is subequal between the EV with integral motor and split motor.
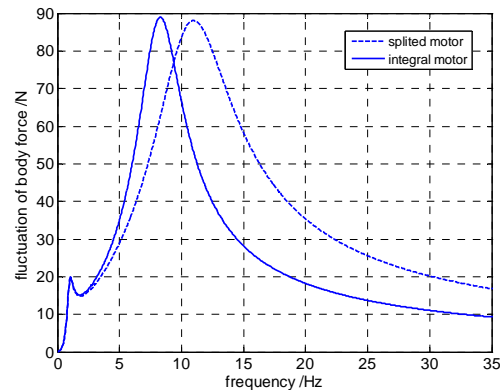

Figure 11. Fluctuation of body force

The fluctuation of body vertical force is shown in Fig.10. The max fluctuation amplitude of EV with integral motor and split motor has less difference.

Based on forgoing analysis, the increasing of unsprung mass for electric vehicle with integrated-motorized wheel has less influence on road holding performance and vehicle comfort when the stiffness of suspension and tire is low.

IV VIBRATION ANALYSIS OF SUSPENDED-MOTOR MOTORIZED WHEEL

The same as the engine, suspension of motor used in connection with wheel can downsize the vibration degree of unsprung mass.

A. Vibration model of EV with suspended-motor motorized wheel

Fig.11 shows the model of suspended motor connected with the wheel. Because the motor is suspended up, the system becomes the 3-degree model in which the motor has its independent vibration degree. In the model, k_3, c_3 are the elastic constant and damping of the suspension between the in-wheel motor and wheels. m_3 is the mass of in-wheel motor. x_3 is the vertical displacement of in-wheel motor. F_d is the force acting on the motor. Other parameters are defined the same as in Fig.5. Here the m_1 does not include the motor mass.
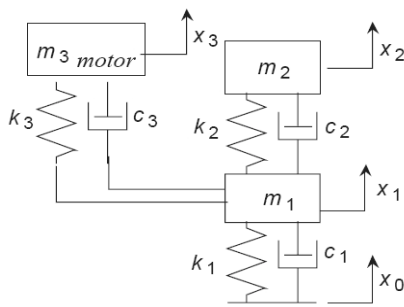
Figure 12. 1/4 vibration model of vehicle with suspended-motor motorized wheel

$$m_1 \ddot{x}_1 + F_n + F_t + F_d = 0$$

$$m_2 \ddot{x}_2 + c_2 \left(\dot{x}_2 - \dot{x}_1 \right) + k_2 \left(x_2 - x_1 \right) = 0$$

$$m_3 \ddot{x}_3 + c_3 \left(\dot{x}_3 - \dot{x}_1 \right) + k_3 \left(x_3 - x_1 \right) = 0$$

$$F_t = k_1 \left(x_1 - x_0 \right) + c_1 \left(\dot{x}_1 - \dot{x}_0 \right)$$

$$F_n = k_2 \left(x_1 - x_2 \right) + c_2 \left(\dot{x}_1 - \dot{x}_2 \right)$$

$$F_d = k_3 \left(x_1 - x_3 \right) + c_3 \left(\dot{x}_1 - \dot{x}_3 \right)$$

B. Analysis of road holding performance and comfort

Simulation model were made using simulink software. After providing the model the same excitation signals indicated before, simulation analysis was fulfilled for the frequency of the road holding force fluctuation and the vehicle body acceleration. In simulation, $m_3 = 20$kg, $k_3 = 32000$ (N/m), $c_3 = 1000$ (N/(m/s)), other

parameters are shown in Table.I and Tab.III. Results are shown in Fig.13 and Fig.14.

From Fig.13, the road holding performance was improved a lot by the suspending of in-wheel motor. The amplitude near the resonant frequency had also been downsized a lot compared with before. The peak value of road holding force fluctuation was smaller than that of detached type.
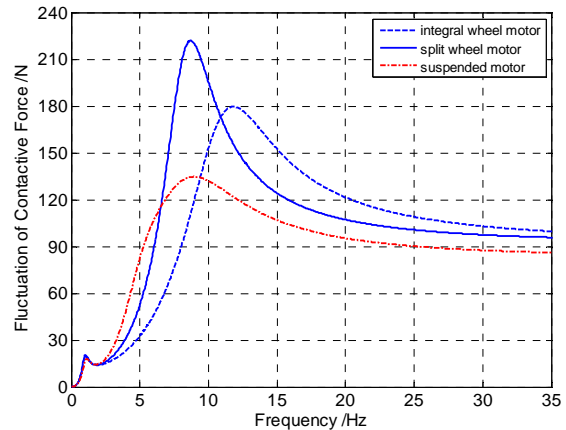

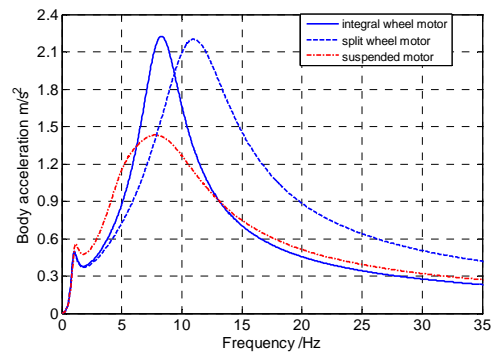Figure 13. road holding performance by Suspended-motor

Figure 14. Vehicle Body Acceleration with Suspended-motor

From Fig.14, it can be seen that the vehicle body acceleration using suspended-motor motorized wheels is much smaller than that using integrated-motorized wheel and that using detached-motorized wheels.
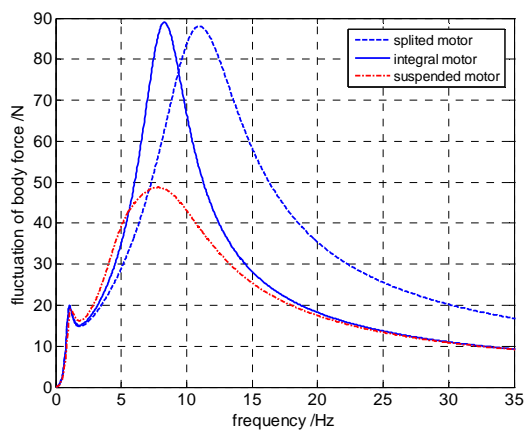

Figure 15. Fluctuation of body force with suspended motor

From Fig.15, the body vertical force fluctuation amplitude was decreased a lot by the suspending of in-wheel motor. The amplitude near the resonant frequency had also been downsized a lot compared with before.

Based on the analysis above, the adverse effects on vehicle's road holding performance and comfortableness caused by the increasing of unsprung mass for EV with integrated-motorized wheels can be improved remarkably by suspending of in－wheel motor The stiffness and damping of suspended part should match the ones of the wheel and the suspension.

V. SIMULATION WITH WHOLE VEHICLE MODEL

A. Vehicle body model

The whole vehicle model is shown in Fig.16.
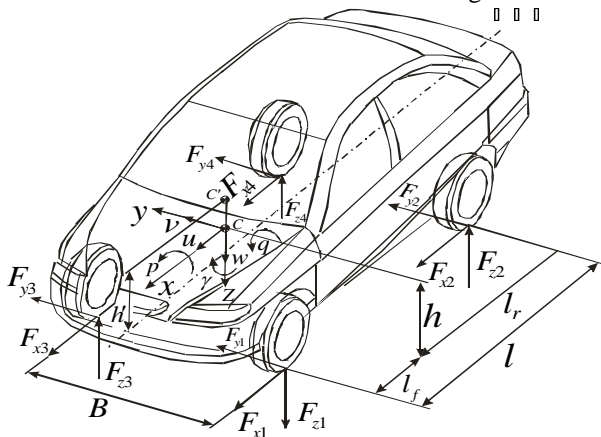
Figure 16. vehicle body model

To make comprehensive analysis for performance of electric vehicle with motorized wheels, the six freedom is included in body model. u , v , w is longitudinal speed, side speed and vertical velocity of sprung mass respectively. r , p , q is yaw velocity, roll velocity and pitch velocity of sprung mass. And ϕ , θ is roll angle and pitch angle of sprung mass. ψ is the yaw angle of vehicle.

F_{xi}, F_{yi}, F_{zi} ($i=1,2,...4$) is the force acting on the vehicle by tire along coordinate axis. C is the center of vehicle mass. h is the height of vehicle mass center, C' is the height of sprung mass, h' is the distance from center of sprung mass to roll axis. B is the track width，l is the wheelbase. l_f, l_r is the distance from mass center to front axle and to rear axle respectively.

B. Suspension model

The action point and direction of the force between the body and wheels is decided by suspension model. To describe the influence on vehicle performance for increasing unsprung mass by using motorized wheels, the suspension model is made as shown in Fig.17[8]. The

C' is the center of sprung mass, K_{ui}, C_{ui} ($i=1,2,...4$) are the vertical stiffness and damping coefficient of the tires. K_{si}, C_{si} ($i=1,2,...4$) are the stiffness and coefficient of suspensions, Z_{ui} is the height of unsprung mass, Z_{ri} is the roughness of the road. Z_s is the height of sprung mass.
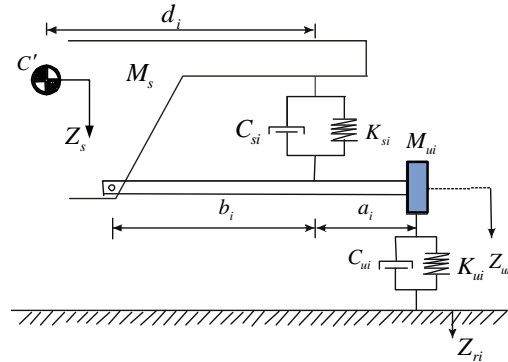

Figure 17. suspension model

The detailed description inclusive of the dynamics equation of the vehicle model is presented in literature 11.

C. Step response simulation

A condition that vehicle is driving on a road with a jump is simulated. The sudden change for the height of jump is shown in Fig.18. The road height changes to 0.1m steeply at 6s during vehicle driving and decreases to 0 m from 6.5s quickly.
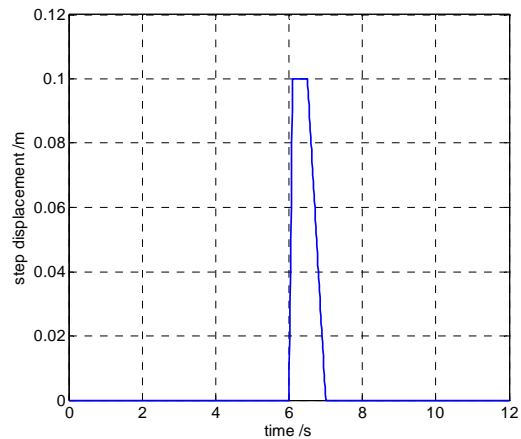

Figure 18. Jump during driving

The comparative result of the vertical acceleration of vehicle mass center between the vehicle with split in-wheel motor and integral in-wheel motor is shown in Fig.19. It is can bee seen from result that the acceleration of vehicle with split motor is bigger than with integral motor when the vehicle driving on the jump. The acceleration of wheel center of vehicle with split motor is bigger than with integral motor too, as is shown in Fig.20. It can be concluded that the vibration is improved when driving on the jump for increasing unsprung mass by integral in-wheel motor.
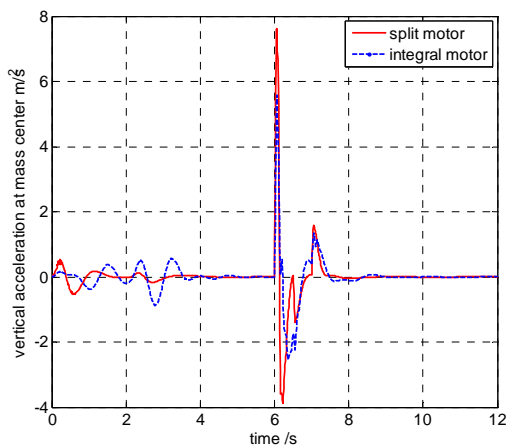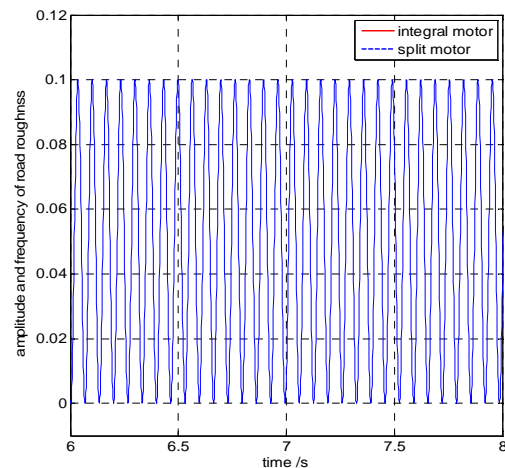
Figure 19. Vertical acceleration of vehicle mass center

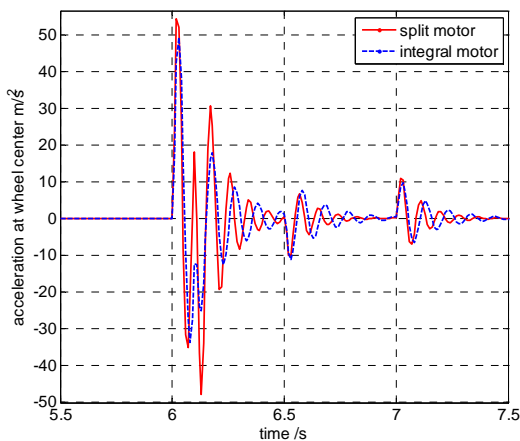

Figure 21. Road roughness of sine with 15Hz
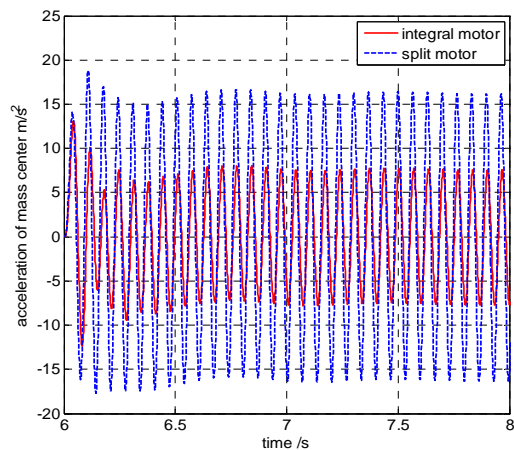

Figure 20. Vertical acceleration of wheel center

Figure 22. Vertical acceleration of vehicle mass center

D. Driving on the rough road with sine of 15hz

The road roughness during simulation is shown in Fig.21.

The comparative acceleration of vehicle mass center with split motor and with integral motor is shown in Fig.22. The vertical acceleration of vehicle mass center is decreased by using integral motor comparing with split motor when driving on the rough road with with sine of 15Hz. Similarly, shown in Fig.23, the vertical acceleration of wheel center is decreased too by using integral motor. And so the vehicle comfort performance and road holding performance is improved by increasing unsprung mass for using integral motor at rough road of sine with 15Hz.

As shown in Fig.6 and Fig.7, the frequency of 15Hz is bigger than the resonant frequency of vehicle and wheel vibration system. And the vertical acceleration of vehicle mass center with integral motor is less than with split motor at this frequency. The same result is achieved by simulation with quarter vehicle model and whole vehicle model.
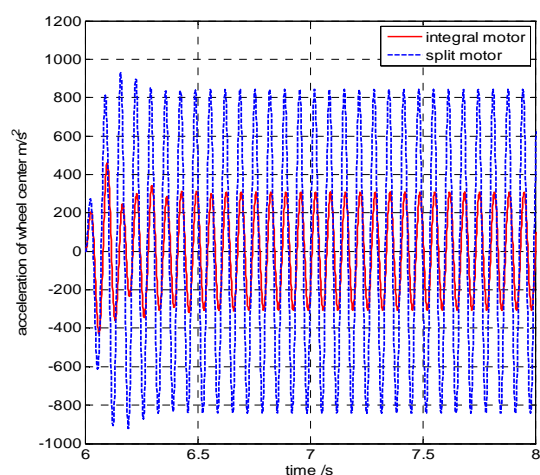

Figure 23. Vertical acceleration of wheel center

E. Driving on the rough road with sine of 10Hz

The road roughness during simulation is shown in Fig.24.

The acceleration of vehicle mass center with split motor and with integral motor is shown in Fig.25. the vertical acceleration of vehicle mass center is increased obviously by using integral motor comparing with split motor when driving on the rough road with sine of 10Hz. Similarly, the vertical acceleration of wheel center with integral motor is bigger than with split motor too, as shown in Fig26. And so the vehicle comfort performance and road holding performance is worsen by increasing unsprung mass for using integral motor at rough road of sine with 10Hz.

As shown in Fig.6 and Fig.7, the frequency of 10Hz is close to the resonant frequency of vehicle and wheel vibration system. And the vertical acceleration of vehicle mass center with integral motor is bigger than with split motor at this frequency. The same result is achieved by simulation with quarter vehicle model and whole vehicle model.
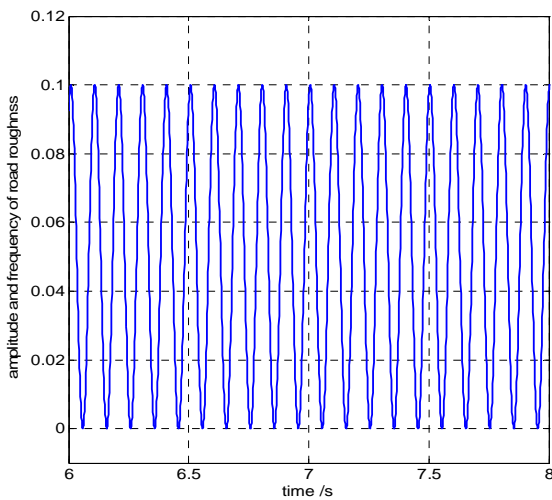

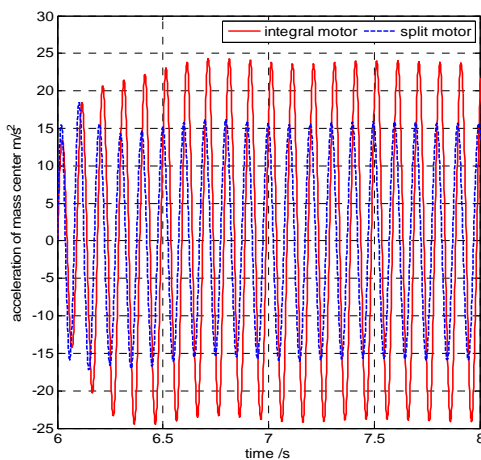Figure 24. Road roughness of sine with 10Hz

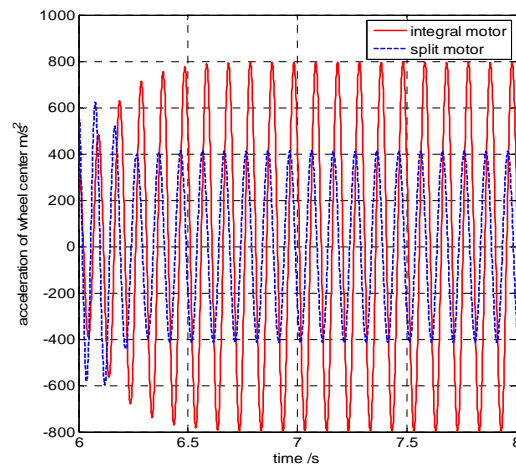Figure 25. Vertical accelertation of mass center

Figure 26. Vertical acceleration of wheel center

The increasing of unsprung mass by using integral motor deteriorates the vehicle comfort and vibration performance at low excitation frequency of road. But it is improved by using integral motor at higher excitation frequency of road, while unsprung mass is increased by using integral motor. The fluctuation amplitude of vibration at high frequency is much more big than at low frequency. In this view, the in-wheel motor that is integrated into wheels is acceptable structural concept for electric vehicle with motorized wheels with lower cost than suspended in-wheel motor.

V CONCLUSION

In this paper, structures of motorized wheels used in EV driving system were analyzed. Both detached-motorized wheel and integrated-motorized wheel have their merits and demerits. Detached-motorized wheel is easier to be mounted; Motor could works under good conditions; It is also easy to modified a conventional vehicle into a vehicle with motorized wheels. Integrated-motorized wheel requires much less place in layout. But the motor would works under terrible conditions, so high motor performance is required. To eliminate the adverse effect of increasing of unsprung mass, motor in integrated-motorized wheel should be suspended up. The vehicle with suspended in-wheel motor is much better than split motor and integral motor in comfort and road holding performance. Also, the stiffness and damping matching of suspended part should be taken with the wheel and suspension system. For high cost of suspended in-wheel motor, The effect of increasing of unsprung mass by using integral motor is analyzed. The more unsprung mass of integral motor deteriorates the vehicle comfort and vibration performance at low excitation frequency of road. But it is improved by using integral motor at higher excitation frequency of road, while unsprung mass is increased by using integral motor. The fluctuation amplitude of vibration at high frequency is much more big than at low frequency. In this view, the conclusion that in-wheel motor integrated into wheels is acceptable structural concept for electric vehicle with motorized

wheels with lower cost than suspended in-wheel motor is achieved.

Motorized wheel drive system used in EV can help downsize the overall mass, reduce the developing costs and improve the vehicle's performance. It is a kind of technology that of great potential of application.

REFERENCES

[1] Peter C. Cho, William Wylam, David Crecelius. Advanced Hybrid Electric Vehicle Propulsion Systems With Individual Wheel Brushless Traction Motors. SAE 2000-01-3110

[2] Se-Kyo Chung, Hyun-Soo Kim, Chang-Gyun Kim, and Myung-Joong Youn. A New Instantaneous Torque Control of PM Synchronous Motor for High-Performance Direct-Drive Applications. IEEE TRANSACTIONS ON POWER ELECTRONICS, VOL. 13, NO. 3, MAY 1998

[3] Rao S. Zhou and Fukuo Hashimoto. Highly Compact Electric Drive for Automotive Applications. SAE2004-01-3037

[4] JIN Li-qiang, WANG Qing-nian, SONG Chuan-xue. Simulation for Optimal PD Control for Dynamics Control of EV with Motorized Wheels. Journal of System Simulation，Vol.19 No.10，2007

[5] Y. Hori, Y. Toyoda, and Y. Tsuruoka. Traction control of electric vehicle: Basic experimental results using the test EV "UOT Electric March". IEEE Transaction. Industry. Application., Vol. 34, No. 5, pp. 1131–1138, 1998

[6] F.Carricchi, F.crescimbini, E.santini. Axial Flux Electromagnetic Differential induction motor. 'Electric Machines and drives' Conference Publication No.412,11-13 September 1995, IEE, P1－5

[7] Yee-Pien Yang _, Chun-Pin Lo.Current distribution control of dual directly driven wheel motors for electric vehicles. Control Engineering Practice. 16 (2008) 1285－1292

[8] Masayuki Terashima, Tadashi Ashikaga. Novel Motors and controllers for high Performance Electric Vehicle with Four In-Wheel Motors. IEEE Transactions on Industrial Industrial electronics, VOL. 44, NO. 1, FEBRUARY 1997. p28-38

[9] T. Ram Mohan Rao and G. Venkata Rao. PARAMETRIC SENSITIVITY ANALYSIS OF A HEAVY DUTY PASSENGER VEHICLE SUSPENSION SYSTEM. ARPN Journal of Engineering and Applied Sciences. NO. 8, VOL. 4, OCTOBER 2009

[10] Hyeongcheol Lee and Masayoshi Tomizuka. Adaptive Vehicle Traction Force Control for Intelligent Vehicle Highway Systems (IVHSS). IEEE Transactions on Industrial Industrial electronics, NO.1, VOL.50, FEB 2003 p37-47

[11] Jin Li-qiang , Wang Qing-nian , Song Chuan-xue.Dynamic simulation model and experimental validation for vehicle with motorized wheels. Journal of Jilin University (Engineering and Technology Edition). No.4, Vol.37, 2007, p745-750

JIN Li-qiang received the B.S. from College of Mechanical Engineering, Hebei University of Technology, Tianjin, China in 1999. And Received M.S. and Ph.D. degrees from the College of Automotive Engineering, Jilin University, Changchun, China, in 2002 and 2006, respectively.

He is currently a Assistant Professor with the College of Automotive Engineering, Jilin University. His research areas include modeling and control of vehicle systems, Advanced Technology and theory of Electric vehicle with motorized wheels.

SONG Chuan-xue received the B.S. M.S. and Ph.D. degrees from the College of Automotive Engineering, Jilin University, Changchun, China, in 1982 1985 and 1991, respectively.

He is currently a Professor with the College of Automotive Engineering, Jilin University. His research areas include Theory and Control of Vehicle Dynamics.

WANG Qing-nian received the B.S. M.S. and Ph.D. degrees from the College of Automotive Engineering, Jilin University, Changchun, China, in 1978 1983 and 1987, respectively.

He is currently a Professor with the College of Automotive Engineering, Jilin University. His research areas is Theory and Control of Hybrid Electric vehicle.

HHT Fuzzy Wavelet Neural Network to Identify Incipient Cavitations in Cooling pump of Engine

LI Li-hong[1, 2], XU Xiang-yang[1], LIU Yan-fang[1], GUO Qian-jin[1], LI Xiao-li[1]

1. School of Transportation Science and Engineering, Beihang University, Beijing, P.R.China

2. College of Vehicle and Motive Power Engineering, Henan University of Science and Technology, Luoyang, P.R.China

lyh_5204@163.com

Abstract—**Incipient cavitations identification is very practical and academic significance for cavity research in cooling pump of engine but it is very complicated. In this paper, a Hilbert-Huang transform(HHT) fuzzy wavelet neural network (FWNN) is proposed for incipient cavitations identification. The main incipient cavitations feature was extracted from entrance pressure fluctuation by the HHT. This FWNN uses wavelet basis function as membership function which shape can be adjusted on line so that the networks have better learning and adaptive ability and at the same time combine the wavelet neural network with fuzzy logical theory to deal with complicated nonlinear, uncertain and fuzzy problem. At last the experiment showed that this identification model can provide fast and reliable incipient cavitations identification with minimum assumptions and minimum requirements for modeling skills.**

Index Terms—**incipient cavitations identification; Hilbert-Huang transform; fuzzy wavelet neural network; wavelet basis function**

I. INTRODUCTION

The water pump is an important part in cooling system of engine, which performances directly impacts on the engine dynamic, economic and durability. The cavitations in water pump are the key reason to result in performance decline, which can be controlled effectively by identifying incipient cavitations. Conventionally, the incipient cavitations were confirmed by measuring pump external performance. At the beginning of cavitations the pump head changes little and the incipient cavitations are unlikely to be diagnosed in time. But when the pump head change distinctly, the cavitations have been very serious. So it is very important to identify incipient cavitations accurately, effectively and in time for engineering practice. Perovic, S, Unsworth, P.J used spectral analysis of their phase current waveform to diagnose Cavitations in Cooling pump of engine [1]. Liu yajun and Hao Dian identified incipient cavitations by analyzing wavelet package theory [2]. But the spectral analysis can't preciously tell the incipient cavitations. The wavelet analysis deal with sample signals as random ones supposing their frequency uniformly distribute in [1]whole

band. So they have localization. In this paper, the Hilbert-Huang transform(HHT) is proposed to extract main signal feature from the entrance pressure fluctuation and then combine with the fuzzy wavelet neural network to optimize global parameters and can preciously confirm the incipient cavitations. HHT can dispose nonlinear signal to a series of intrinsic component features which can more reflect signal intrinsic character and the result has more generality [3]. The cavitations occur possibly due to pool running instance or external fuzzy circumstance such as engine and radiator vibration. Fuzzy technology is an effective tool for dealing with complex nonlinear processes that are characterized with ill-defined and uncertain factors. Fuzzy controllers are often used in achieving a desired performance, the rule base of which is created on the base of the knowledge of human experts. However, for some complicated processes, this knowledge may not be sufficient. The FWNN can use their advantage adequately and combined the wavelet neural network with fuzzy logical theory to dispose the complex, uncertain and fuzzy problems. This model was tested that can more accurately and effectively to identify incipient cavitations in water pump than other model such as WNN and BP.

II. HILBERT-HUANG TRANSFORM

A. Hilbert-Huang Transform Principle

In 1998, Norden E. Huang, who worked in NASA Goddard space flight center, proposed Hilbert-Huang transform(HHT)to process those signals with nonlinear and non-stationary characteristics [4]. The HHT is composed of empirical mode decomposition (EMD) and Hilbert spectral analysis (HSA). The goal of the EMD is to extract a series of IMF from the underlying signal. EMD adopts feature scale parameters based on extremes point [5]. The basic variable described by EMD is instantaneous frequency (IF). The IMF is a function that satisfies two conditions [6]:

1) The number of extrema and zero-crossings of the function must be equal or differ by no more than one. That is that the number of extrema equaled the number of zero points in local group with the symmetry

---

characteristics of the upper envelope line and the lower envelope line.

2) The mean at any point of the envelope defined by the local maxima and local minima is zero.

When we have decomposed the signal into a series of IMF, Hilbert transform can be carried out on each IMF to get a series of instantaneous frequency and amplitude

$$z(t) = s(t) + jH(s(t)) = A(t)\exp(j\phi(t)) \quad (1)$$

Where amplitude function:

$$A(t) = \sqrt{s^2(t) + H^2(s(t))} \quad (2)$$

And angle function:

$$\phi(t) = \arctan\frac{H(s(t))}{s(t)} \quad (3)$$

Instantaneous frequency function:

$$f(t) = \frac{1}{2\pi}\frac{d\phi(t)}{dt} \quad (4)$$

In fact, any signal maybe not satisfies these conditions because of its complexity in the process of pump running. Yet, Norden E. Huang supposed that any signal was composed by some different intrinsic modes, which might be linear or nonlinear [8].At any time, a signal might include many intrinsic modes. If these modes overlapped with each other, they might format compound signals. Hence, any signal may be decomposed into infinite IMFs, which could be achieved by the method bellowed.

(1) For any signal $s(t)$, firstly determine all extreme points of $s(t)$ in local group, and then connect these maximal points to format upper envelope line by cubic spline lines. Meanwhile, connect those minimal points to format lower envelope line by cubic spline lines. These two lines envelope all data of the signal.

(2) The mean value of the two envelope lines is denoted by $m_1(t)$, the difference $h_1(t)$ of $s(t)$ and $m_1(t)$ can be written as:

$$s(t) - m(t) = h_1(t) \quad (5)$$

(3) Judge whether $h_1(t)$ is IMF. If $h_1(t)$ satisfies the IMF's condition, let $c_1(t) = h_1(t)$. If $h_1(t)$ doesn't satisfy the IMF's condition, $h_1(t)$ as the original data, the mean value is denoted by $m_{11}(t)$, then calculate

$$\begin{cases} h_{11}(t) = s(t) - m_{11}(t) \\ ... \\ h_{1k}(t) = s(t) - m_{1k}(t) \end{cases} \quad (6)$$

Repeat step (1) and (2) until $h_1(t)$ satisfies the IMF's condition. Now, let $c_1(t) = h_{1k}(t)$ as the first IMF which denotes high-frequency component of $s(t)$.

(4) Extract $c_1(t)$ from $s(t)$, then achieve signal $r_1(t)$ without high-frequency component.

$$r_1(t) = s(t) - c_1(t) \quad (7)$$

As the original data, $r_1(t)$ is processed according to the steps (1), (2) and (3) to achieve the second IMF, denoted by $c_2(t)$. Repetitions of procession, n IMF are achieved.

$$\begin{cases} r_1(t) - c_2(t) = r_2(t) \\ ... \\ r_{n-1}(t) - c_n(t) = r_n(t) \end{cases} \quad (8)$$

When $c_n(t)$ or $r_n(t)$ satisfies termination condition that $r_n(t)$ becomes a monotonic function, repetition is terminated. From equations (2) and (3), we may have

$$s(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \quad (9)$$

Here, $r_n(t)$ is the residual function that denotes the trend of signal. Yet, each IMF $c_1(t)$, $c_2(t)$, …, $c_n(t)$ denotes different frequency band component from high to low respectively. EMD could decompose the signal into waveform and trend as different scales, and then a series of IMFs with different feature scale are achieved. Each IMF focuses on local characteristic of the signal. Then, features extraction could be more accurate and effective.
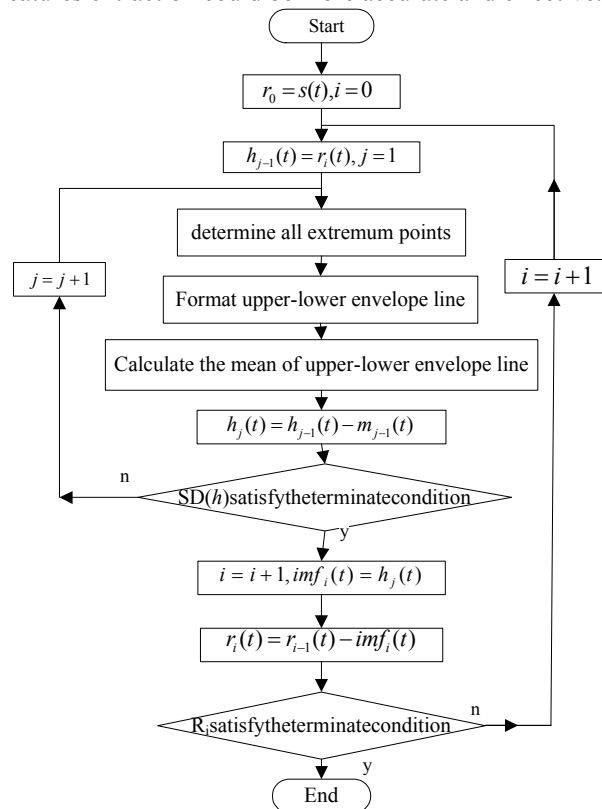

Fig.1 EMD flow chat

B. Extracting Feature Energy from Pressure Signal

When the gas-liquid two-phase flow regime transforms, the differential pressure fluctuation signal can put up different feature. The EMD method can placidly linearize unsteady and nonlinear data and preserve data feature. After decomposed, each component of the pressure

fluctuation signal respectively represents a group of steady signals with feature scale. The energy variety of every frequency band represents different feature, the fore eight energy features scale with main information and the remnant function were elected as network feature vectors to identify cavitations. The steps of energy extraction based on EMD were followed as:

(1)Decompose original sample signal of the differential pressure fluctuation and selecting the fore 8 components with main feature scale.

(2)Resolve total energy of each components

$$E_i = \int_{-\infty}^{+\infty} |c_i(t)|^2 dt \quad (I=1,2,\ldots,8,r) \qquad (10)$$

Where $c_i(t)$ is EMD component

(3)Construct a feature vector with energy as element.

$$T = [E_1, E_2, \ldots, E_7, E_8, E_r] \qquad (11)$$

Because of the bigger energy value, normalize T for subsequent data analysis and disposal. Set:

$$E = (\sum_{i=1}^{8} |E_i|^2 + E_r)^{1/2} \qquad (12)$$

Follow:

$$T' = [E_1/E, E_2/E, \ldots, E_8/E, E_r/E,] \qquad (13)$$

Where the Vector T' is the normalized vector. According to the method above, the normalized data fall in the range [0, 1].

III. IDENTIFICATION WFNN

The neural network and fuzzy theory are two effective identification methods. The fuzzy wavelet neural network (FWNN) integrates the advantages of both fuzzy technique and wavelet neural network and can denote qualitative knowledge. It is capable of self-learning and processing quantificational data. It excels over the normal neural network at learning time, training steps and precision [6].Wavelet transform has the ability to analyze nonstationary signals to discover their local details. Fuzzy logic can reduce the complexity of the data and deal with uncertainty. Their combination can develop a system with fast learning capability that can describe nonlinear systems characterized with uncertainties. Fuzzy neural network can approach any nonlinear system, but its member function is a Gaussian function. Parameter is not easily to adjust. The FWNN with Gaussian function replaced by wavelet function has favorable scale transform and dilation-translation characters. It can learn on smooth function at large scale and on local fantastic function at small scale. The parameters can be easily adjusted.

A. FWNN constructing

FWNN put the basic factors: membership function, fuzzy logical rules, fuzzy mapping and reversal fuzzy mapping and hidden fuzzy rules into relative structure and make it possess learning capability for membership function and fuzzy logical rules. FWNN integrates wavelet and neural network fuzzy model. The kernel of the fuzzy system is the fuzzy knowledge base that

consists of the input–output data points of the system interpreted into linguistic fuzzy rules. The consequent parts of fuzzy IF–THEN rules are represented by either a constant or a function. These fuzzy networks do not provide full mapping capabilities, and, in the case of modeling of complex nonlinear processes, may require a high number of rules in order to achieve the desired accuracy. Increasing the number of the rules leads to an increase in the number of neurons in the hidden layer of the network. The use of wavelet (rather than linear) functions are proposed to improve the computational power of the neuron-fuzzy system. The basic configuration of the FWNN system includes a fuzzy rule base, which consists of a collection of fuzzy IF-THEN rules in the following form:

R^l : IF x_1 is F_1^l and ... and x_m is F_m^l THEN y_1 is G_1^l and ... and y_n is G_n^l

where Rl are the lth rule ($1 \le l \le s$), $\{xi\}i=1\ldots m$ the input variables, $\{yj\}j=1\ldots n$ the output variables, F_i^l are the labels of the fuzzy sets characterized by the membership functions(MF) $\mu_{F_i^l}(x_i)$, G_1^l are the labels of the fuzzy subsets in the output space.

The structure of FWNN proposed in this paper for identification. It includes four layers structure with an input layer, wavelet layer (membership layer), fuzzy inference and non-fuzzy layer. A schematic diagram of the four-layered FWNN is shown in Fig.2.
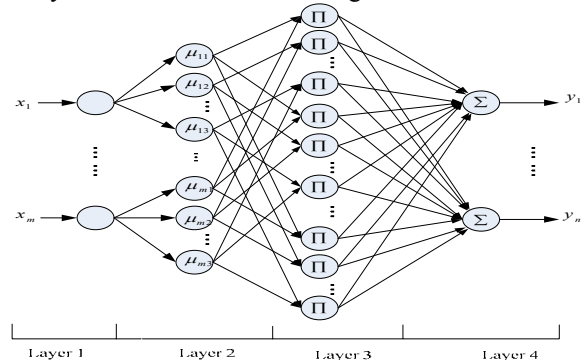

Fig.2. FWNN topology structure for cavitations identification

The first Layer (input layer): In this layer, the number of nodes is equal to the number of input feature energy. These nodes are used for signals feature. For every node i, the net input and the net output are related by

$$I_i^{(1)} = x_i$$
$$O_i^{(1)} = I_i^{(1)} \qquad i = 1,2,\ldots, m, \qquad (14)$$

where $I_i^{(1)}$ and $O_i^{(1)}$ denote ith input node and ith output node, respectively.

The second layer (membership layer): In this layer, the input vector is described as fuzzy member and the output is the corresponding membership degree. The wavelet basis function is adopted as the membership function. For every input vector, three linguistic fuzzy variable are adopted as no, just and over. The member relation of Input vector x_i and the corresponding linguistic fuzzy variable is written as:

$$\mu(x_i)=\psi_{a_{ij},b_{ij}}(x_i)=\psi\left(\frac{x_i-b_{ij}}{a_{ij}}\right)=\cos(0.25\bullet\frac{x_i-b_{ij}}{a_{ij}})\bullet\exp\left[-\frac{(x_i-b_{ij})^2}{2a_{ij}^2}\right]$$

$$(15)$$

where $i=1,2,\dots,m$; $j=1,2,\dots,r$, $r(r=5)$ represents the number of linguistic values for each input, a_{ij} and b_{ij} are dilation parameter and translation parameter correspondingly.

The relation between the input and output is represented as

$$I_{ij}^{(2)}=O_i^{(1)}$$

$$O_{ij}^{(2)}=\mu_{ij}(I_{ij}^{(2)})=\cos(0.25\bullet\frac{O_i^{(1)}-b_{ij}}{a_{ij}})\bullet\exp\left[-\frac{(O_i^{(1)}-b_{ij})^2}{2a_{ij}^2}\right]\quad(16)$$

where $i=1,2,\dots,m; j=1,2,\dots,r$.

The third layer (rule layer): this fuzzy inference layer completes fuzzy operation with fuzzy logical rules. The number of nodes corresponds to the number of rules .Each node represents one fuzzy rule. Here, to calculate the values of the output signals of the layer AND (min) operation is used. The normal AND (min) operation includes intersection calculation and algebraic calculation.

If intersection is used [11-14], we have

$$O_i^{(3)}=\min(O_{1i_1}^{(2)},O_{2i_2}^{(2)},\dots,O_{mi_m}^{(2)})\qquad(17)$$

On the other hand, if algebraic operation is used, we have

$$O_i^{(3)}=O_{1i_1}^{(2)}O_{2i_2}^{(2)}\dots O_{mi_m}^{(2)}\qquad(18)$$

Each node in this layer is denoted by $\prod$, the relation between the input and output is represented as

$$I_i^{(3)}=O_{1i_1}^{(2)}*O_{2i_2}^{(2)}*\dots*O_{mi_m}^{(2)}=\min\{O_{1i_1}^{(2)},O_{2i_2}^{(2)},\dots,O_{mi_m}^{(2)}\}\quad(19)$$

The net output is written as

$$O_i^{(3)}=I_i^{(3)}\qquad(20)$$

where $i=1,2,\dots,s$, $s=\prod_{i=1}^{m}r^m=\prod_{i=1}^{m}5^m$

The fourth layer (non-fuzzy layer): this layer is to cancel the fuzzy factors of output. Nodes in this layer represent the output variables of the system.

$$I_i^{(4)}=\sum_{j=1}^{S}W_{ij}O_j^{(3)}\qquad(21)$$

The net output is written as

$$Y_i=O_i^{(4)}=I_i^{(4)}$$
$$i=1,2,\dots,n\quad j=1,2,\dots,s\qquad(22)$$

where w_{ij} are the connection weights, Y_i represents the ith output to the node of layer 4.

B. Training Algorithm for FWNN

In this study, a gradient-descent (GD) learning algorithm with adaptive learning rate is adopted. The latter guarantees the convergence and speeds up the learning of the network. In addition, a momentum is used to speed-up the learning process. The parameter to be trained are the parameters of wavelets (a_{ij}, b_{ij}, w_{ij}, $i=1,\dots,n,j=1,\dots,m$) in the consequent part. The initial values are generated randomly.

Here, the gradient descent learning algorithm with adaptive learning rate is introduced. The learning error function can be written as

$$E=\frac{1}{2}(D-Y)^T(D-Y)\qquad(23)$$

where D is the desired output acquired from specialists, and Y is the FWNN's current output, $Y=[y_1,y_{2,\dots,}y_n]$.

The weights in the output layer are trained by:

$$\Delta W_{ij}=-\frac{\partial E}{\partial W_{ij}}=-\frac{\partial E}{\partial O_i^{(4)}}\bullet\frac{\partial O_i^{(4)}}{\partial I_i^{(4)}}\bullet\frac{\partial I_i^{(4)}}{\partial W_{ij}}=(D_i-Y_i)\bullet O_j^{(3)}\quad(24)$$

The weights of the output layer are updated according to the following equation:

$$W_{ij}(t+1)=W_{ij}(t)+\eta\bullet\Delta W_{ij}=W_{ij}+\eta\bullet(D_i-Y_i)\bullet O_j^{(3)}\quad(25)$$

where $i=1,2,\dots,n$; $j=1,2,\dots,s$; η is the learning-rate parameter.

One of the important problems in learning algorithms is the convergence. The convergence of the gradient descent method depends on the selection of the initial values of the learning rate and the momentum term. Usually, these values are selected in the interval [0–1]. A large value of the learning rate may lead to non-stable learning; a small value of learning rate leads to slow learning speed. In this paper, an adaptive approach is used for updating these parameters.

For this fuzzy wavelet neural network, which is very important is to select wavelet base which is decided by actual situation. At present, many wavelet function can be supplied such as Morlet，Harr，Mexican Hat and Meyer. Morlet wavelet has been chosen to serve as an adoption basis function [10] which has been proved to be concise and applicable.

$$\psi_{a,b}(t)=\cos(0.25t_z)e^{-\frac{t_z^2}{2}}\qquad(26)$$

where $t_z=(t-b)/a$

$$\psi_{a,b}'(t)=-\left(0.25\sin(0.25t_z)\exp(-\frac{t_z^2}{2})+\cos(0.25t_z)\exp(-\frac{t_z^2}{2})t_z\right)$$

$$(27)$$

Training the dilation factor and translation factor, and the modification of a_{ij} and b_{ij} are expressed as:

$$\Delta a_{ij}=-\frac{\partial E}{\partial a_{ij}}=-\frac{\partial E}{\partial O_i^{(4)}}\bullet\frac{\partial O_i^{(4)}}{\partial I_i^{(4)}}\bullet\frac{\partial O_k^{(3)}}{\partial O_k^{(3)}}\bullet\frac{\partial I_k^{(3)}}{\partial I_k^{(3)}}\bullet\frac{\partial O_{ij}^{(2)}}{\partial O_{ij}^{(2)}}\bullet\frac{\partial I_{ij}^{(2)}}{\partial I_{ij}^{(2)}}\bullet\frac{\partial I_{ij}^{(2)}}{\partial a_{ij}}$$

$$=2\sum_l\left[(D_l-Y_l)\bullet\sum_{k_1,\dots,k_{i-1},k_{i+1},\dots,k_m}(O_{1k_1}*\dots*O_{i-1k_{i-1}}*O_{i+1k_{i+1}}*\dots*O_{mk_m})\bullet W_{lk}\right]\bullet\left[-\frac{(O_i^{(1)}-b_{ij})^2}{2a_{ij}^2}\right]$$

$$\bullet\exp\bullet\frac{O_i^{(1)}-b_{ij}}{a_{ij}^2}\bullet\left[0.25\sin(0.25\bullet\frac{O_i^{(1)}-b_{ij}}{a_{ij}})+\cos(0.25\bullet\frac{O_i^{(1)}-b_{ij}}{a_{ij}})\bullet\frac{O_i^{(1)}-b_{ij}}{a_{ij}}\right]$$

$$(28)$$

$$\Delta b_{ij}=-\frac{\partial E}{\partial b_{ij}}=-\frac{\partial E}{\partial O_i^{(4)}}\bullet\frac{\partial O_i^{(4)}}{\partial I_i^{(4)}}\bullet\frac{\partial O_k^{(3)}}{\partial O_k^{(3)}}\bullet\frac{\partial I_k^{(3)}}{\partial I_k^{(3)}}\bullet\frac{\partial O_{ij}^{(2)}}{\partial O_{ij}^{(2)}}\bullet\frac{\partial I_{ij}^{(2)}}{\partial I_{ij}^{(2)}}\bullet\frac{\partial I_{ij}^{(2)}}{\partial b_{ij}}$$

$$=2\sum_l\left[(D_l-Y_l)\bullet\sum_{k_1,\dots,k_{i-1},k_{i+1},\dots,k_m}(O_{1k_1}*\dots*O_{i-1k_{i-1}}*O_{i+1k_{i+1}}*\dots*O_{mk_m})\bullet W_{lk}\right]$$

$$\bullet\exp\left[-\frac{(O_i^{(1)}-b_{ij})^2}{2a_{ij}^2}\right]\bullet\frac{1}{a_{ij}}\bullet\left[0.25\sin(0.25\bullet\frac{O_i^{(1)}-b_{ij}}{a_{ij}})+\right.$$

$$\left.\cos(0.25\bullet\frac{O_i^{(1)}-b_{ij}}{a_{ij}})\bullet\frac{O_i^{(1)}-b_{ij}}{a_{ij}}\right]$$

$$(29)$$

Then the wavelet node parameters are updated as follows:

$$a_{ij}(t+1) = a_{ij}(t) + \eta \Delta a_{ij} \qquad (30)$$

$$b_{ij}(t+1) = b_{ij}(t) + \eta \Delta b_{ij} \qquad (31)$$

where $i=1,2,\ldots,m$; $j=1,2,\ldots,r$; $k=1,2,\ldots,s$; $l=1,2,\ldots,n$; η is the learning rate. We let the learning rate η vary to improve the speed of convergence, as well as the learning performance.

IV. HT FWNN IDENTIFICATION SYSTEM EXPERIMENT ANALYSIS

Test system is composed of data sampling and data processing. Data sampling is hard part of the system which consists of signal measure and data sampling as Fig.3. signal measure is made up of pressure signal picker as PD-23 difference transformer from Switzerland KELER company, its capacity 0~10kPa, output 1V~5V,precision 0.5%, frequency more than 1000Hz. It has adequate precision and desirable dynamic response characters. The data sampling is composed of computer, IMP3595A data sampling card, IMP35951C data sampling transformer and data line, power supply. Sampling frequency is 1000—10000Hz. It was set as 1000Hz. IMP data sampler has an intelligent input-output top. Every IMP of system does not only contain microprocessor and logical array door, but also EPROM and RAM. So, it is capable of processing data and programming. Meanwhile, every IMP can make some data pretreatment by inner intrinsic software and delivery part data processed by main computer to data sampler in addition to complete data sampling and communication.
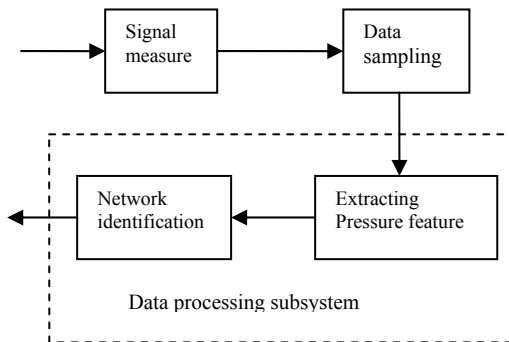

Fig.3 Cavitations identification system structure chat

A. Experiment structure

The water pump cavitations experiment structure was adopted as Fig. 4. The normal speed of pump is 3000rpm, charge Q=18m^3 / h, the number of blade z=7; The work medium is water. The entrance section introduced transparent organic glasses convenient for watching the cavitations of pump entrance. During the test processing, firstly regulated pump charge to prescriptive amount and keep it. When the pump run at normal and the pressure difference of input-output was stable, the entrance pressure at normal running (the vacuum pressure was zero) was considered as reference point and then adjust vacuum pump to change entrance pressure. Meanwhile noting the cavitations condition and external character

curve, the pressure fluctuation value at entrance was picked in real-time at the pump head descended to none, 3% and 6% respectively.
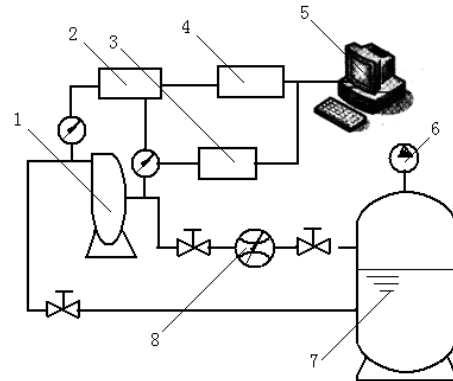
1 test pump, 2 pressure transform meter, 3 data picker1, 4 data picker2, 5 computer, 6 vacuum pump 7 water tank, 8 vortex flow meter

Fig.4 The water pump cavitations experiment

B. Data disposing and feature extracting

Fig.5, Fig.6 and Fig.7 are pressure signal composing figure at the pump head descended to none, 3% and 6% respectively. From figures, EMD decomposed pressure signal to eight IMF components and a residual function. The different IMF component included different time scale which displayed signal features at different differentiating rate. The part of every layer energy values were calculated by (10) shown as table 1, the corresponding normalized values were written in table 2.
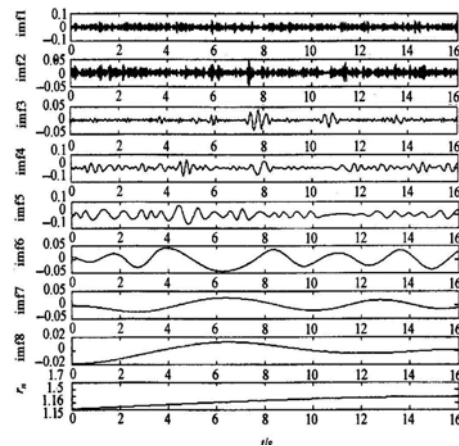

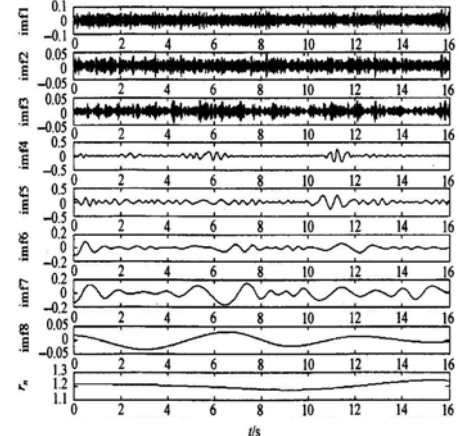Fig.5 IMFs of pressure signal at no head decline

Fig.6 IMFs of pressure signal at 3% head decline

Table 1 the fore feature energy and remnant function energy value

E1	E2	E3	E4	E5	E6	E7	E8	r
1.2949	0.8108	0.3797	0.2877	0.1589	0.1234	0.0995	0.0465	2.9082
1.8368	1.8802	1.4236	1.333	0.713	0.6965	0.1567	0.0105	1.2902
1.6558	1.2289	0.9216	0.8311	0.6504	0.3765	0.4560	0.0201	2.0451
2.6989	1.8221	1.5897	1.3110	0.8394	0.5563	0.2042	0.0854	1.8946
1.1755	1.2389	1.0422	0.9609	0.6313	0.5285	0.4366	0.0665	1.9625
1.9100	1.1342	1.8600	0.6875	0.6101	0.4490	0.1032	0.0672	1.8399
1.9302	1.2564	1.0760	1.0010	0.8749	0.4548	0.4043	0.0977	2.1084
1.7989	1.1371	0.9199	0.5103	0.2406	0.1287	0.0977	0.0318	1.6692
2.187	1.1916	1.1437	0.8157	0.7220	0.2454	0.0920	0.0612	2.2227
1.6910	0.9272	0.8442	0.5307	0.2448	0.0937	0.0370	0.0027	2.1736
1.4392	1.3643	1.2802	0.9240	0.3849	0.0803	0.0891	0.0783	1.5113
1.5449	1.2164	1.2110	0.9372	0.6563	0.6265	0.5274	0.0973	1.3434
2.5723	1.9423	1.5321	0.9909	0.8269	0.7881	0.5260	0.0680	2.7540
0.9203	0.542	0.3700	0.4554	0.3454	0.3031	0.2136	0.0467	2.2241
1.6911	1.2982	1.2609	1.2145	1.1614	0.9756	0.3894	0.0323	3.2611

Table 2 the normalized values in table 1

| e1	e2	e3	e4	e5	e6	e7	e8	er
0.3891	0.2436	0.1141	0.0864	0.0477	0.0370	0.0299	0.0139	0.8740
0.5018	0.5137	0.3889	0.3642	0.1948	0.1903	0.0428	0.0028	0.3525
0.5050	0.3748	0.2811	0.2535	0.1983	0.1148	0.1390	0.0061	0.6238
0.6111	0.4125	0.3599	0.2968	0.1900	0.1259	0.0462	0.0193	0.4290
0.3783	0.3988	0.3354	0.3093	0.2032	0.1701	0.1405	0.0214	0.6317
0.5330	0.3165	0.5190	0.1918	0.1702	0.1252	0.0287	0.0187	0.5134
0.5342	0.3477	0.2978	0.2770	0.2421	0.1258	0.1118	0.0270	0.5835
0.6167	0.3898	0.3153	0.1749	0.0824	0.0441	0.0334	0.0109	0.5722
0.5906	0.3218	0.3088	0.2202	0.1949	0.0662	0.0248	0.0165	0.6002
0.5483	0.3006	0.2737	0.1721	0.0793	0.0303	0.0119	0.0008	0.7049
0.4830	0.4578	0.4296	0.3101	0.1291	0.0269	0.0299	0.0262	0.5072
0.5113	0.4026	0.4008	0.3102	0.2172	0.2073	0.1745	0.0322	0.4446
0.5376	0.4059	0.3202	0.2071	0.1728	0.1647	0.1099	0.0142	0.5756
0.3557	0.2095	0.1430	0.1760	0.1335	0.1171	0.0825	0.0180	0.8598
0.3717	0.2853	0.2771	0.2669	0.2552	0.2144	0.0855	0.0070	0.7168
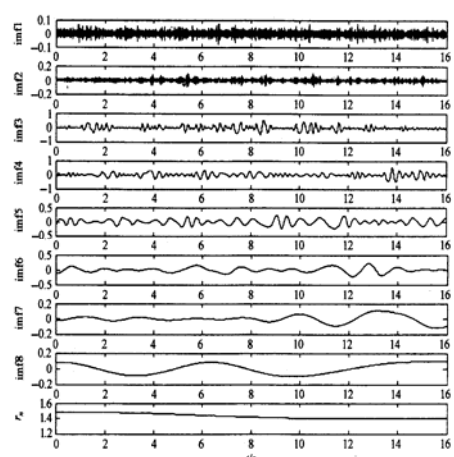
Fig.7 IMFs of pressure signal at 3% head decline

C. Membership function

According to table 2, the sets of sample in Table1 can be normalized between 0 and 1. These normalized vectors are used as input and output of fuzzy sets. Described as Section Ⅲ, the relative identifying knowledge is necessary to be expressed to fuzzy mode rules when constructing FWNN identification of cavitations system and then by which the new learning and training samples can be constructed. So, three fuzzy sets are used for fuzzy identification rules, corresponding to no cavitations, just cavitations, over cavitations (labeled as n, j, and o respectively). Morlet wavelet function is chosen as the membership functions for these three fuzzy sets. Here, the three fuzzy sets for the input and output linguistic variables of the FWNN identification system have both been designed in the same range of [0,1] (Fig. 8). $b_{i1} = 0$, $b_{i2} = 0.5$, $b_{i3} = 1$. Once the shape of the fuzzy sets is given, the relationship between input and output variables of the FWNN based identification system is defined by a set of linguistic statements that are called fuzzy rules.
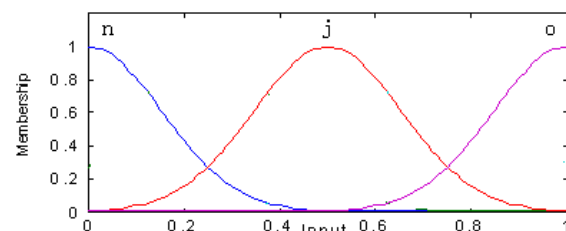

Fig.8 The wavelet based membership functions

D. Network training and determination

Network initialization initializing the wavelet network parameters is an important issue. A random initialization of all the parameters to small random values (as usually done with neural networks) is not desirable since this may

make some wavelets too local (small dilations) and make the components of the gradient of the cost function very small in areas of interest.

In [14], a method was proposed for dilation and translation initialization. This work uses this method to initialize the dilation aj and translation bj. In general, one wants to take advantage of the input space domains where the wavelets are not zero. According to the wavelet theory, given the t* as the center of the time domain, $\Delta\psi$ as the radius of the mother wavelet, the time domain so can be written as:

$$\left[b + at^* - a\Delta\psi, b + at^* + a\Delta\psi \right] \tag{32}$$

In order to guarantee that the wavelets extend initially over the whole input domain, the initialization of the dilation and translation parameters must agree to the following equation:

$$\begin{cases} b_j + a_j t^* - a_j \Delta\psi = \sum_{i=1}^{I} w_{ji} x_{i\min} \\ b_j + a_j t^* + a_j \Delta\psi = \sum_{i=1}^{I} w_{ji} x_{i\max} \end{cases} \tag{33}$$

Thus

$$\begin{cases} a_j = \dfrac{\sum_{i=1}^{I} w_{ji} x_{i\max} - \sum_{i=1}^{I} w_{ji} x_{i\min}}{2\Delta\psi} \\ b_j = \dfrac{\sum_{i=1}^{I} w_{ji} x_{i\max}(\Delta\psi - t^*) + \sum_{i=1}^{I} w_{ji} x_{i\min}(\Delta\psi + t^*)}{2\Delta\psi} \end{cases} \tag{34}$$

The choice of the learning rate factor η has an important effect on stability and convergence rate. If it is too small, too many steps are needed to reach an acceptable solution. On the contrary, a large learning rate may possibly lead to skip the optimal solution and cause oscillation. In the experiment, η are initialized to 0.9.

The choice of the weights is less critical: these parameters are initialized to small random values.

Network training and determination In the experiment, 200 groups of sample data are acquired on Fig 3. 160 groups of the data are used as the training sets, the remaining 40 groups are used as the validate sets. Expecting output error threshold is 0.001. Sample vectors are used as input and output of FWNN. After all possible normal operating modes of the identification are learned, the system enters the identification stage.

FWNN start as a network of nodes arranged in four layers--the input, membership layer, rule layer, fuzzy-cancelling layer. The input and fuzzy-cancelling and rule layers serve as nodes to buffer input and fuzzy-cancelling for the model, respectively, and the membership layer serves to provide a means for input relations to be represented in the output. Here, the network construction used for cavitations identification consists of 9 inputs corresponding to the 8 feature energy value and residual energy value that have been given (listed in table2), and one outputs corresponding to one unknown condition.

To demonstrate the performance of the FWNN-based approach on cavitations classification, comparisons are made with two types of artificial neural networks, namely wavelet neural networks and BP networks. Fig.9, demonstrate the training history and the performance of the FWNN, WNN and BP networks by using GD algorithm respectively. Table3 are the comparison results of FWNN, WNN and BP methods on identification. The second column in this table lists the identification accuracy on the 40 actual sample data. The third column in this table lists the average number of error function identification used during training, until running terminated with the network converged or the epochs exceeding the maximum epochs. Compared with the BP method, the FWNN method has more than 4.0% improvement on the identification accuracy, and compared with the WNN method, the FWNN method has 2% improvement on the identification accuracy. The test results confirm that, in all compared cases, the proposed FWNN method has a better capability for generalization than the other methods.
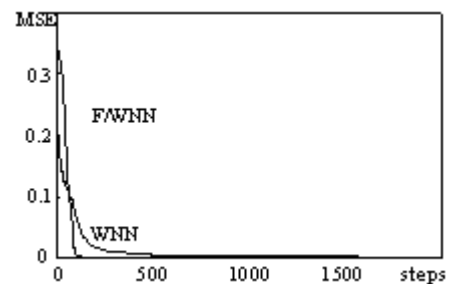
(a) The compare between FWNN and WNN
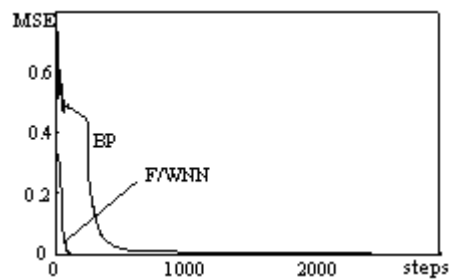
(b) The compare between FWNN and BP
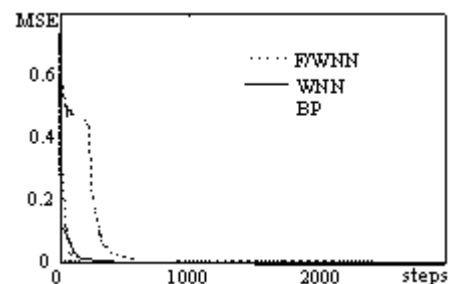
(c) The compare among FWNN ,WNN and BP

Fig.9   Convergence curves for ANNs

Table3 Comparisons of FWNN, WNN and BP method

Method	Estimation accuracy (%)	Sum error	The number of Training steps
BP	92.0	0.001	2490
WNN	94.0	0.001	1550
FWNN	96.0	0.001	135

From the comparison of various methods, it can be seen that, the FWNN method outperformed all the other construction. The FWNN construction is decidedly superior, the errors are smaller than the WNN and BP construction. Another benefit of the FWNN approach is the reduced training time and steps of training.

V. CONCLUSION

Experiment result showed that onefold neural network was adopted to identify the cavitations in the water pump, the precision is very low and can't satisfy controlling the water pump in time and accurately. But the method introduced above that the sample signals are preprocessed by HHT firstly and then identified by FWNN can not only attain higher precision but also speed up the identification.

Based on fuzzy neural network, a FWNN model was build and used the wavelet function as subjection function according to favorable scale transform and dilate translation characters. This model can be used as incipient cavitations identifier. After simulation, it is shown that at the same condition the FWNN model structure is simple, learning algorithms effective and identifying capacity high comparing with other approaches such as WNN and BP.

REFERENCE

[1] Perovic, S., Unsworth, P.J., Higham, E.H. "Fuzzy logic system to detect pump faults from motor current spectra", Thirty-Sixth IAS Annual Meeting. Conference Record of the 2001 IEEE. Industry Applications Conference 2001 ,pp: 274 - 280 vol.1,doi: 10.1109/IAS.2001.955423

[2] Sirok, B.,Hocevar, M., Kern, I., Novak, M."Monitoring of the cavitation in the Kaplan turbine", ISIE '99. Proceedings of the IEEE International Symposium on. 1999, pp. 1224 - 1228 vol.3, doi: 10.1109/ISIE.1999.796873

[3] HAO Dian, ZHOU Chang-jing, ZHANG Zhi-fe．Diagnosis of inception cavitation in centrifugal pump using wavelet, combined with autocorrelation method [J].Journal of the University of Petroleum,China 2005 vol 29(2) pp.78—82

[4] Guo Qian-jin, YU Hai-bin, XU Ai-dong. "Wavelet fuzzy neural network for fault diagnosis," Communications, Circuits and Systems, Proceedings. 2005 International Conference on vol 2, doi: 10.1109/ICCCAS.2005.1495274

[5] N. Subhasis, H. A. Toliyat, "Condition monitoring and fault diagnosis of electrical machines - A review" IAS,pp. 197 - 204, Vol. 1 1999. doi: 10.1109/IAS.1999.799956

[6] Abiyev, R.H., Kaynak, O. "Fuzzy Wavelet Neural Networks for Identification and Control of Dynamic Plants—A Novel Structure and a Comparative Study", Industrial Electronics, IEEE Transactions on 2008, vol 55 9(8) pp:3133-3140,doi: 10.1109/TIE.2008.924018

[7] F. J. Lin, C. H. Lin, and P. H. Shen, "Self-constructing fuzzy neural network speed controller for permanent-magnet synchronous motor drive," IEEE Trans Fuzzy Syst., vol. 9, pp. 751–759, Oct. 2001. doi:10.1109/91.963761

[8] C. H. Lee and C. C. Teng, "Identification and control of dynamic systems using recurrent fuzzy neural networks," IEEE Trans. Fuzzy Syst., vol. 8, pp. 349–366, Aug. 2000,doi: 10.1109/CERMA.2007.4367670

[9] S. F. Su and F. Y. Yang, "On the dynamical modeling with neural fuzzy networks," IEEE Trans. Neural Networks, vol. 13, pp. 1548–1553, Nov.2002.doi: 10.1109/TNN.2002.804313

[10] Monroy, P.E.M., Perez, H.B. "A recurrent fuzzy-neural model for dynamic system identification," IEEE Trans. Syst., Man, Cybern, vol.32, pp.112–117, Apr.2002. doi: 10.1109/CERMA.2007.4367670

[11] E. Sblomot, V. Cuperman, and A. Gersho, "Hybrid coding: combined harmonic and waveform coding of speech at 4 kh/s" Speech and Audio Processing. IEEE Transactions on, Volume: 9 Issue: 6, pp. 632, Sept. 2001, doi: 10.1109/89.943341

[12] W. T. Thomson, "A review of ON-LINE condition monitoring techniques for three phase squirrel cage induction motors - past, present and future", The 1999 IEEE International Symposium on Diagnosis for Electrical Machines, Power and Drives IEEE SDEMPED'99, Spain, pp. 3-18

[13] F.C.A. Brooks, L. Hanzo, "A multiband excited waveforminterpolated 2.35-khps speech codec for andlimited channels" Vehicular Technology. IEEE Transactions on, Volume: 49 Issue: 3, pp. 766 -777, May 2000

[14] O. Gottesman, A. Gersho. "Enhanced waveform interpolative coding at low hit-rate "Speech and Audio Processing, IEEE Transactions on. Volume: 9 Issue: 8, pp. 786 -798, Nov. 2001,doi: 10.1109/89.966082

[15] Sokka, S.; Gauthier, T.P. "Hynynen, K.Spatial control of cavitation: theoretical and experimental validation of a dual-frequency excitation method",Ultrasonics Symposium, 2004pp. 878 - 881 Vol.2, doi: 10.1109/ULTSYM.2004.1417875

LI Li-hong, female, 1968.10.23, born in Luoyang city , Henan province.

Educational background: 1990.6, bachelor's degree, graduated from School of Mechatronics Engineering, Henan University of Science and Technology, Luoyang, China. 2004.6, master's degree, graduated from Vehicle & Motive Power Engineering College, Henan University of Science and Technology, Luoyang, China. From 2007.9. untill now Candidate for doctor's degree, study in School of Transportation Science and Engineering, Beihang University, Beijing, China.

Work experience: 1990.7-2001.8, technician, engineer, section chief in sequence, worked in technology department, No.1 tractor Corp Ltd., China. 2004.7-2007, teacher, associate professor in sequence, in Vehicle & Motive Power Engineering College, Henan University of Science and Technology, China.

XU Xiangyang, Male, 1965.5.9 born in Laiyang, Shandong province, China. Professor of School of Transporta- tion Science and Engineering, Beihang University, Beijing, China.

The Smith-PID Control of Three-Tank-System Based on Fuzzy Theory

Jianqiu Deng

Tsinghua University/ School of Automation, Beijing, China
Naval Aeronautical Engineering Institue,Shandong, China
Email: djq06@mails.tsinghua.edu.cn

Cui Hao

Beijing Information Science & Technology University / School of Automation, Beijing, China
E-mail:haocui0433234@163.com

Abstract—**According to the character of the volume-lag of the controlled process of three-tank-system, Smith predictor was adopted to compensate three-tank-system fuzzy adaptive control system. Fuzzy adaptive Smith predictive control system is composed with Smith predictor and fuzzy adaptive controller. The PID parameters were setting on line. This algorithm uses fuzzy adaptive PID control to improve the resistance ability to random disturbance and Smith predictive control to overcome the time-delay character of controlled object. Simulation results showed that this control algorithm has the advantages of strong adaptive ability and noise immunity.**

Index Terms -**fuzzy control, adaptive, Simth-PID, three-tank-system**

I. INTRODUCTION

As shown in [1], the controlled process of three-tank-system has obviously character of volume-lag. Due to the complexity of the controlled object, traditional PID control can't satisfy the control requirements of the system. Smith—PI controller was used in [1]. Because the parameter K_p and K_i were man-set, the control effect was not satisfactory. Fuzzy adaptive controller was used in [2]. The random disturbance was inhibited effectively, but the control effect should be improved because of the delays of the system. The schemes of the control of the three-tank-system are mainly the two above by now.

In recent years, the research of advanced control was increase day by day, and fuzzy adaptive control is one of them. Fuzzy adaptive control which is often composed into synthetic controllers with other control schemes has great potential for the control of nonlinear and uncertain systems. A fuzzy adaptive controller was designed in [3] based on the dynamic model and the kinematic controller of the mobile robot of non-holonomic constraint. The controller can overcome the influence of the disturbance and the unknown of the parameters of the robot model[3]. Fuzzy adaptive controller was applied to the reconfigurable manipulator in [4]. The fuzzy adaptive

system can approximate the dynamic model of the subsystem well. Fuzzy adaptive sliding mode control scheme and it's strengthen scheme were used in uncertain systems and nonlinear uncertain systems in [5] and [6]. Fuzzy adaptive control was applied to the control system of DC motor in [7], which can overcome disturbance effectively. The stability of the closed-loop system was proved with the use of Lyapunov analysis. A fuzzy adaptive controller was proposed in the control of wheeled inverted pendulums in [8]. Based on Lyapunov synthesis, the fuzzy control ensures that the system outputs track the given bounded reference signals to within a small neighborhood of zero, and overcome disturbance effectively. In addition, fuzzy adaptive control was used in nonlinear SISO systems[9,10,11] and MIMO systems[12,13], and obtained ideal control effects.

The traditional Smith controller can control the delay system effectively. The fuzzy control is changing the human fuzzy language to the machine language that computers can distinguish and controlling the complex object through processing the fuzzy information according to the human experience. So, the fuzzy controller has good control effect to the system with random disturbance. The fuzzy adaptive control and Smith predictive control are combined together in this paper, which improves the robustness of Smith control greatly.

II. PROBLEM DESCRIPTION

A. Controlled Object Description

Shanghai new Aotuo three-tank-system is a nonlinear coupling controlled model with multi-input and multi-output. The subject of the experimental device is composed by the three same columniform tanks which are made by transparent organic glass, a reservoir, and corresponding actuator and sensors. Such as Figure1, T1, T2 and T3 are three water containers. V1, V2 and V3 are junction valves between any two of them. V7 and V8 are two adjustable proportional valves. The water is injected into T1 and T2 through pumps P1 and P2. V4, V5 and V6 are three manual regulating valves, which can discharge

the water of the three containers into tank. Through setting the open or close state of the six manual valves, different object models can be combined.

In this system, the manual valves V4, V5 and V6 are closed, pump P2 are closed, and the manual valves V1, V2 and V3 are adjusted to the appropriate opening (demanding that: V1>V2>V3). The liquid in the reservoir is extracted by pump P1, injected into the container T1 through the proportional valve V7, and then the liquid injected into the container T3 through the manual valve V1, reached container T2 through manual valve V2, and injected into the reservoir through the manual valve V3. In the object model, the input variable is the flow Q_1 which is injected into container T1, the controlled variable is the liquid level h_2 of the container T2. That composes the three-tank-control-system. The system is single-input single-output third-order object model with volume time-delay.

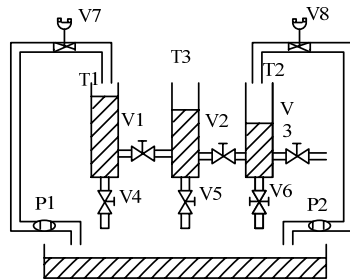The control schematic diagram of three-tank-system is shown in Figure 1.

Figure 1. The schematic diagram of the control device of three-tank-system

B. Mathematical Modeling

Variables and parameters are defined as below: h_i-liquid level, Q_{ij}-flow, Q_1-inflow, A-cross sectional area of the container, R_i-liquid resistance of the manual valve V_i, i=1,2,3; (i,j)$\in$ {(1,3);(3,2);(2,0)}. Q_{13} represents the flow from T1 to T3, Q_{32} represents the flow from T3 to T2, and Q_{20} represents the flow from T2.

The total delay time of the three-tank-system is τ, and the volume time-delay of each container is $\tau/3$. Suppose that the flow injected into T1 is $Q_1(t)$. Because the volume time-delay of the container T1 is $\tau/3$, its liquid level can be expressed as $h_1(t-\tau/3)$. The flow from T1 to T3 is expressed as $Q_{13}(t-\tau/3)$. Because the volume time-delay of the container T3 is $\tau/3$, its liquid level of T3 can be expressed as $h_3(t-\tau/3)$. The flow from T3 to T2 is expressed as $Q_{23}(t-2\tau/3)$. The output liquid level of T2 is $h_2(t-\tau)$. The flow of T2 is $Q_{20}(t-\tau)$. In this system, the inflow Q_1 is input of the controlled process, The liquid level h_2 of T2 is the output, and the mathematical model is the mathematical expression between h_2 and Q_1. Three balance equations of T1, T2 and T3 are established:

$$A\frac{dh_1(t-\frac{\tau}{3})}{dt} = Q_1(t) - Q_{13}(t-\frac{\tau}{3}) \tag{1}$$

$$A\frac{dh_3(t-\frac{2\tau}{3})}{dt} = Q_{13}(t-\frac{\tau}{3}) - Q_{32}(t-\frac{2\tau}{3}) \tag{2}$$

$$A\frac{dh_2(t-\tau)}{dt} = Q_{32}(t-\frac{2\tau}{3}) - Q_{20}(t-\tau) \tag{3}$$

Expressing (1), (2) and (3) in incremental forms:

$$A\frac{dh_1(t-\frac{\tau}{3})}{dt} = \Delta Q_1(t) - \Delta Q_{13}(t-\frac{\tau}{3}) \tag{4}$$

$$A\frac{dh_3(t-\frac{2\tau}{3})}{dt} = \Delta Q_{13}(t-\frac{\tau}{3}) - \Delta Q_{32}(t-\frac{2\tau}{3}) \tag{5}$$

$$A\frac{dh_2(t-\tau)}{dt} = \Delta Q_{32}(t-\frac{2\tau}{3}) - \Delta Q_{20}(t-\tau) \tag{6}$$

According to fluid dynamics, there is linear relationship between the liquid level and flow in the turbulence situation. Through linear processing:

$$\Delta Q_{13}(t-\frac{\tau}{3}) = \frac{\Delta h_1(t-\frac{\tau}{3})}{R_1} \tag{7}$$

$$\Delta Q_{32}(t-\frac{2\tau}{3}) = \frac{\Delta h_3(t-\frac{2\tau}{3})}{R_2} \tag{8}$$

$$\Delta Q_{20}(t-\tau) = \frac{\Delta h_2(t-\tau)}{R_3} \tag{9}$$

After Laplace transformation, the mathematical model of the three-tank-system is:

$$W(s) = \frac{H_2(s)}{Q_1(s)} \tag{10}$$

$$W(s) = \frac{K}{(T_1s+1)(T_2s+1)(T_3s+1)}e^{-\tau s} \tag{11}$$

Thereinto, T_1 is the time constant of the first container, $T_1=R_1A$; T_2 is the time constant of the second container, $T_2=R_2A$; T_3 is the time constant of the third container, $T_3=R_3A$; K is the magnification factor, $K=R_3$; τ is the lag time.

Suppose that: the transfer function of the system in this paper can be represented by third-order inertial lag link:

$$W(s) = \frac{2e^{-2s}}{125s^3+75s^2+15s+1} \tag{12}$$

III. FUZZY CONTROLLER

Fuzzy control is a computer intelligent control based on fuzzy set theory, fuzzy language variable and fuzzy logic inference. The basic conception of fuzzy control is proposed by the famous professor of California University of America L.A.Zadeh at the first time. Passes more than twenty years' development, significant success has been obtained in the aspect of fuzzy control theory and application research.

The basic principle block of the fuzzy control is shown in Figure 2. The core part is fuzzy controller. The control law of the fuzzy controller is realized by the computer program. The process of realization of the fuzzy control algorithm is described as follows: The accurate quantity of
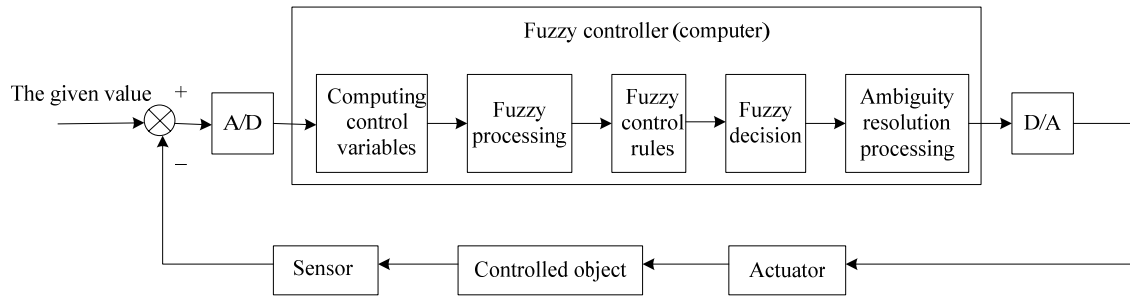
Figure 2. The basic principle block diagram of the fuzzy control

the controlled variable is sample obtained by computer, and then compares the value with the given value to obtain the error signal E which is usually selected as one of the input values of the fuzzy controller. The accuracy value of the error signal E is fuzzed to the fuzzy value. The fuzzy value of the error signal E can be expressed by the corresponding fuzzy language to obtain a subset of the fuzzy language set e (e is a fuzzy vector), and then e and the fuzzy control rule R (fuzzy operator) are used to obtain the fuzzy control value u based on the combination rule of the inference:

$$u = e \circ R \qquad (13)$$

The main difference between the fuzzy control system and the usual computer digital control system is that the fuzzy control system adopts fuzzy controller. Fuzzy controller is the core of the fuzzy control system. The performance of the fuzzy control system mainly depends on the structure of the fuzzy controller, the fuzzy rules, the compositional inference algorithm and the fuzzy decision method, ect.

Fuzzy controller is also called fuzzy logic controller. Because the fuzzy control rules are described by the fuzzy condition sentences of the fuzzy theory, the fuzzy controller is a language controller. It is also called fuzzy language controller.

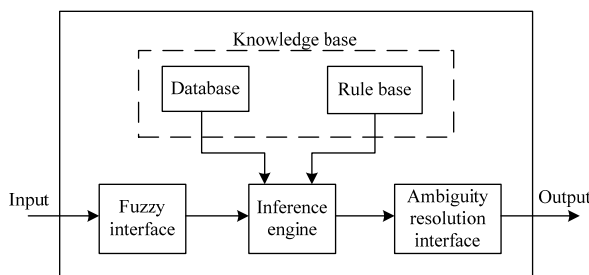The composition block diagram of the fuzzy controller is shown in Figure 3.

Figure 3. The composition block diagram of the fuzzy controller

Fuzzy interface:

The inputs of the fuzzy controller must be fuzzed to solve the output of the control. So, it is the input interface of the fuzzy controller. Its main role is changing the real input value to a fuzzy vector. For the fuzzy input variable e, its fuzzy subset is usually divided as follows:

e={Negative Big, Negative Small, Zero, Positive Small, Positive Big}={NB, NS, ZO, PS, PB}

e={Negative Big, Negative Middle, Negative Small, Zero, Positive Small, Positive Middle, Positive Big}={NB, NM, NS, ZO, PS, PM, PB}

e={Negative Big, Negative Middle, Negative Small, Negative Zero, Positive Zero, Positive Small, Positive Middle, Positive Big}={NB, NM, NS, NZ, PZ, PS, PM, PB}

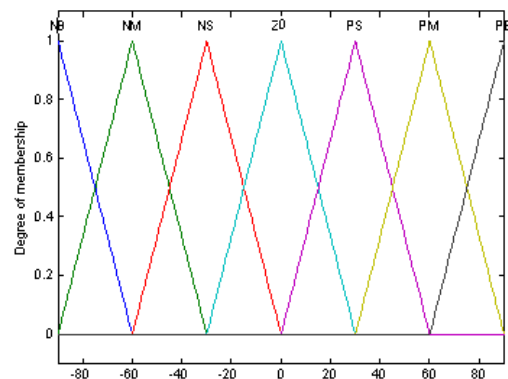The membership functions are usually expressed using triangles.

Figure 4. The membership functions of variable

Knowledge Base:

The knowledge base is composed by database and rule base.

Database:

The database stores the membership function vectors of the fuzzy subset of all the input variables and output variables. If the universe is the continuous domain, the database stores membership function. It supplies data for the inference engine in the solving process of fuzzy relationship equations of the rule inference.

Rule base:

The rules of the fuzzy controller are based on the experience of the manual operators or the knowledge of experts. It is a language expression form based on people's intuition. Fuzzy rules are usually connected by a series of relational words, such as if_then, else, also, end and or, ect. The relational words must through translating to make the fuzzy rules numerical. The commonly used relational words are if_then and also. For the multi variable fuzzy control system, the commonly used relational words are and, ect. For example, the input

variables of the fuzzy control system are e(error) and e_c(error change). Their corresponding language variables are E and EC. A group of fuzzy rules are:

R1: IF E is NB and EC is NB then U is PB

R2: IF E is NB and EC is NS then U is PM

The if⋯ part is usually called "premise part", and the then⋯ part is usually called "conclusion part". Its basic structure can be concluded to: If A and B then C. Thereinto, A is a fuzzy subset of universe U, and B is a fuzzy subset of universe V. According to the manual control experience, the control decision table R can be organized off-line. R is a fuzzy subset of the Cartesian product set U×V. The control value can be expressed as:

$$C = (A \times B) \circ R \qquad (14)$$

Thereinto, × is the fuzzy direct product operation. ∘ is the fuzzy compose operation.

The rule base is used to store the fuzzy control rules. It supplies control rules for the inference engine. The number of the rules is related to the fuzzy subset division of the fuzzy variables. The accuracy of the rule base is connected with the accuracy of the knowledge of experts.

Inference and Defuzzy-interface:

Inference is the function part of obtaining the fuzzy control values through fuzzy control rules based on the input fuzzy values and the fuzzy control rules. In the fuzzy control, the inference methods with simple operation will be adopted considering the inference time. The most basic inference is Zadeh approximate inference. It includes forward inference and backward inference. The forward inference is usually used in the fuzzy control, and the backward inference is usually used in the expert systems of the engineering fields.

The acquirement of the inference conclusion expresses that the rule inference function of the fuzzy control has been completed. But the result is a fuzzy vector. The accuracy output control value has been obtained through defuzzy. The output terminal with converting function is called Defuzzy-interface.

IV. FUZZY ADAPTIVE PID CONTROLLER

With the development of computer technology, artificial intelligence method is used to store the adjusting experience of operators to the computer. According to the site actual situation, computer can adjust PID parameters automatically. That is the expert PID controller. The controller combined the classical PID control with the advanced expert system, and realized the optimal control of the system. This control method determines the object model accurately, models the experience of operators, and adjusts PID parameters by inferring.

Because the experience of operators can't be described accurately, and the various semaphores and the evaluation index can't be expressed quantificationally, the expert PID method is limited. Fuzzy theory is the efficient path to solve the problem. The basic theory and methods of fuzzy mathematic are used to construct fuzzy sets to express the conditions of rule and operations. Then, the computer controls the systems based on the actual response

condition. Fuzzy adaptive PID control has many kinds of structure forms right now, but their working principles are basically identical.

Fuzzy adaptive PID control is on the basis of PID algorithm, through calculating the error e and the error rate e_c of the current system, reasoning using fuzzy inference rules and adjusting parameters by searching fuzzy matrix tales. In the design of fuzzy adaptive controller, the error e and the error rate e_c of the current system are inputs, which can meets the demand of self-tunning of the PID parameters based on different e and e_c in different time. The fuzzy adaptive controller modifies PID parameters on line based on fuzzy control rules. Its structure is shown in Figure 5.

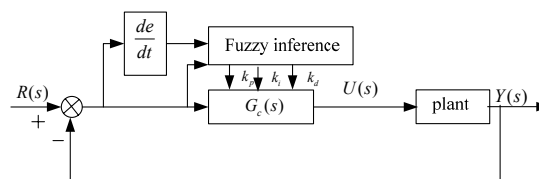

Figure 5.The structure of the fuzzy adaptive controller

The fuzzy adaptive controller is finding the relationship between PID parameters and different e, different rate e_c, detecting e and e_c constantly during the operation, and modifying the three parameters on line based on fuzzy control theory to meet different demands of control parameters for different e and different rate ec. In that way, the controlled object has good dynamic and static performance[14].

A. Fuzzy Inputting Method

Fuzzy controller was adopted in adjusting the proportion parameter k_p, integral parameter k_i and differential parameter k_d on line. The different e and different rate ec of input are fuzzed by 7 fuzzy sets. They are "PB", "PM", "PS", "Z", "NS", "NM", "NB". The output proportion parameter k_p, integral parameter k_i and differential parameter k_d are fuzzed by 7 fuzzy sets. They are "PB", "PM", "PS", "Z", "NS", "NM", "NB". The discrete universe of e and ec is {-3, -2, -1, 0, 1, 2, 3}. The discrete universes of outputs are: k_p {-0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3}, k_i {-0.06, -0.04, -0.02, 0, 0.02, 0.04, 0.06}, and k_d {-3, -2, -1, 0, 1, 2, 3}. There are 49 fuzzy control rules which are shown in table 1[14].

The fuzzy variables of the fuzzy adaptive controller adopt forms of triangle functions, S mode functions and Z mode functions.

B. Fuzzy Rules

Consider from the stability, response velocity, overshoot and steady state accuracy, the rules of K_p, K_i and K_d are:

The role of proportional coefficient K_p is speeding up the response velocity of system and improving the adjusting accuracy of the system. The bigger the K_p, the quicker the response velocity of the system, the higher the

TABLE I.   THE FUZZY INFERENCE RULES TABLE OF THE KP, KI ,KD

e_c e $\quad$ $k_p/k_i/k_d$	NB	NM	NS	Z	PS	PM	PB
NB	PB/NB/PS	PB/NB/NS	PM/NM/NB	PM/NM/NB	PS/NS/NB	Z/Z/NM	Z/Z/PS
NM	PB/NB/PS	PB/NB/NS	PM/NM/NB	PS/NS/NM	PS/NS/NM	Z/Z/NS	NS/Z/Z
NS	PM/NB/Z	PM/NM/NS	PM/NS/NM	PS/NS/NM	Z/Z/NS	NS/PS/NS	NS/PS/Z
Z	PM/NM/Z	PM/NM/NS	PS/NS/NS	Z/Z/NS	NS/PS/NS	NM/PM/NS	NM/PM/Z
PS	PS/NM/Z	PS/NS/Z	Z/Z/Z	NS/PS/Z	NS/PS/Z	NM/PM/Z	NM/PB/Z
PM	PS/Z/PB	Z/Z/NS	NS/PS/PS	NM/PM/PS	NM/PM/PS	NM/PB/PS	NB/PB/PB
PB	Z/Z/PB	Z/Z/PM	NM/PS/PM	NM/PM/PM	NM/PM/PS	NB/PB/PS	NB/PB/PB

adjusting accuracy of the system, the easier the production of overshoot, and what's more lead to the unstable of the system. If the value of K_p is too small, the adjusting accuracy will be reduced, the response velocity will be slow, the adjusting time of the system will be prolonged, and the static and dynamic characteristics will become worse.

The role of integral coefficient K_i is eliminating the steady-state error of the system. The bigger the K_i, the eliminating of the steady-state error of the system is more quickly. If K_i is too big, integral saturation phenomenon will appear in the early stage of the response process and the overshoot of the response process will be big. If K_i is too small, the steady-state error of the system will be hard to eliminate, the adjusting accuracy of the system will be affected.

The role of differential coefficient K_d is improving the dynamic characteristics of the system. The role is mainly inhibiting the deviation changing towards any direction. It can forecast the changing of the deviation. If K_d is too big, the response process will brake in advance, the adjusting time will be prolonged, and the noise immunity of the system will be decreased.

The effects of the three parameters and the relationships between them have to be considered when setting the PID parameters. The relationships among e, e_c and K_p, K_i, K_d has been concluded as follows:

When |e(t)| is big, K_p should be big too, so that the time constant and the damping coefficient of the system will decrease. The values should be not too big, otherwise, the system will be unstable. To avoid the out of range control of the system in the beginning, K_d should be small. To avoid the big overshoot, K_i can be set to zero.

When |e(t)| is medium sized, K_p should be relatively small. So, the overshoot of the system will be small. The value of K_d is important to the system this time. The value of K_d should be proper to guarantee the response speed of the system. The value of K_i can be increased properly, but not too big.

When |e(t)| is small, the values of K_p and K_i should be bigger to guarantee the good steady performance of the system. The value of K_d should be proper to avoid concussion of the system in the equilibrium.

Based on the relationships among e, e_c and K_p, K_i, K_d which were summarized above, combined with the analysis and the actual operation experience of the engineers and technicians and considering the effect of the different rate $|e_c(t)|$, table 1 was obtained which are the

fuzzy rules of regulating the three parameters of the PID controller[15].

C. Fuzzy Inference

Inference is the process of accomplishing fuzzy inference and obtaining the fuzzy control variables based on the fuzzy inputs and fuzzy control rules which are used to solve the fuzzy relation equations. Mamdani organum is adopted in this paper. Mamdani organum is a common organum in fuzzy control. It is a compositional rule of inference which takes different form in fuzzy implication relations.

Suppose that the fuzzy condition sentences are:

$\tilde{R}_i$: if x is $\tilde{A}_i$ and y is $\tilde{B}_i$ then z is $\tilde{C}_i$ (i=1,2,3 ··· n)

The inference process is:

Calculating the fuzzy implication relations: $\tilde{R}_i$ (i=1,2,3 ··· n)

Suppose that: $\tilde{A}_i = (a_1 \quad a_2 \quad \cdots \quad a_m)$, which is m dimensions row vector. $\tilde{B}_i = (b_1 \quad b_2 \quad \cdots \quad b_n)$, which is n dimensions row vector. $\tilde{C}_i = (c_1 \quad c_2 \quad \cdots \quad c_t)$, which is t dimensions row vector. Calculating the fuzzy relations of the fuzzy sets in the antecedent of $\tilde{R}_i$: $\tilde{P}_i = \tilde{A}_i \times \tilde{B}_i = \tilde{A}_i^T \wedge \tilde{B}_i$. Thereinto, $\tilde{A}_i \times \tilde{B}_i$ is the cartesian product of fuzzy vector, and the result is $m \times n$ fuzzy matrix.

Calculating the fuzzy implication relations $\tilde{P}_i \rightarrow \tilde{C}_i$ which between the antecedent and consequent of $\tilde{R}_i$. For easy, the fuzzy matrix $\tilde{P}_i$ is straightened to m × n dimensions fuzzy row vector. $\tilde{R}_i = \tilde{P}_i \rightarrow \tilde{C}_i = \tilde{P}_i^T \wedge \tilde{C}_i$.

Calculating the total fuzzy implication relation $\tilde{R}$

$$\tilde{R} = \tilde{R}_1 \cup \tilde{R}_2 \cup \cdots \cup \tilde{R}_i \cup \cdots \qquad (15)$$

Calculating the fuzzy relation between the new inputs $\tilde{A}'$ and $\tilde{B}'$.

$$\tilde{P}' = \tilde{A}' \times \tilde{B}' \qquad (16)$$

Calculating the fuzzy sets of the output value

$$\tilde{C}' = (\tilde{A}' \times \tilde{B}') \bigcirc \tilde{R} = \tilde{P}' \bigcirc \tilde{R} \qquad (17)$$

D. Ambiguity Resolution of the Output

The inference result represents the achievement of the rule inference of the fuzzy control. But the result is still a fuzzy vector. In the actual fuzzy control, there must be a certain value to control or drive the actuator. So, the fuzzy vector obtained through fuzzy inference must be

converted to obtain the clear control outputs. This process is called ambiguity resolution.

The selection of ambiguity resolution method is related with the selection of the shape of membership functions and the inference methods. Matlab provides five kinds of ambiguity resolution method: (1) centroid (area centroid method), (2) bisector (area-equal-dividing method), (3) mom (average maximum membership degree method), (4) som (small maximum membership degree method), (5) lom (big maximum membership degree method).

Centroid anti-fuzzy method is adopted in this paper.

To obtain accuracy control variables, the fuzzy method is asked to express the calculation results of the membership degree functions of the outputs. The area centroid method is taking the centroid of the area which is enclosed by the membership degree function curve and the abscissa as the finally output.

$$v_0 = \frac{\int_V v \mu_v(v) dv}{\int_V \mu_v(v) dv} \qquad (18)$$

For the discrete filter situation:

$$v_0 = \frac{\sum_{k=1}^{m} v_k \mu_v(v_k)}{\sum_{k=1}^{m} \mu_v(v_k)} \qquad (19)$$

Compared with the maximum membership degree method, the area centroid method has more smooth output inference control. The output of the system will change even if the input signal has little change.

The single loop control system with pure delay is shown in Figure 6, whose closed-loop transfer function is shown as follows:

$$\phi(s) = \frac{Y(s)}{R(s)} = \frac{G_c(s)G_0(s)e^{-\tau s}}{1 + G_c(s)G_0(s)e^{-\tau s}} \qquad (20)$$

It's characteristic equation is:

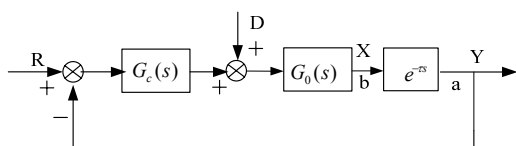$$1 + G_c(s)G_0(s)e^{-\tau s} = 0 \qquad (21)$$

Figure 6. The single loop control system with pure delay

Obviously, there is pure delay in the characteristic equation. If τ is big enough, the system will be unstable. That is the essence that the long time delay process is hard to control[9]. $e^{-\tau s}$ appears in the characteristic equation, because the feedback signal is cited from point a of the system. If the feedback signal is cited from point b, the pure delay part is removed outside of the control circuit. As shown in Figure 7. After time delay τ, the controlled variable Y will repeat changes of X.
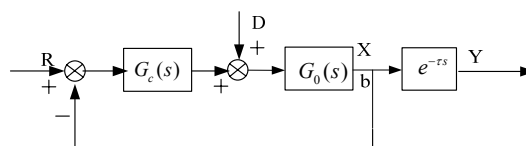
Figure 7. Improved single loop control system with pure delay

Because the feedback signal X has no delay, the response of the system is greatly improved. Point b is not exist or limited by the physical conditions in the practical system. So, the feedback signal can't be cited from point b. According to this problem, Smith proposed artificial model method. The structure is shown in Figure 8.
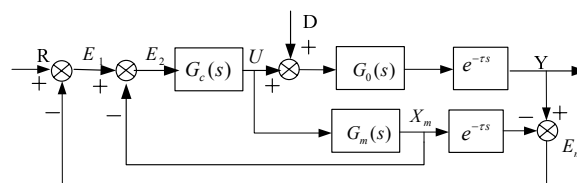

Figure 8. Smith predict control system

If the model is accurate, such as $G_0(s) = G_m(s)$, $\tau = \tau_m$, and there is no load disturbance(D=0). $Y = Y_m$, $E_m = Y - Y_m = 0$, $X = X_m$. So, X_m can change X as the first feedback loop and the pure delay part is moved outside of the control loop. If the model is inaccurate or there is load disturbance in the control process, X is not equal to X_m, $E_m = Y - Y_m \neq 0$, and the control precision is not a great satisfaction. So, E_m is used to realize the second feedback loop. This is the control strategy of Smith predictor controller[15].

V.   FUZZY ADAPTIVE SMITH CONTROLLER

The pute Smith predictive controller compensates the delay part of the time-delay system. Its PID control demands that the system must have accuracy transfer function. For the plant model with time-varying transfer function or time-varying parameters, Smith predictive controller can't achieve the ideal control effect. What's more, the Smith predictive controller is sensitive to the error. If the error is too big, the control effect will be decreased. Because the fuzzy adaptive controller adopts fuzzy set theory, it has some tolerance degree to the interference factors of itself and external. It allows the transfer function of the controlled object varying with time. The fuzzy adaptive control can work effectively whether the controlled object is linear or nonlinear. The combination of Smith predictive controller and fuzzy adaptive controller can control the overshoot of the system effectively and achieve a good control effect.

If the pute fuzzy adaptive controller is used to control the large time-delay system, the concussion of the system will be exacerbated, the adjusting time will be prolonged, and the control performance will become worse, because the control variables will have effect on the controlled object after a long time. The control variable must be responded in advance. If the Smith predictive controller is added, the delay time of the system is compensated; the

control variable can control the controlled object in real time, and the ideal control effect will be obtained.

The control problem of the pure delay time-varying system can be solved effectively when fuzzy adaptive PID controller is combined with Smith predict controller. The

adjusting time of the system will be shortened, the concussion and the overshoot of the system will be eliminated, the anti-interference ability will be enhanced, and the robustness and adaptive ability will be improved. The structure of the control system is shown in Figure 9.
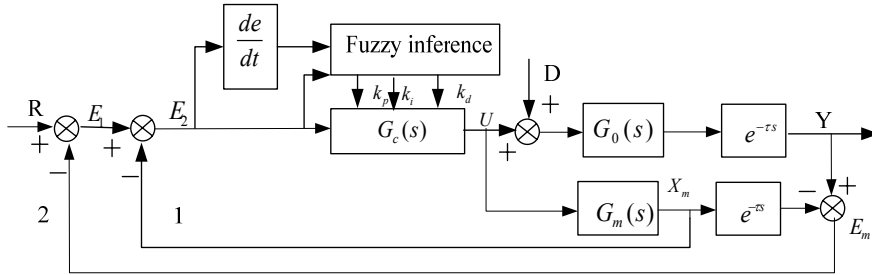

Figure 9. Fuzzy adaptive Smith-PID control system

The basic principle of the method is that: fuzzy control is used to adjust the parameters k_p 、 k_i 、 k_d of the PID controller $G_c(s)$. The output of the PID controller $G_c(s)$ is shown as follows:

$$u(k) = u(k-1) + k_p[e'(k) - e'(k-1)] +$$
$$k_i e'(k) + k_d[e'(k) - 2e'(k) + e'(k-2)] \qquad (22)$$

In the equation (21), k_p is proportion parameter, k_i is integral parameter, and k_d is differential parameter. u(k) is the output of the controller in the k th sampling time, and u(k-1) is the output of the controller in the (k-1) th sampling time.

$$e'(k) = E_m(k) \qquad (23)$$

If the model of the controlled object is accurate, and there is no load disturbance, the first feedback loop should be used, m=2. Otherwise, the second feedback loop should be used, m=1.

VI. FUZZY ADAPTIVE SMITH CONTROL SIMULATION

The mathematical model of one three-tank-system is $W(s) = \dfrac{2e^{-2s}}{125s^3 + 75s^2 + 15s + 1}$. The simulation process was performed as follows: Smith—PI method, fuzzy adaptive PID method, and fuzzy adaptive Smith—PID method were used in this paper. One unit pulse interference signal was added in the middle of the simulation process. The simulation was carried with MATLAB7.4.0. The graph of the fuzzy inference system is shown in Figure 10. The graph of the membership functions of the fuzzy adaptive PID controller is shown in Figure 11~15. The curves of the outputs of the fuzzy adaptive PID controller are shown in Figure 16~18.

The values of K_p0 and $Ki0$ were taken as 0.5 and 0.05 when taking the method of Smith—PI. The values of K_p0, K_i0 and K_d0 were taken as 0.5, 0.042 and 1.4 when taking the method of fuzzy adaptive Smith—PID. The step response curves of the system with the method of Smith—PI and fuzzy adaptive PID were shown in Figure 19 and Figure 20. The step response curve of the system with the

method of fuzzy adaptive Smith—PID was shown in Figure 21.
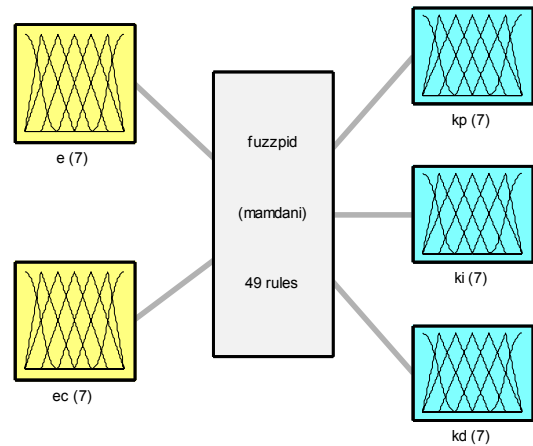
System fuzzpid: 2 inputs, 3 outputs, 49 rules

Figure 10. Fuzzy inference system

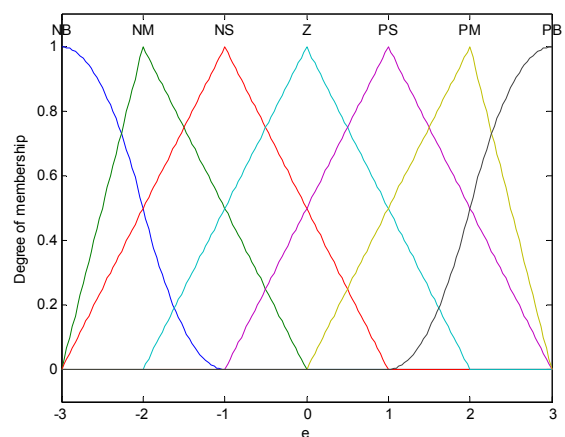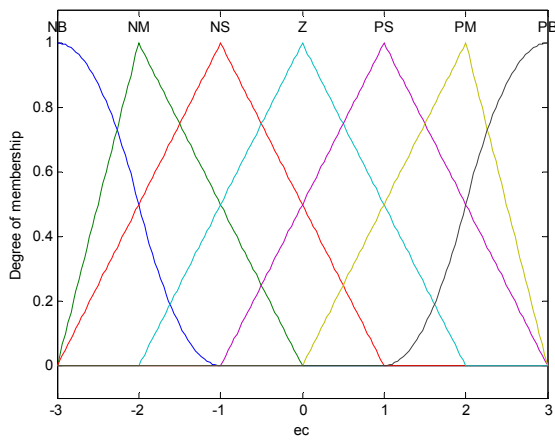Figure 11. The membership function of input e

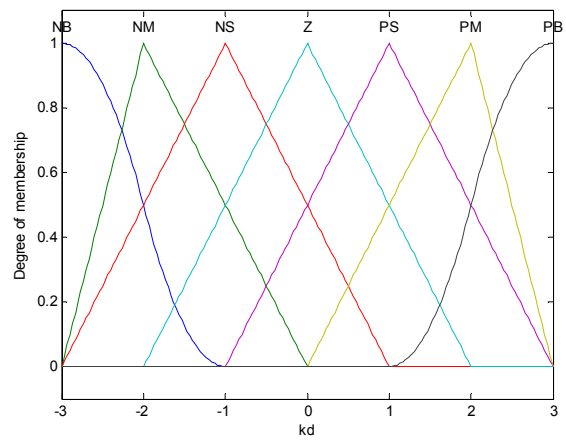Figure 12. The membership function of input ec

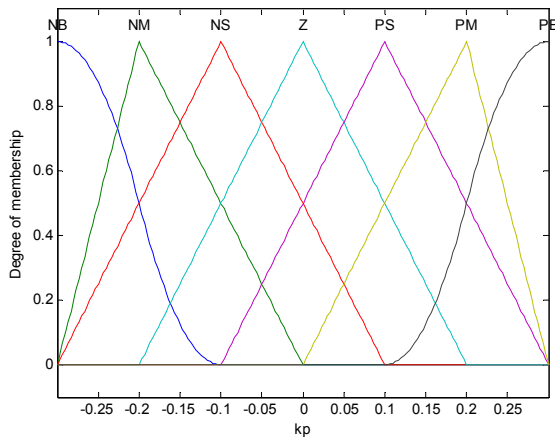Figure 15. The membership function of output K_d

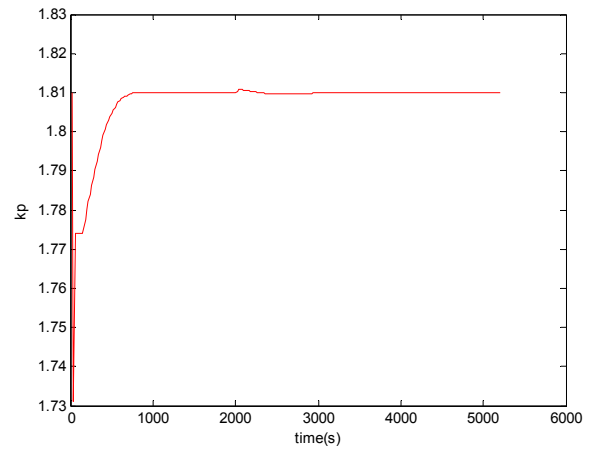Figure 13. The membership function of output K_p

Figure 16. The curve of the output K_p of the fuzzy adaptive PID controller
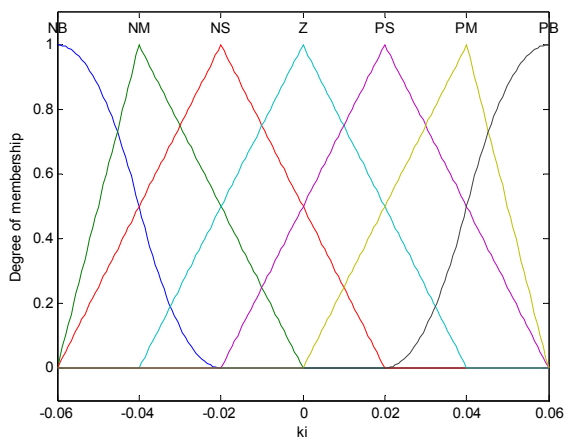

Figure 14. The membership function of output K_i
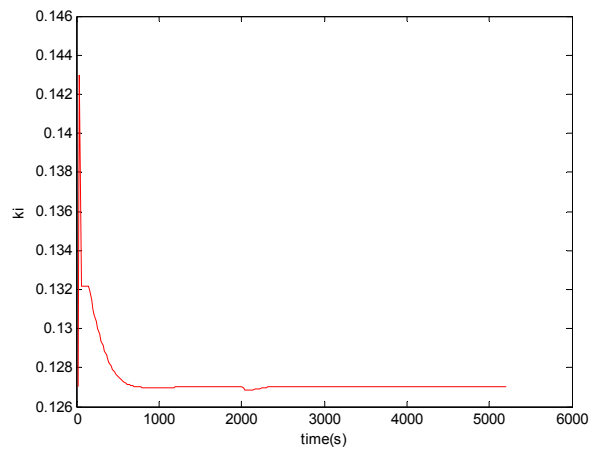
Figure 17. The curve of the output K_i of the fuzzy adaptive PID controller
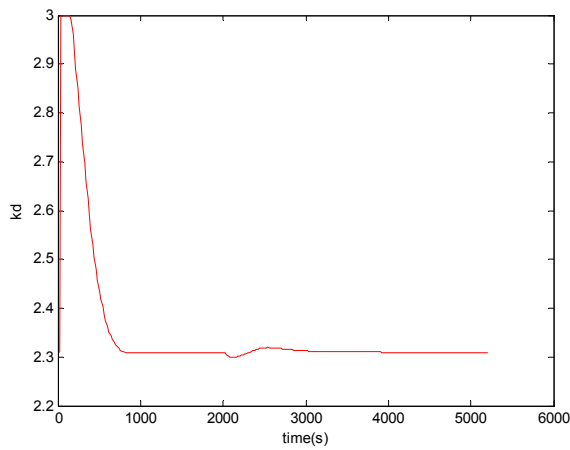
Figure 18. The curve of the output K_d of the fuzzy adaptive PID controller

Figure 19. The output curve of the system with Smith—PI method
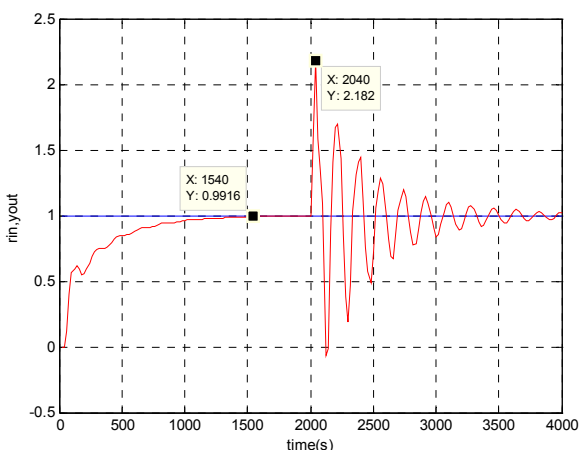
Figure 20. The output curve of the system with fuzzy adaptive PID method

Figure 21. The output curve of the system with fuzzy adaptive Smith—PID method

The figures show that: the response process of the system has no overshoot when using the method of fuzzy adaptive Smith—PID, and can reach stable in 40s. The response process of the system has big overshoot when using the method of Smith—PI, and can reach stable in 82s. The response process of the system has no overshoot when using the method of fuzzy adaptive PID, and can reach stable in 1540 s. For anti-interference, the system has better reaction effect when using the method of fuzzy adaptive Smith—PID. The simulation shows that the method of fuzzy adaptive Smith—PID can solve the contradiction of the rapidity and the stability.

VII. CONCLUSIONS

Fuzzy adaptive Smith predictor was used to control three-tank-system with delay in this paper. The fuzzy adaptive PID part can improve the resisting ability and adaptive ability of the system to random disturbance, and the Smith predict control part can overcome the delay characteristic of the controlled object. MATLAB7.4.0 is used as development platform and three-tank-system as research object in this paper. Smith—PI method, fuzzy adaptive PID method, and fuzzy adaptive Smith—PID method were used in the simulation. The simulation results show that fuzzy adaptive Smith—PID method can make the system has better adaptive ability, shorter settling time, better stability, and stronger anti-interference ability.

REFERENCES

[1] WA NG Qing-lan, Xie Xue-jun, "Smith Control of Three Tank Water System Based on C++ Builder", Journal   of Qufu Normal University(in Chinese), vol.1.2008, pp. 33-35.

[2] SUN Hong-ying, YAN De-wen, LI Bin, "The Level Control of Three Water Tanks Based on Self-Tunning Fuzzy-PID Controller", Electric Application(in Chinese), vol.25.2006, pp. 97-99.

[3] WANG Yu-hua, YU Shuang-he, DU Jia-lu, TIAN Yuan, HU Ying, "Adaptive Fuzzy Control of Uncertain Nonholonomic Mobile Robots", Journal of System Simulation(in Chinese), vol.21.2009, pp. 469-473.

[4] ZHU Ming-chao, LI Ying, LI Yuan-chun, JIANG Ri-hua, "Observer-based decentralized adaptive fuzzy control for reconfigurable manipulator", Control and Decision(in Chinese), vol.24.2009, pp. 429-434.

[5] Navid Noroozi, Mehdi Roopaei and M. Zolghadri Jahromi, "Adaptive fuzzy sliding mode control scheme for uncertain systems", Communications in Nonlinear Science and Numerical Simulation, vol.14. 2009, pp. 3978-3992.

[6] Mehdi Roopaei, Mansoor Zolghadri ,Sina Meshksar, "Enhanced adaptive fuzzy sliding mode control for uncertain nonlinear systems", Communications in Nonlinear Science and Numerical Simulation, vol.14. 2009, pp. 3670-3681.

[7] Gerasimos G. Rigatos, "Adaptive fuzzy control of DC motors using state and output feedback", Electric Power Systems Research, vol.79.2009, pp. 1579-1592.

[8] Zhijun Li, Chunquan Xu, "Adaptive fuzzy logic control of dynamic balance and motion for wheeled inverted pendulums", Fuzzy Sets and Systems, vol. 160.2009, pp. 1787-1803.

[9] Salim Labiod, Thierry Marie Guerra, "Adaptive fuzzy control of a class of SISO nonaffine nonlinear systems", Fuzzy Sets and Systems, vol.158.2007, pp. 1126-1137.

[10] Shaocheng Tong, Yongming Li, Peng Shi, "Fuzzy adaptive backstepping robust control for SISO nonlinear system with dynamic uncertainties", Information Sciences, vol.179.2009, pp. 1319-1332.

[11] H.F. Ho, Y.K. Wong, A.B. Rad, "Adaptive fuzzy sliding mode control with chattering elimination for nonlinear SISO systems", Simulation Modelling Practice and Theory, vol.17.2009, pp. 1199-1210.

[12] Tong Shaocheng, Li Changying, Li Yongming, "Fuzzy adaptive observer backstepping control for MIMO nonlinear systems", Fuzzy Sets and Systems, vol.160.2009, pp. 2755-2775.

[13] Chaio-Shiung Chen, "Dynamic structure adaptive neural fuzzy control for MIMO uncertain nonlinear systems", Information Sciences, vol.179.2009, pp. 2676-2688.

[14] LIU Jin-kun, "Advanced PID control and MATLAB simulation" (in Chinese), The second edition. Publishing House of Electronics Industry, May,2007, pp. 116.

[15] SHI Xin-min, HAO Zheng-qing, "Fuzzzy control and MATLAB simulation" (in Chinese), Publishing House of QingHua University, Mar,2008, pp. 124

Jianqiu Deng (Changde city, Hunan province, China, August 4 1974) Master's degree, Navigation, Guidance and Control, Naval Aeronautical Engineering Institute, Shandong, China, 2000. Doctorial student, Tsinghua University, Beijing, China. The major field of study: cybernation.

He teaches at Naval Aeronautical Engineering Institue now. The position is teacher. The institue is located in Shandong province of China. The article he has published is: The NCS controller based on GA optimize (Journal of Central South University, 2009). The current and previous research interests are: software engineering and software architecture.

Cui Hao (Qinhuangdao city, Hebei province, China, Mar.2 1986) Bachelor degree, Automation, Beijing Information Science and Technology University, Beijing, China, 2008. Postgraduate, Detection Technology and Automatic Equipment, Beijing Information Science and Technology University, Beijing, China. The major field of study: intelligent control and its application.

She is doing part-time job in the Zhong Neng Qing Yuan Company now. The position is **Software Engineer**. The company is located in Souhu Mansion, Beijing, China. The articles she has published are: Research on Designing of Soccer Robot's Color Tag and Identifying Algorithm (Hangzhou, China, Proceedings of 2009 International Conference on Intelligent Human –Machine Systems and Cybernetics, 2009); Fuzzy Immune Adaptive Smith-PID Control for Water Quality Adjusting System of Thermal Power Plant (Wuhan, China, The 2nd International Workshop on Intelligent Systems and Applications (ISA2010), 2010); The current and previous research interests are: the intelligent adaptive control based on the feature model; the analysis and design of the network control system.

Ms. Hao awarded the third prize in the 2006th Electronics Design Contest in Beijing.

A Self-Adaptive Differential Evolution Algorithm with Dimension Perturb Strategy

Wei-Ping Lee
Information Management Department
Chung Yuan Christian University
Chung li, Taiwan
wplee@cycu.edu.tw

Chang-Yu Chiang
Information Management Department
Chung Yuan Christian University
Chung li, Taiwan
rlong7579@hotmail.com

Abstract—**Differential Evolution (DE) has been proven to be an efficient and robust algorithm for many real optimization problems. However, it still may converge toward local optimum solutions, need to manually adjust the parameters, and finding the best values for the control parameters is a consuming task. In this paper that proposed a dimension perturb strategy and self-adaptive F value in original DE to increase the exploration ability and exploitation ability. Self-adaptive has been found to be highly beneficial for adjusting control parameters. The performance of self-adaptive differential evolution algorithm with dimension perturb strategy (PSADE) is showed on the following performance measures by benchmark functions: the solution quality and solution stability. This paper has found that PSADE can efficiently find the global value of these functions.**

Index Terms—**Differential Evolution, Dimension Perturb Strategy, Self-adaptive**

I. INTRODUCTION

The differential evolution algorithm (DE) was proposed by Storn and Price in 1996 [12] [14] [15], the algorithm is a popular optimization algorithm in recent years. Previous studies can be found that DE shows better results than other algorithms such as genetic algorithm (GA) and particle swarm optimization (PSO), in addition, DE is used in various fields widely.

Differential evolution is based on random population based global algorithm, like other evolutionary algorithm framework. It is not only well-known, robust, easy to use, fast convergence, but also requiring relatively little control parameters in solving optimization problems. Although DE has the advantages mentioned above, it encounters the crucial flaws of evolutionary computation, like trapped in the local solution easily, and adjust the parameters manually [12] [17].

In the past studies, the differential algorithm had much in common the improved methods that has improve control parameters [1] [4] [7] [16], modify the algorithm framework for process [3] [11] and with other algorithms [9] [17], the main purpose of these methods are trying to improve exploration ability and exploitation ability. Exploration means that the search to the global optimum capacity in the solution space, when the exploration capability is poor, the algorithm is easily trapped in the local optimum. And the exploitation ability means in the vicinity of global optimum solution can dig out the better solutions, if without adequate exploitation capacity that will cause instability in convergence, stretching computation time and lowering the quality of the solution.

Furthermore, the user must find the best parameter values for different problems. Finding the best values for the control parameters is a time consuming task [17].

Therefore, this paper presented a modified the control parameter and dimension perturb strategy. Through the above mentioned ways to improve

DE has to manually adjust the parameters and the optimal solution trapped into the local optimum issues. Followed by high-dimensional experiment to prove the modified algorithm can achieve a balance between the exploration ability and the exploitation ability.

The following of this paper is organized as follows: this paper briefly introduce the original DE algorithm in section II. Section III, A self-adaptive differential evolution algorithm with dimension perturb strategy (PSADE) is presented in detail. In section IV, the experimental settings and the experiment results are reported the PSADE performed on the benchmark functions. And this paper elaborates the findings in final section.

II. LITERATURE REVIEW

A. Evolution Computation

Evolutionary Computation (EC) is not only popular innovative search technology in the last decade, but also significantly alter the structure of the existing engineering computing. Evolutionary computation using computer models of evolution and selection process, to simulate Darwin "survival of the fittest" as the basis of natural biological evolution. The computer model, so-called

evolutionary algorithm, is for solving many NP-hard optimization problems in engineering and science related fields [8].

Evolutionary algorithms (EAs) are population-based search algorithms to simulate the evolution of individual select, mutate and recombine process. According to the concept of evolutionary algorithm, the developed specifications include the following steps [8]:

■ The initial phase: the initial population randomly generated.

■ Assessment phase: assessment fitness for each individual of population. If meet the termination conditions will be terminated. Otherwise, continue the following steps.

■ The selection phase: first, the individual selected from the group as a parent. Second, the parent through the various genetic operators to produce new offspring. Finally, assess the fitness value of offspring.

■ Production phase: the decision to replace some or entire individual of the current group to become the next generation and then back to assessment phase.

B. Differential Evolution Algorithm

Differential evolution is a floating-point encoding of evolutionary algorithms for continuous global optimization solution space, it also can be discrete coding manner [2] [7] [10]. Differential evolution algorithm is similar to genetic algorithm used by the individuals composed of population to search for the optimal solution. But the biggest difference between genetic algorithm and differential algorithm is the mutation operators. Genetic algorithm only use small probability to implement the mutation operator.

On the other hand, mutation in the DE is the use of arithmetic formula to combine the individual in each generation. The mutation plays the role of explores at the beginning and becomes a developer at later evolution process, its individuals within their group will become increasingly similar [6] [12].

The original DE is introduced in more detail with reference to three main operators: mutation, recombination and selection [13] [14] [15].

■ Mutation

For each individual vector, $x_i(t)$, of generation t, a mutant vector is created by

$$v_i(t) = x_{r1}(t) + F(x_{r2}(t) - x_{r3}(t)) \qquad (1)$$

where $x_{r1}(t)$, $x_{r2}(t)$ and $x_{r3}(t)$ are randomly selected with $i \neq r_1 \neq r_2 \neq r_3$, and the F is scale factor used to effect the amplification of the difference vector, $x_{r2} - x_{r3}$, $F \in [0,2]$.

■ Recombination

DE adds the diversity of population that uses a discrete recombination way where elements from the target vector, $x_i(t)$, are combined with elements from the mutant vector, $v_i(t)$, to produce the trial vector, $u_i(t)$.

$$u_{ij}(t) = \begin{cases} v_{ij}(t) & if \ rand < CR \ or \ j = r \\ x_{ij}(t) & otherwise \end{cases} \qquad (2)$$

where $j=1,...,Nd$ refers to a specific dimension, Nd is the number of parameters of single individual, and $rand \sim U(0,1)$, in the above, CR is the probability of crossover, $CR \in [0,1]$, and r is the randomly selected index, $r \sim U(1,...,Nd)$. In other words, the trial vector inherits directly some of elements of the mutant vector, Thus, even $CR=0$, at least one of the elements of the trail vector randomly selected.

■ Selection

To select the trial vector $u_i(t)$ or target vector $x_i(t)$, DE employs a very easy selection process. Only when the fitness of the trial vector is better than the target vector that the generated trial vector $u_i(t)$ replaces the target vector $x_i(t)$ be a member of the next generation. For example, if this paper has the same functions in this paper for a minimization problem, the following selections rule such as:

$$x_i(t+1) = \begin{cases} u_i(t), & f(u_i(t)) < f(x_i(t)) \\ x_i(t), & else \end{cases} \qquad (3)$$

where if the trial vector has smaller or equal fitness value than the corresponding target vector, the trial vector will replace the vector, otherwise, the target vector still remain in the population.

Recently, self-adaptive has been proven to be very useful in the automatic and dynamic adjustment of the evolution algorithm of control parameters such as mutation rate and crossover rate. Self-adaptive evolution strategy allows re-configuration to adapt to any general problems [2].

In DE, self-adaptive often used in mutation and recombination of the control parameters, the differential evolution algorithm to set the control parameters mutation and recombination than the third set control parameters on number of population is more sensitive [2].

Sum up, we can be divided into two main types of setting parameter values; parameter tuning and parameter control, parameter tuning is a more common approach is to find out the true meaning of the parameter values before the implementation of the algorithm, and then adjustment algorithm using these values. The parameter control is the parameter value during generation will change and this change can be divided into three [2][5].

■ Deterministic parameter control:

It's changed with a number of decisive rules for parameter value.

■ Adaptive parameter control:

Parameter value assignment strategy may involve some form of feedback mechanism used to determine the direction of the search with (or) significant changes in policy parameters.

■ Self-adaptive parameter control:

The evolutionary algorithm can be applied to the evolution of the concept of self-adaptive parameters. Here the parameters are coded to the individual (genes), the better values of these encoded parameters lead to

better individuals. In other words, it will be easier to survive and produce offspring.
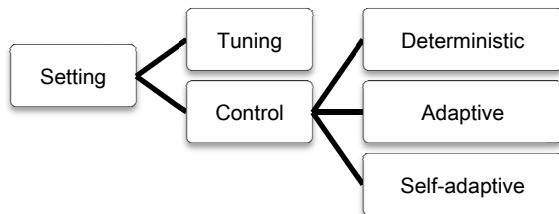

Figure 1. Types of setting parameters [5]

III. A SELF-ADAPTIVE DIFFERENTIAL EVOLUTION ALGORITHM WITH DIMENSION PERTURB STRATEGY(PSADE)

In this section, this research will improve the structure of differential algorithm and import dimension perturb strategy. This research anticipates that when the individuals trap into the local optimum, it can give individuals the opportunity to escape the local optimum. Hence, the DE can increase the diversity of population, and to continue the evolution in the same time.

In addition, this study take the most widely used DE/rand/1 strategy of differential algorithm with self-adaptive in F, F values are about with the convergence rate in solving different optimization problems. If scholars want to manually adjust to the best parameter in the implementation process, they will have to by very many experiments can be found. In other words, finding the best setting parameter value from different experiment is a very time-consuming work. Therefore, the concept is to retain good F values and let the individual of the next stage as a reference.

A. Dimension Perturb Strategy

First, the current iteration of the best individual vector carry out the action dimension perturb, that is, randomly selected to perform the exchange of two elements of the action from the best individual vector. Its main objective is that at the beginning of evolution, due to the differences of each individual vector is very large, such a mechanism can be achieved through the jumping ability to global exploration. Then in the latter part of evolution, the differences between individual vectors will slowly

shrink, and then through the perturb mechanism can enhance the population diversity and the ability to escape from the local optimum.

B. F value of Self-adaptive

The F value is related to the solution convergence rate, and different test functions may have the best value itself. The concept is to retain good F values and let the individual of the next stage as a reference.

This study assumes F normally distributed in a range with mean Fa and standard deviation 0.15. In the beginning, Fa is set at 0.5 and different F values conform this normal distribution are for each individual in the current population. The method is then calculated as

$$v_i(t) = x_{r1}(t) + F_i(t)*(x_{r2}(t) - x_{r3}(t)) \qquad (4)$$

where
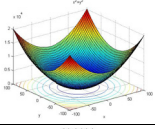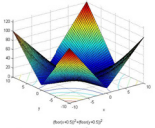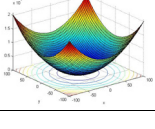
$$F_i(t) = N(Fa, 0.15) \qquad (5)$$
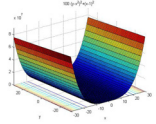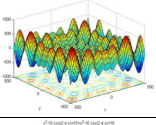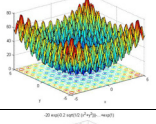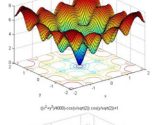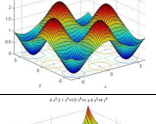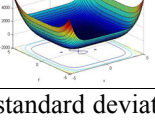
These F values for all individuals remain for several generations and then a new set of F values is generated under the same normal distribution. During every generation, the F values associated with trial vectors successfully entering the next generation are saved. After a specified number of generations, F has been changed for several times under the same normal distribution with mean Fa and standard deviation 0.15, and after every trail vector to be successful reservations via selection mechanism to the next generation, F value will be recorded along. With this new normal distribution's mean and the standard deviation 0.15, we repeat the above process. In addition, if there are no successful individuals during this period, the Fa will not change. As a result, the proper F value range for the current problem can be learned to suit this problem.

IV. EXPERIMENT RESULT

In this study, the test function can be divided into the unimodal as well as the multimodal functions, which contains small and large local optimum of two optimal types, the assessment criteria will be used to solve the solution to the mean and standard deviation for precision and stability of performance assessment.

Table I. Benchmark functions

Benchmark functions		f_{min}	S	Graphic
Sphere	$f_1(x) = \sum_{i=1}^{n} x_i^2$	0	$[-100,100]^D$	
Schwefel's problem 2.22	$f_2(x) = \sum_{i=1}^{n} |x_i| + \prod_{i=1}^{n} |x_i|$	0	$[-10,10]^D$	
Step	$f_3(x) = \sum_{i=1}^{n} (\lfloor x_i + 0.5 \rfloor)^2$	0	$[-100,100]^D$	

Rosenbrock	$f_4(x) = \sum_{i=1}^{n-1}\left(100\left(x_i - x_{i-1}^2\right)^2 + \left(x_{i-1} - 1\right)^2\right)$	0	$[-30,30]^D$	
Schwefel 's Prblem 2.26	$f_5(x) = -\sum_{i=1}^{n}\left(x_i \sin\left(\sqrt{|x_i|}\right)\right)$	-12569.5	$[-500,500]^D$	
Rastrigin	$f_6(x) = \sum_{i=1}^{n}\left(x_i^2 - 10\cos\left(2\pi x_i\right) + 10\right)$	0	$[-5.12,5.12]^D$	
Ackley's	$f_7(x) = -20\exp\left(-0.2\sqrt{\frac{1}{D}\sum_{i=1}^{D}x_i^2}\right) - \exp\left(\frac{1}{D}\sum_{i=1}^{D}\cos(2\pi x_i)\right) + 20 + e$	0	$[-32,32]^D$	
Griewank	$f_8(x) = \frac{1}{4000}\sum_{i=1}^{n}x_i^2 - \prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$	0	$[-600,600]^D$	
Six-hump Camel-back	$f_9(x) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1 x_2 - 4x_2^2 + 4x_2^4$	-1.0316285	$[-5,5]^D$	

A. Experiment with Original Differential Evolution

Table II. Experiment parameters setting with original differential evolution

NP	50
Dimension	30
F	0.5
CR	0.9; PSADE1~N(0.5,0.1)
Initial Fa	0.5
Update Fa Generation	50
Generation	1000

This section will be conducted comparison experiment with DE/rand/1, DE/best/2 and DE / current to best / 1. The CR 0.9 of PSADE refer to Storn with Price, and the PSADE1 refer to Qin and Huang, who made use of the normal distribution of mean 0.5 and standard deviation 0.1 for differential individuals in the current population [11]. Other parameters set as shown in table II.

Table V summarizes the results obtained by applying the different methods to the unimodal problems. The result shows that PSADE and PSADE1 performed better than (or at least equal to) the other strategies in all the benchmark functions except the Rosenbrock function where DE/best/2 outperformed the other strategies. However, PSADE1 performed significantly better than PSADE.

Table V summarizes the results obtained by applying the different methods to the multimodal problems. The result shows that PSADE1 performed consistently better than the other strategies in all the test functions. The improvement is even more significant for the Rastrigin function. Furthermore, examining the standard deviations of all the algorithms, PSADE1 achieved the smallest standard deviation, illustrating that PSADE and PSADE1 are more stable and thus more robust than the other versions of DE.

Fig. 2 shows average best fitness curves for the different strategies with 30 independent runs for selected benchmark functions. Fig. 2 shows that PSADE1 has a faster convergence rate than DE/rand/1, DE/best/2, DE/Current to best/1 and PSADE. On the other hand, the figure shows that PSADE1 converges insignificantly faster than DE/best/2. Hence, from table V and Fig. 2 it can be concluded that, PSADE1 reached the global optimum solution faster than the other strategies in all functions except for the Rosenbrock function.

B. Solution of the effectiveness of Fa update frequency

This section will examine the updated *Fa* effectiveness. To find out the self-adaptive update frequency will affect or not affect the performance of the solution.

Set of experimental parameters such as table III.

Table III. Experiment parameters setting with update *Fa* frequency

NP	50
Dimension	30
F	0.5
CR	0.9; PSADE1~N(0.5,0.1)
Initial Fa	0.5
Update Fa Generation	25;50;100
Generation	1000

PSADE1 can offer the best output by updating the *Fa* value strategy with every 25 iterations, on the other hand, PSADE1 advantage with every 100 iterations to update the *Fa* value strategies offer more prominent effect output in Rosenbrock function.

Solving multi-modal functions, can still be observed PSADE1 with updates every 25 iteration strategy *Fa* values are output to provide better accuracy and stability.

Fig. 3 shows average best fitness curves for the different strategies with 30 independent runs for selected benchmark multimodal functions. Fig. 3 shows that PSADE1 (25) and PASDE1 (50) achieved faster convergence rates than others. Besides, PSADE1 (25) converged insignificantly faster than PSADE1 (50).

Table VI shows the dynamic *CR* values strategy is better than fix *CR* values strategy. And different updating strategy of *Fa* frequency will also affect the accuracy and stability of the solution. Whether in the unimodal or multimodal of the benchmark functions, the updating the *Fa* value strategy with every 25 iterations provided the most outstanding results.

C. With BBDE in high-dimension experiment

Table IV. Experiment parameters Setting with BBDE[10]

NP	50
Dimension	100
F	0.5
CR	0.9; PSADE1~N(0.5,0.1)
Initial Fa	0.5
Update Fa Generation	25
NFEs	50000

In this section, the function dimension raised to 100. This paper will choose the Omran et al, who published BBDE [10] algorithm to compare the object in 2009. The article has put forward an improved mechanism and control parameters. The BBDE is based on self-adaptive differential algorithm with almost free parameter fine-tuning of the hybrid differential improvement algorithm.

The main motivation is that when encountering new problems, PSO and DE are required to optimize the control parameters set, it can be regarded as the best set parameters dependence. In addition, self-adaptive differential algorithm have been studied to show good solution results, and it can reduce the time that to find the best parameter value of each problems. Therefore, this study will compare this research, to verify the method of this study and conduct precision on the performance comparison.

Set of experimental parameters in table IV. The four of the unimodal functions, the BBDE and PSADE1 presented results superior to the traditional differential algorithm. Hence, comparison results can be seen that BBDE and PSADE1 has the advantage to solve Sphere and Schwefel's problem 2.22 through the data.

The Step function is not a continuous problem type of the unimodal function, PSADE and PSADE1 still can maintain the effectiveness of the stability of its solution.

But for the Rosenbrock function, BBDE are more prominent than other algorithms. According to unimodal functions of experimental data, PSADE1 and BBDE have faster convergence, more stable and more accurate than the traditional differential algorithm.

For the multimodal functions, the Rastrigin function and Griewank function has a lot of local optimum solutions. Table VII can be found that the PSADE1 has better accuracy than BBDE and traditional differential algorithm. On the other hand, BBDE and the traditional differential algorithm are not particularly outstanding to solve these functions.

For the Ackley function, BBDE and PSADE1 are able to meet the better stability and accuracy. Therefore, in multimodal test functions, PSADE1 algorithm presented the excellent results in the same condition. In other words, when PSADE1 encounter more complex and more local optimum solution multimodal functions, the application of perturb strategy have the opportunity to escape from the region and self-adaptive of *F* values can fast convergence in every function.

V. CONCLUSION

The differential algorithm through the self-adaptive parameter *F* and the dimension perturb strategy is to make up for the shortcomings. In order to enhance the exploration ability, exploitation ability and increase the population diversity, so that algorithms can be more appropriate speed convergence towards the optimal solution and effectively promote to escape from the local optimum.

Expectations effectively enhance the effectiveness of the algorithm for solving accuracy and stability. Meanwhile, the conclusions are sorted out from the experiment.

First, the parameters tuning is a time-consuming work, therefore, this study use self-adaptive mechanism for *F* value, it can reduce the need for different functions to do manually adjust the parameters to find the best setting time. Second, when trapped into the local optimum, the algorithm with dimension perturb strategy can enhance the capacity of the escaping from local optimum area. Third, if the problem complexity increases, the PSADE1 is able to maintain a stable and accurate status for solving effectiveness. Finally, the convergence evolution diagram can understand that the proposed algorithm has fast convergence speed. In other words, PSADE1 can be reached with less iteration to evolution.

REFERENCES

[1] Abbass, H.A. The self-adaptive Pareto differential evolution algorithm. in Evolutionary Computation, 2002. CEC '02. Proceedings of the 2002 Congress on. 2002.

[2] Brest, J., et al., Performance comparison of self-adaptive and adaptive differential evolution algorithms. Soft Computing - A Fusion of Foundations, Methodologies and Applications, 2007. 11(7): p. 617-629.

[3] Changshou, D., et al. New Differential Evolution Algorithm with a Second Enhanced Mutation Operator. in

Intelligent Systems and Applications, 2009. ISA 2009. International Workshop on. 2009.

[4] Das, S., A. Konar, and U.K. Chakraborty. Improved differential evolution algorithms for handling noisy optimization problems. in Evolutionary Computation, 2005. The 2005 IEEE Congress on. 2005.

[5] Eiben, A.E., R. Hinterding, and Z. Michalewicz, Parameter control in evolutionary algorithms. Evolutionary Computation, IEEE Transactions on, 1999. 3(2): p. 124-141.

[6] Gong, W., Z. Cai, and L. Jiang, Enhancing the performance of differential evolution using orthogonal design method. Applied Mathematics and Computation, 2008. 206(1): p. 56-69.

[7] Junhong, L. and L. Jouni. A fuzzy adaptive differential evolution algorithm. in TENCON '02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering. 2002.

[8] Kicinger, R., T. Arciszewski, and K.D. Jong, Evolutionary computation and structural design: A survey of the state-of-the-art. Computers & Structures, 2005. 83(23-24): p. 1943-1978.

[9] Luitel, B. and G.K. Venayagamoorthy. Differential evolution particle swarm optimization for digital filter design. in Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress on. 2008.

[10] Omran, M.G.H., A.P. Engelbrecht, and A. Salman, Bare bones differential evolution. European Journal of Operational Research, 2009. 196(1): p. 128-139.

[11] Qin, A.K., V.L. Huang, and P.N. Suganthan, Differential Evolution Algorithm With Strategy Adaptation for Global Numerical Optimization. Evolutionary Computation, IEEE Transactions on, 2009. 13(2): p. 398-417.

[12] Salman, A., A.P. Engelbrecht, and M.G.H. Omran, Empirical analysis of self-adaptive differential evolution. European Journal of Operational Research, 2007. 183(2): p. 785-804.

[13] Storn, R. On the usage of differential evolution for function optimization. in Fuzzy Information Processing Society, 1996. NAFIPS. 1996 Biennial Conference of the North American. 1996.

[14] Storn, R. and K. Price. Minimizing the real functions of the ICEC'96 contest by differential evolution. in Evolutionary Computation, 1996., Proceedings of IEEE International Conference on. 1996.

[15] Storn, R.,K. Price., Differential Evolution - A simple and efficient adaptive scheme for global optimization over continuous spaces. Technical report, California: International Computer Science Institute, Berkeley., 1995.

[16] Das, S., A. Konar, and U.K. Chakraborty. Two improved differential evolution schemes for faster global search. GECCO '05: Proceedings of the 2005 conference on Genetic and evolutionary computation, 2005.

[17] Zhi-Feng, H., G. Guang-Han, and H. Han. A Particle Swarm Optimization Algorithm with Differential Evolution. in Machine Learning and Cybernetics, 2007 International Conference on. 2007.

Wei-Ping Lee received the the Ph.D. degrees in Institute of Computer Science and Information Engineering from National Chiao Tung University, Hsinchu, Taiwan, R.O.C.

Currently, he has been an Assistant Professor in the Department of at Chung Yuan Christian University Chung li, Taiwan, R.O.C. His research interests include the global optimization, evolutionary algorithms, data mining and AI applications.

Chang-Yu Chiang received the master degree in Information Management from Chung Yuan Christian University, Chung li, Taiwan, R.O.C.

His interests include the global optimization, evolutionary algorithms and data mining.

Table V. Mean and standard deviation (SD) of the optimization results for 30 runs

	rand1	best2	currenttobest1	PSADE	PSADE1
Sphere	3.123e-011 (4.097e-011)	6.352e-021 (9.549e-021)	1.867e+000 (8.064e-001)	1.818e-019 (2.848e-019)	**1.712e-025** **(3.139e-025)**
Schwefel's problem 2.22	1.154e-006 (4.293e-007)	9.996e-012 (1.202e-011)	1.053e+000 (2.267e-001)	6.534673e-014 (6.144e-014)	**4.392549e-017** **(2.794e-017)**
Step	**0** **(0)**	3.633e+000 (4.319e+000)	6.100e+000 (2.482e+000)	**0** **(0)**	**0** **(0)**
Rosenbrock	1.299e+002 (2.729e+002)	**1.685e+001** **(1.661e+001)**	3.207e+002 (3.162e+002)	2.569e+002 (4.459e+002)	3.767e+001 (2.721e+001)
Schwefel 's Prblem 2.26	-7.402e+003 (9.969e+002)	-9.888e+003 (6.148e+002)	-4.958e+003 (4.100e+002)	**-1.256e+004** **(6.702e-008)**	**-1.256e+004** **(1.098e-009)**
Rastrigin	1.631e+002 (2.888e+001)	1.951e+002 (1.743e+001)	2.315e+002 (1.582e+001)	7.153e-013 (2.334e-012)	**0** **(0)**
Ackley's	1.426e-006 (7.270e-007)	1.199e+000 (7.978e-001)	2.289e+000 (5.257e-001)	8.465e-011 (9.985e-011)	**7.170e-014** **(3.272e-014)**
Griewank	2.053e-003 (4.344e-003)	1.206e-002 (1.088e-002)	9.675e-001 (5.725e-002)	6.583e-004 (2.586e-003)	**0** **(0)**
Six-hump Camel-back	**-1.0316285** **(4.5168e-016)**	**-1.0316285** **(4.5168e-016)**	**-1.0316285** **(4.5168e-016)**	**-1.0316285** **(4.5168e-016)**	**-1.0316285** **(4.5168e-016)**

(a)Sphere　　　　　　　　　　　　　　　　　　　　(b)Schwefel's problem 2.22

(c)Rastrigin　　　　　　　　　　　　　　　　　　　(d)Ackley's
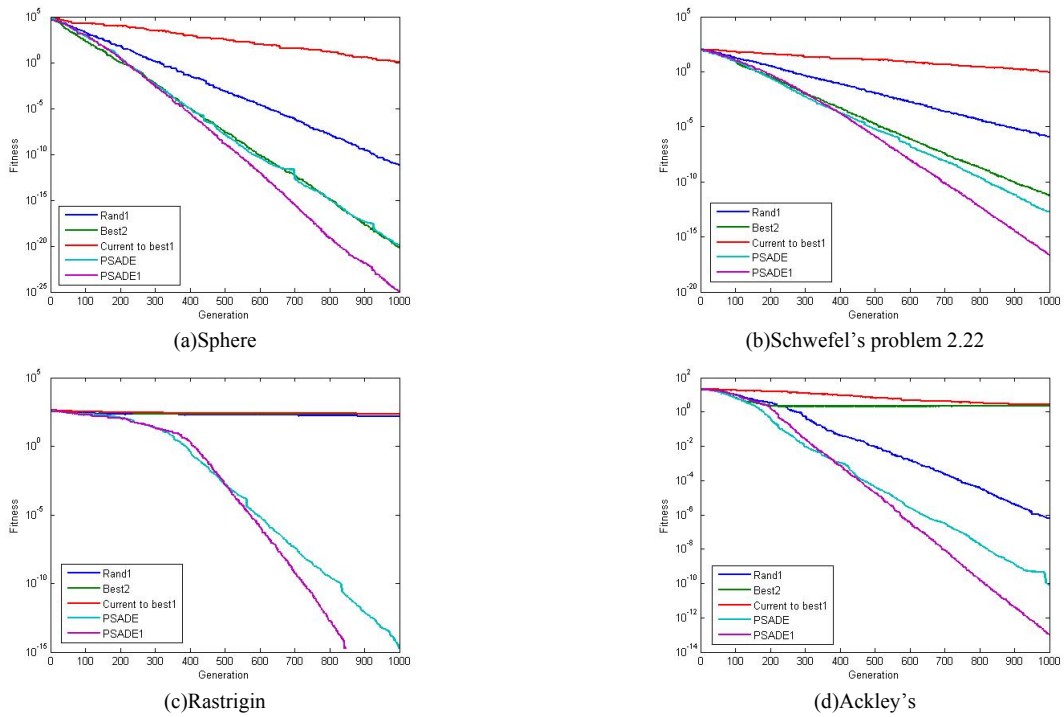
Figure 2. Convergence Graph for the optimum solution

Table VI. Mean and standard deviation (SD) of the optimization results for 30 runs

	PSADE (25)	PSADE1 (25)	PSADE (50)	PSADE1 (50)	PSADE (100)	PSADE1 (100)
Sphere	2.470e-024 (4.586e-024)	**7.899e-028** **(8.669e-028)**	1.818e-019 (2.8488e-019)	1.712e-025 (3.1394e-025)	2.858e-023 (8.199e-023)	1.572e-022 (1.349e-022)
Schwefel's problem 2.22	4.121e-016 (4.555e-016)	**6.771e-019** **(2.867e-019)**	6.534e-014 (6.144e-014)	4.392e-017 (2.794e-017)	1.585e-014 (8.835e-015)	2.490e-014 (1.273e-014)
Step	**0** **(0)**	**0** **(0)**	**0** **(0)**	**0** **(0)**	**0** **(0)**	**0** **(0)**
Rosenbrock	1.073e+002 (1.596e+002)	5.379e+001 (4.030e+001)	2.569e+002 (4.459e+002)	3.767e+001 (2.721e+001)	**3.522e+001** **(2.009e+001)**	**2.700e+001** **(4.043e-001)**
Schwefel 's Prblem 2.26	-1.233e+004 (1.288e+003)	**-1.256e+004** **(2.026e-012)**	**-1.256e+004** (6.702e-008)	**-1.256e+004** (1.098e-009)	**-1.256e+004** (4.262e-003)	**-1.256e+004** (5.210e-004)
Rastrigin	**0** **(0)**	**0** **(0)**	7.153e-013 (2.334e-012)	**0** **(0)**	1.275e-011 (6.833e-011)	1.776e-016 (7.151e-016)
Ackley's	3.457e-013 (3.114e-013)	**1.048e-014** **(2.822e-015)**	8.465e-011 (9.985e-011)	7.170e-014 (3.272e-014)	9.140e-013 (9.041e-013)	2.677e-012 (1.012e-012)
Griewank	4.351e-003 (7.216e-003)	**0** **(0)**	6.583e-004 (2.586e-003)	**0** **(0)**	2.465e-004 (1.350e-003)	**0** **(0)**

(a)Sphere

(b)Schwefel's problem 2.22

(c)Rastrigin

(d)Ackley's
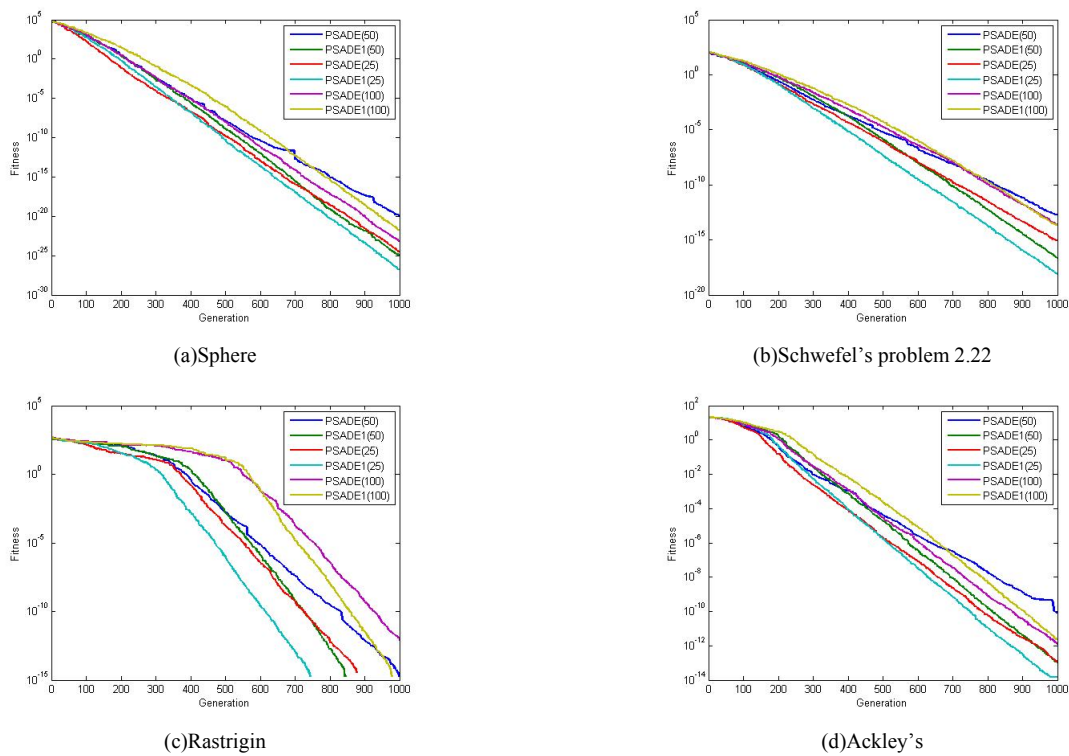
Figure 3. Convergence Graph for the optimum solution

Table VII. Mean and standard deviation (SD) of the optimization results for 30 runs

	rand1	best2	BBDE[10]	PSADE	PSADE1
Sphere	32.091642 (40.145155)	1.959712 (2.750338)	**0.000000 (0.000000)**	0.000046 (0.000045)	**0.000000 (0.000000)**
Schwefel's problem 2.22	1.544770 (1.189442)	0.760741 (0.671501)	**0.000000 (0.000000)**	0.000151 (0.000071)	**0.000000 (0.000000)**
Step	158.8333 (220.6941)	932.7 (522.7707)	2.70000 (5.01824)	**0.000000 (0.000000)**	**0.000000 (0.000000)**
Rosenbrock	85618.17 (168799.9)	72253.2 (93453.1)	**312.63207 (195.54631)**	835.0324 (525.9118)	562.1907 (475.7651)
Rastrigin	774.7837 (71.39962)	806.7959 (232.3332)	616.194754 (38.11584)	1.292809 (3.239625)	**0.001120 (0.001623)**
Ackley's	2.435670 (0.472041)	6.881239 (1.129922)	**0.000000 (0.000001)**	0.000528 (0.000187)	**0.000004 (0.000001)**
Griewank	1.168208 (0.225261)	0.465042 (0.252081)	0.001640 (0.005296)	0.009988 (0.016258)	**0.000000 (0.000000)**

Study on Partial Least-Squares Regression Model of Simulating Freezing Depth Based on Particle Swarm Optimization

Tianxiao Li

School of Water Conservancy & Civil Engineering / Northeast Agricultural University,
Harbin, China
Email: litianxiao.888@163.com

Qiang Fu [*], Fanxiang Meng, Zilong Wang and Xiaowei Wang
School of Water Conservancy & Civil Engineering / Northeast Agricultural University,
Harbin, China
Email: fuqiang100@371.net, mengfanxiang100@163.com

Abstract—In order to improve fitting and forecasting precision and solve the problem that some data with less sensitivity lead to low simulation precision of partial least-squares regression (PLS for short) model, the new method of simulating freezing depth is presented according to ground temperature of different depths, air temperatures, surface temperatures and the like. Firstly, the PLS model, which is built by virtue of the ideas of principal component analysis and canonical correlation analysis, can be adopted to solve the multi-correlation among each factor effectively by extracting principal components. And the interpretation ability of each principal component to freezing depth can be obtained by assistant analysis. Meanwhile, particle swarm optimization algorithm (PSO) is adopted to optimize partial regression coefficient, and then the PLS model based on PSO can be built. Compared with traditional PLS model, the optimized model has more reliability and stability, and higher precision.

Index Terms—freezing depth, particle swarm optimization, simulation, partial least-squares regression, principle component

I. INTRODUCTION

Frozen soil is a phenomenon that soil containing water is freezing when the temperature drops to 0 ℃ or below 0℃. Frozen soil is a kind of extremely sensitive soil medium to temperature change [1]. Therefore, soil temperature has an important effect on frozen soil, and temperature change can affect regional distribution of soil and freezing depth. The area of frozen soil, including permafrost and seasonal frost, accounts for about 20% of the earth's land area [2]. Frozen soil processes in cold regions play an important role in climate change and weather forecasting [3-5]. Seasonal freeze-thaw layer is above annual change layer of temperature, near to surface, which is more sensitive to temperature change. It is generally known that frozen soil is an important part of soil situation and soil freezing depth is closely related to agriculture, architecture, railway design, road and bridge design and so on. The impact of temperature change on frozen soil not only affects those industries above, but also environment [6-7]. Therefore, it is of great significance to study the change of freezing depth.

The traditional regression method filters the pre-decided factors and solves coefficients, but if there is a severe relevance among the pre-decided factors, the analysis results will be bad or even invalid. However, Partial Least-Squares Regression (PLS) can solve this problem better [8-10]. Because the PLS model has better simulation to some sensitivity data and the less simulation to some insensitivity data, the accuracy of partial regression analysis modeling is limited to some degree [11].

Particle Swarm Optimization (PSO), which is an evolutionary computation technology based on the swarm intelligence, is proposed by Kennedy and Eberhart in 1995 [12]. The basic idea is a swarm intelligence and parallel searching algorithm through cooperation and competition among particles. PSO has many advantages such as the faster convergence rate, satisfying results in multi-dimensional function space optimization, dynamic target optimization, so it has been widely applied in many fields [13]. In that case, PSO is adopted to optimize partial regression coefficient, and then the PSO-PLS model is built on the basis of making PLS model. At the same time, this model is applied to forecast freezing depth.

II. THE MODELING IDEAS OF PLS MODEL

A. A brief introduction of PLS model

Partial least squares (PLS) regression is a new statistical method for modeling quantitative relationships between two blocks of variables Y and X, developed by H. Wold [14-16] in 1983 and it is used as alternative to other different methods: least squares regression and principal components regression [17-20]. With this linear regression model, each dependent (response) variable Y_i in the block Y can be obtained from a linear combination of the p independent (predictor) variables X_i in the X block according to equation [21]:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (1)$$

Wherein, β_i are the regression coefficients and are determined with the calibration set and ε is the error.

B. working target

When there is only one response variable (there is only one response variable in this paper, namely, freezing depth), PLS model is short for PLS1 model [22-25].Suppose $y \in R^n$, response variable set $X = [x_1, x_2, \cdots, x_p]$, $x_j \in R^n$, if we want to adopt least squares regression to build the regression model of y to $x_1, x_2, \cdots, x_p$, the p-estimator of regression coefficients is $B = (X^T X)^{-1} X^T Y$. When all response variables are completely correlative, $(X^T X)$ is a noninvertible matrix. So the regression coefficient B can not be obtained by this formula. When all response variables are more correlative, the value of $\left| X^T X \right|$ is close to zero, then the inverse matrix of $\left| X^T X \right|$ has serious rounding error. Therefore, the least squares regression is invalid, parameter estimation can be destroyed and the robustness of model can be lost. If we adopt these data to model, the regression coefficient of x_j is always hard to explain, even appears opposite sign in real life.

This problem can be well solved by PLS model. Compared with traditional multivariate statistical methods, there are some prominent characteristics:

1) Under the condition of existing serious multi-correlation among response variables, PLS can regress.

2) PLS can be applied to regress the place where the sample number is less than the number of response variable.

3) The last model of PLS consists of all response variables

4) PLS model is easier to distinguish information and noise of system

5) Each regression coefficient is easier to be interpreted in the PLS model

C. Modeling method

Given one dependent variable y, p independent variables $\{x_1, x_2... x_p\}$ and n observing sample points, the table between $X = [x_1, x_2,...,x_p]_{n \times p}$ and $Y = [y]_{n \times 1}$ can be formed. PLS model extracts the components t_1 and u_1 from X and Y respectively. In order to meet the need of regression analysis, the following requirements must be satisfied [22-25]:

1) t_1 and u_1 ought to carry as much varied information as possible.

2) The correlation between t_1 and u_1 must reach the maximum.

PLS model does the regression of X on t_1 and Y on t_1 after extracting the first component t_1 and u_1, and the algorithm ends if the accuracy of regression equation meets the demand. Otherwise, the residual information of X and Y, which is explained by t_1, is done the second round of extracting the components and repeated until the precision reaches the satisfactory accuracy. If $t_1, t_2... t_m$ are extracted at last, PLS model will do the regression of Y on $t_1, t_2... t_m$, and then the regression equation of y about $x_1, x_2... x_p$ is represented. In the process of extracting components, the cross effectiveness is adopted to fix on the number of components.

D The cross effectiveness discriminant

Usually, PLS regression equation does not need to use all components ($t_1, t_2, \cdots, t_A$) to model and it adopts the interception way to select the previous m components ($m < A, A = rank(X)$). The model with better prediction performance can be obtained by only using these m components ($t_1, t_2, \cdots, t_A$). If the follow-up components can not offer more meaningful information for interpreting F_0 , the way of adopting more components will only undermine the understanding of the statistical trends and leading to wrong forecasting results [22-25].

In the process of building partial least squares regression model, how many components should be selected on earth? Whether it has remarkable improvement for the model's forecasting function after adding one component. Then we adopt the cross effectiveness to discriminate: all samples set after removing some sample points can be seen as a sample and a regression equation can be fitted using h components and then the eliminating sample points can be put into the previous fitting equation, the fitting value $\widehat{y}_{h(-i)}$ of y_i in the sample point i can be obtained. Each sample point repeats the above calculation. Define the sum of prediction error squares of y_i as $press_h$, namely,

$$press_h = \sum_{t=1}^{n} (y_i - \widehat{y}_{h(-i)})^2 \quad (2)$$

In addition, regression equations containing h components are fitted by all sample points. If $\hat{y}_{hi}$ is the predictive value of the ith sample point, then sum of error square of y_i can be defined as ss_h, namely,

$$ss_h = \sum_{i=1}^{n} (y_i - \hat{y}_{hi})^2 \qquad (3)$$

Generally speaking, $press_h > ss_h$, and $ss_h < ss_{h-1}$, but which is bigger, ss_{h-1} or $press_h$? ss_{h-1} is fitting error of equation with $(h-1)$ components fitted by all sample points. $press_h$ added a component b, but containing disturbance error of sample point. If $press_h$ is smaller than ss_{h-1}, to some extent, then it is considered that forecasting precision can be obviously improved by adding a component t_h. Therefore, the smaller $\dfrac{press_h}{ss_{h-1}}$ is, the better it is, namely,

$$\frac{press_h}{ss_{h-1}} \le 0.95^2 \qquad (4)$$

The cross effectiveness discriminant can also be defined as,

$$Q_h^2 = 1 - \frac{press_h}{ss_{h-1}} \qquad (5)$$

If $Q_h^2 \ge$ (1-0.9025)=0.0975, it indicates that the quality of model can be improved by adding component, otherwise not.

III. THE FUNDAMENTAL OF PSO

Suppose in D-dimension target searching space, there is a community including N particles. The position of the ith particle is represented a D-dimension vector $X = (x_{i1}, x_{i2}\dots x_{id})$; the historical optimal position of the ith particle is represented $P_i = (P_{i1}, P_{i2}\dots P_{id})$. The optimal position searched by all particles from now on is notated as $P_g = (P_{g1}, P_{g2}\dots P_{gi})$. The speed of the ith particle is also a D-dimension vector $V = (v_{i1}, v_{i2}\dots v_{id})$, it determines the displacement of particles in the searching space. Therefore, PSO is an arithmetic based on iteration idea. PSO adjusts their corresponding positions according to the following formula (6) and (7) [26-29]:

$$V_{id}^{k+1} = V_{id}^{k} + c_1 r_1 \left(p_{id}^{k} - x_{id}^{k} \right) + c_2 r_2 \left(p_{gd}^{k} - x_{gd}^{k} \right)$$

$$\begin{cases} v_{id} = -v_{\max} & v_{id} < -v_{\max} \\ v_{id} = v_{\max} & v_{id} > v_{\max} \end{cases} \qquad (6)$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \qquad (7)$$

Wherein, $1 \le i \le N$, $1 \le d \le D$, k is the iteration times($k \ge 0$); The acceleration constants c_1 and c_2 are nonnegative number; r_1 and r_2 are the random number between 0 and 1; $v_{\max}$ is a constant, which limits the max speed value. The speed of the particle is bigger when $v_{\max}$ is bigger, which is suitable for global search. But it is possible to fly across the optimal solution. If $v_{\max}$ is smaller the particle may be in a specific area to search carefully. The arithmetic ends when the max iteration time reaches the iterative limiting value or the global optimal solution is steady.

IV. THE MODELING STEPS OF PLS BASED ON PSO

Partial regression coefficients, which are the decision variables of the PSO optimization equation, are retained after building PLS model. And then the objective function on the decision variables can be obtained. The neighboring area of partial regression coefficients is the definitional domain, it will be adjusted according to the computing result and the new partial regression coefficients can be obtained. Compared with the original PLS model, the algorithm will be ended until the optimization results are satisfied.

The main modeling steps are as follows:

1) According to the ideas of PLS modeling, the freezing depth forecasting model is established. And then the partial regression coefficients are regarded as the decision variables of PSO optimization equation;

2) Fix on constraints and fitness function. The constraints are the definitional domain of each decision variable. Defined the fitness function as follows:

$$\min f_1(x) = \frac{1}{m} \sum_{i=1}^{m} [(y_{ni} - y_i)/y_i] \qquad (8)$$

$$\min f_2(x) = \frac{1}{n} \sum_{i=1}^{m} [(y_{di} - y_i)/y_i] \qquad (9)$$

Wherein, $\min f_1(x)$ is the minimum fitting relative error; $\min f_2(x)$ is the minimum reserved inspection relative error; m is the sample size; n is the reserved inspection sample size; y_{ni} is the fitting value of freezing depth, mm; y_{di} is the reserved inspection value of freezing depth, mm; y_i is the measured value of freezing depth, mm. In order to find the optimal results, (8) and (9) all need to reach minimum. Now the two fitness functions are to combine to form one function by the weight method [30], the last fitness function is as follows:

$$\min f(x) = \alpha_1 \min f_1(x) + \alpha_2 \min f_2(x)$$

$$\alpha_1 + \alpha_2 = 1 \qquad (10)$$

Wherein, α_1 and α_2 are the weights, which are fixed on by the trial method.

3) Fix on the process and control parameters of PSO, including the initialized swarm number N, the acceleration constants c_1 and c_2, the maximum flying speed limiting value v_{max} and the end iterating time k.

4) Optimization and solution build PSO-PLS model and print optimization results.

V. EXAMPLE ANALYSIS

There are many factors affecting freezing depth of soil, but soil temperature is the most important one. In winter, the surface and soil temperature reduces gradually with the decreasing of air temperature. When soil temperature is lower than freezing temperature of soil, the soil begins to freeze. Therefore, the frozen-thaw process of soil is closely related to soil temperature.

A. Data source

Freezing depth measured data (y) (For convenience, freezing layer thickness is used here) is regarded as the dependent variable, some factors such as air temperature (x_1), land surface temperature (x_2), soil temperature of 5cm depth (x_3), soil temperature of 10cm depth (x_4), soil temperature of 15cm depth (x_5), soil temperature of 20cm depth (x_6), soil temperature of 40cm depth (x_7), soil temperature of 80cm depth (x_8), soil temperature of 100cm depth (x_9), soil temperature of 140cm depth (x_{10}), soil temperature of 180cm depth (x_{11}), are selected as the independent variables. The stage data from Nov. 6, 2008 to Mar. 4, 2009 are used to build model, the data from Mar. 5, 2009 to Mar. 16, 2009 are used to inspect the accuracy of model.

According to experiment data, frozen-thaw curve and temperature change curve can be drawn as figure 1 and figure 2.

From these figures we can see that the frozen-thaw process of seasonal frozen soil can be divided into two stages, namely, unidirectional freezing stage and bidirectional melting stage. In addition, the temperature curve in whole test cycle all generate fluctuation throughout zero centigrade, causing phase transition of soil moisture and change of soil frozen-thaw process.

(1)

(2)

Figure 2. Soil temperature change curve of different depths in test period

B. Multiple correlation diagnosis

With the variance inflation factor to diagnose each independent variable, multiple correlation among them exists is checked up. Variance inflation factor is named as $(VIF)_j$ [10, 22], the expression is as follows:

$$(VIF)_j = (1 - R_j^2)^{-1} \tag{11}$$

Wherein, R_j^2 are the multiple correlation coefficients taking x_j as the dependent variable to other independent variables. The max $(VIF)_j$ in all x_j is usually regarded as the important index of the multiple correlation. If $(VIF)_j > 10$, namely $R_j^2 > 0.9$, that is to say, the multiple correlation will affect the estimation value of the least squares (LS for short) seriously and there is a highly relevant phenomena among them. The multiple

Figure 1. Soil frozen-thaw curve

correlation coefficient and variance inflation factors of each x_j are shown in Table I.

TABLE I.
MULTIPLE CORRELATION COEFFICIENT AND VARIANCE
INFLATION FACTOR OF EACH INDEPENDENT VARIABLE

variables	x_1	x_2	x_3	x_4	x_5	x_6
R_j^2	0.786	0.970	0.309	0.999	0.999	0.996
$(VIF)_j$	4.662	33.70	1.446	855.76	948.34	254.76
variables	x_7	x_8	x_9	x_{10}	x_{11}	/
R_j^2	0.998	0.998	0.997	0.999	0.997	/
$(VIF)_j$	423.36	406.52	386.07	675.02	366.28	/

From Table I, we can know that the multiple correlation coefficients of x_2, $x_4 \sim x_{11}$ are all above 0.9, and the max $(VIF)_j$ =948.3377>10, that is to say, there is a serious multiple correlation among those independent variables.

C. Build PLS model

First, the series of the independent and dependent variable are normalized by matlab7.1 [31]. Second, the PLS theory is applied to extract three principal components. The cross effectiveness result is shown in Table II.

TABLE II.
THE CROSS EFFECTIVENESS RESULT

Principal component	*1*	*2*	*3*	*4*
Calculated value	0.7785	0.6507	0.3128	-0.8164
Critical value	0.0975	0.0975	0.0975	0.0975

From Table II, we can know that the calculated value is less than 0.0975 when extracting three principal components, namely, that extracting three principal components t_1, t_2, t_3 meet the demands. And then do the linear regression of y on t_1, t_2, t_3, the multiple correlation coefficient can be obtained R=0.9903，F=0，the PLS model is as follows:

$$\hat{y} = -0.6320x_1 + 2.1330x_2 - 2.9009x_3$$
$$+ 3.1353x_4 + 3.0573x_5 + 3.4619x_6 \qquad (12)$$
$$- 0.1009x_7 - 6.7977x_8 - 11.8103x_9$$
$$- 14.3515x_{10} - 15.7887x_{11} + 79.4653$$

D. Precision analysis

The correlation coefficient square of above mentioned three principle components, x_1 to x_{11} , and dependent variable y are shown in Table III.

TABLE III.
THE CORRELATION COEFFICIENT OF PRINCIPLE
COMPONENTS, DEPENDENT VARIABLE AND INDEPENDENT
VARIABLE

r^2	x_1	x_2	x_3	x_4	x_5	x_6
t_1	0.3563	0.4020	0.1155	0.7360	0.8269	0.9407
t_2	0.4568	0.5324	0.2033	0.2283	0.1387	0.0148
t_3	0.0181	0.0001	0.2975	0.0113	0.0167	0.0286
r^2	x_7	x_8	x_9	x_{10}	x_{11}	y
t_1	0.9676	0.9513	0.9095	0.8948	0.8830	0.8190
t_2	0.0002	0.0380	0.0804	0.0867	0.0930	0.1376
t_3	0.0139	0.0005	0.0041	0.0097	0.0120	0.0185

According to Table III, accumulated interpretation ability is calculated and shown as in Table IV. From the Table IV we can see that the accumulated interpretation ability of three components to independent is 93.35%，and to dependent is 97.51%. From Table IV we can also see the independence of independent x_3. Compared with all Rd, only x_3 value is very small, just 0.6163, and the accumulated interpretation ability of three components to other independent variables and dependent variables are all up to 90%, except x_1 is 83.12%.

TABLE IV.
THE CORRELATION COEFFICIENT OF PRINCIPLE
COMPONENTS, DEPENDENT VARIABLE AND INDEPENDENT
VARIABLE

Rd	x_1	x_2	x_3	x_4	x_5	x_6	x_7
t_1	0.3563	0.4020	0.1155	0.7360	0.8269	0.9407	0.9676
t_2	0.8131	0.9344	0.3188	0.9643	0.9656	0.9555	0.9678
t_3	0.8312	0.9345	0.6163	0.9756	0.9823	0.9841	0.9817
Rd	x_8	x_9	x_{10}	x_{11}	X	y	/
t_1	0.9513	0.9095	0.8948	0.8830	0.7258	0.8190	/
t_2	0.9893	0.9899	0.9815	0.9760	0.8960	0.9566	/
t_3	0.9898	0.9940	0.9912	0.9880	0.9335	0.9751	/

E. Build PSO-PLS model

The regression coefficient is regarded as the original value of solving model. To prevent the algorithm precocious phenomena, the initial population number, the iteration time and the acceleration factors are tested and optimized repeatedly in the process of solving model. The final calculation results can be seen: the group size N= 100, the iteration time k=60, the acceleration factors $c_1 = c_2 = 1.5$. The goal function law is shown in figure 3 after

60 times iteration when $\alpha_1 = 0.81$, $\alpha_2 = 0.19$. From figure 3, we can know that the fitness function tents to be stable. The final model after optimization is as follows:

$$\hat{y} = -2.3462x_1 + 5.0112x_2 - 0.8977x_3$$
$$- 0.5006x_4 - 5.4864x_5 + 13.8440x_6 \quad (13)$$
$$+ 11.0425x_7 - 17.3525x_8 - 0.0067x_9$$
$$- 3.1467x_{10} - 39.8283x_{11} + 79.4653$$
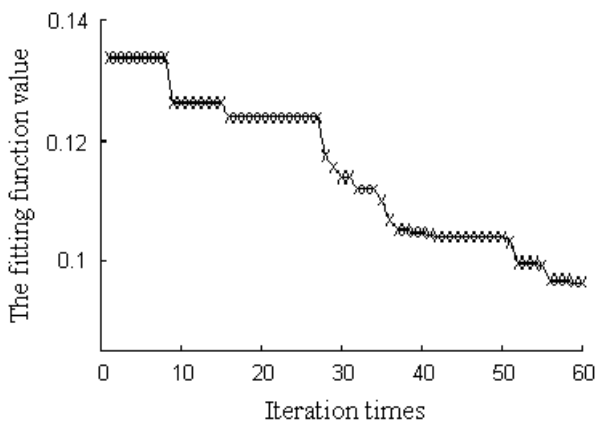
Figure 3. Variation curve of fitness function value

F. Model comparative analysis

The optimization result goes back to (1) and (2), and the fitting relative error after optimization is 10.96%, the reserved inspection relative error is 3.97%. The model precision has greatly improved. In order to compare and analyze conveniently, the comparative analysis results of forecasting model are shown in Table V between in fitting stage and in forecasting stage.

TABLE V.
THE PREDICTION MODEL OF FREEZING DEPTH COMPARISON

Model type	Fitting stage			
	relative error（%）		absolute error（cm）	
	Max	Mean	Max	Mean
PLS forecasting model	936	26.54	24.21	6.08
PSOPLS forecasting model	54.45	10.96	17.34	7.40
Model type	Forecasting stage			
	relative error（%）		absolute error（cm）	
	Max	Mean	Max	Mean
PLS forecasting model	15.14	13.00	22.94	19.59
PSOPLS forecasting model	6.82	3.97	10.20	5.97

From Table V we know that the precision of PSO-PLS model has been improved greatly not only in fitting stage but also in forecasting stage, the average relative error in fitting stage is reduced by 58.7% and 69.4% in forecasting stage. Only in fitting stage the average absolute error of PLS model is better than the PSO-PLS model, which does not affect the overall prediction level. The fitting and forecasting curves are drawn respectively (figure 4 and figure 5). From figure 4 and figure 5 we also can see that the fluctuation amplitude of PSO-PLS model fitting curve is smaller than that of PLS model and the precision of PSO-PLS model is also higher. Thereby the model validity can be proved, and it can be used in the prediction of freezing layer thickness.

Figure 4. Contrast curve of freezing layer thickness in fitting stage

Figure 5. Freezing layer thickness contrast curve in prediction stage

VI. CONCLUSION

Through theoretical analysis and calculation of PSO-PLS model, some conclusions can be obtained:

1) The variable can be processed by PLS model and the best explanation principal components can be extracted. On the one hand, the multiple correlations among variables can be solved; on the other hand, it can provide the foundation for optimizing the regression coefficient.
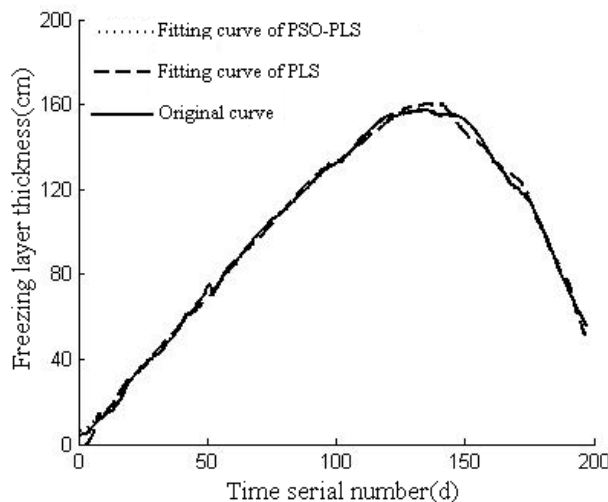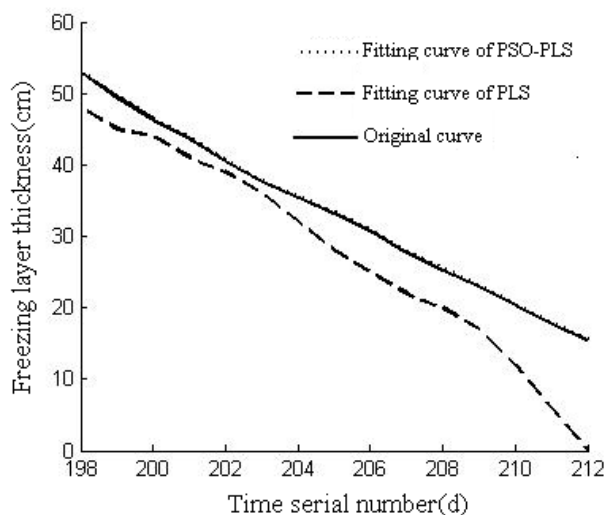
2) By assistant analysis, accumulative interpreting ability of freezing deep impact factor on freezing deep is derived. From the analysis results we can see that temperature has strongly independence, accounting for small proportion in the three principal components. According to actual observation, fluctuation of 5cm ground temperature changes greatly with temperature, which shows certain independence compared with other ground temperatures.

3) The PSO can be introduced into PLS modeling to optimize the regression coefficient. The multi-objective optimization thought is used to set up the fitness function and the multi-objective problem is solved by the trial method. This can not only improve the model fitting precision but also improve the prediction precision. The example shows that this method can provide a new approach for building freezing depth model. Therefore, it can expand the research idea of the freezing depth prediction.

REFERENCES

[1]   Guodong Cheng. Glaciology and Geocryology of China in the Past 40 Years: Progress and Prospect. Journal of Glaciolgy and Geocryology. vol. 20, no. 3, pp. 213-216, 1998

[2]   Peixoto, J., and A. H. Oort Physics of Climate. American Institute of Physics, vol. 200 pp. 20-25 1992

[3]   Mölders, N., and J. E. Walsh. Atmospheric response to soilfrost and snow in Alaska in March. Theor. Appl. Climatol., pp.77-115, 2004

[4]   Poutou, E., G. Krinner, C. Genthon, and N. de Noblet-Ducoudré. Role of soil freezing in future boreal climate change. Climate Dyn., vol. 23, pp. 621-639, 1998

[5]   Viterbo, P., A. Beljaars, J. F. Mahfouf, and J. Teixeira. The representation of soil moisture freezing and its impact on

the stable boundary layer. Quart. J. Roy. Meteor. Soc., vol. 125, pp. 2401-2426, 1999

[6]   YANG Xiaoli, WANG Jing—song. The Change Characteristics of Maximum Frozen Son Depth of Seasonal Frozen Soil in Northwest China. Chinese Journal of Soil Science. vol. 39, no. 2, pp. 238-245, 2008

[7]   XIA ZHANG, SHUFEN SUN, YONGKANG XUE. Development and Testing of a Frozen Soil Parameterization for Cold Region Studies. Journal of Hydrometeorology-special section. vol. 8, pp. 690-672, 2005

[8]   Karen C. Weber · Káthia M. Honório · Aline T. Bruni · Adriano D. Andricopulo · Albérico B. F. da Silva. A partial least squares regression study with antioxidant flavonoid compounds. Struct Chem vol. 17, pp. 307–313, 2006

[9]   Lourdes Cabezas · Miguel Angel González-Viñas · Cristina Ballesteros · Pedro J. Martín-Alvarez. Application of Partial Least Squares regression to predict sensory attributes of artisanal and industrial Manchego cheeses. Eur Food Res Technol, vol. 222, pp. 223-228, 2006

[10] Fu Qiang, Wang Zhiliang and Liang Chuan, Modeling Rice Evaporation with Partial Least-Squares Regression, Transactions of The Chinese Society of Agricultural Engineering, vol. 18, no. 6, pp. 9-12,2002

[11] Deng Nianwu, Chen Zheng and Ye Zerong, Modeling of Partial Least-Squared Regression and genetic algorithm in dam safety monitoring analysis, Dam & Safety, vol. 4, pp. 33-35, 2007

[12] Eberhart R C and Kennedy J, Swarm intelligence. San Francisco: Mornan Kaufmann Publishers, 2001

[13] Wei Xingqiong, Zhou Yongquan and Huang Huajuan. Adaptive particle swarm optimization algorithm based on cloud theory, Computer Engineering and Applications, vol. 45, no. 1, pp. 48-50, 76, 2009

[14] WoldH. In: Gani J (ed). Soft modeling by latents variables; the non-linear iterative partial least squares approach. New York, 1975

[15] I. Durán Merás · A. Espinosa Mansilla · F. Salinas López ·M. J. Rodríguez Gómez. Comparison of UV derivative-spectrophotometry and partial least-squares (PLS-1) calibration for determination of methotrexate and leucovorin in biological fluids. Anal Bioanal Chem, vol. 373 pp. 251–258, 2002

[16] Mateo Vargas · Fred A. van Eeuwijk Jose Crossa · Jean-Marcel Ribaut. Mapping QTLs and QTL × environment interaction for CIMMYT maize drought stress program using factorial regression and partial leastsquares methods. Theor Appl Genet, vol. 112, pp. 1009-1023, 2006

[17] Draper NR, Smith H. Applied regression analysis, 2nd edn. Wiley, New York, 1981

[18] Geladi P, Kowalski BR. Anal Chim Acta. vol. 185, pp. 1-17, 1986

[19] Sharaf MA, Illman DL, Kowalski BR. Chemometrics. Wiley, New York, 1986

[20] Martin-Alvarez PJ. Quimiometria alimentaria. UAM (ed) Madrid, Spain, 2000

[21] Lourdes Cabezas. Miguel Angel González-Viñas. Cristina Ballesteros. Pedro J. Martín-Alvarez. Application of Partial Least Squares regression to predict sensory attributes of artisanal and industrial Manchego cheeses. Eur Food Res Technol. vol. 222, pp. 223-228, 2006

[22] Wang Huiwen. *The application of Partial Least-Squares Regression Method*. Beijing: National Defence Industry Press, 2000

[23] Li Lin and Fu Qiang. Study on the city water resources capacity model with partial least-squares regression,

Advances in Water Science, vol. 16, no. 6, pp. 822-825, 2005

[24] Fu Qiang and Liang Chuan. *Water-Saving Irrigation System Modeling and Optimization Technique. Chengdu*: Sichuan science and technology press, 2002

[25] Fu Qiang. *The data processing method and its application in agriculture*. Beijing: Science Press, 2006

[26] Yin Guoer, Zhang Zhanyu and Zhang Guohua. Multi-objective controlled drainage model based on particle swarm optimized algorithm. Transactions of The Chinese Society of Agricultural Engineering, vol. 25, no. 3, pp. 6-9, 2009

[27] Chen Dachun and Ma Yingjie. Optimized algorithm for estimating parameters by solving Van Genuchten equation based on stochastic particle swarm optimization. Transactions of The Chinese Society of Agricultural Engineering, vol. 22, no. 12, pp. 82-85, 2006

[28] Li Zhi and Zheng Xiao. Application of Improved Particle Swarm Algorithm in Optimization Design of Agricultural Engineering. Transactions of the Chinese Society of Agricultural Engineering, vol. 20, no. 3, pp. 15-18, 2004

[29] Zhang Libiao, Zhou Chunguang and Ma Ming. Solutions of Multi-Objective Optimization Problems Based on Particle Swarm Optimization. Journal of Computer Research and Development, vol. 41, no. 7, pp. 1286-1291, 2004

[30] Xu Shuqin. Study on Water Resources Sustainable Utilization Planning Theory and Application of Irrigation Area. Northeastern Agriculture University, 2008

[31] Zhang Zhiyong. *Proficient MATLAB6.5 version*. Beijing: Beijing University of Aeronautics and Astronautics Press, 2003

Tianxiao Li, male, born in November 1984, master, mainly engaged in agriculture water and soil resources utilization and optimization and system analysis. The author obtained the bachelor degree of agriculture hydraulic project specialty in the Northeast Agricultural University in June 2007 and obtained the master degree of water and soil project in agriculture in the Northeast Agricultural University in June 2010. Recently, the author is a teacher in the Northeast Agricultural University. The number of participating in provincial projects is 2 and fulfilling actual production project is 6. More than 10 academic articles are published, in these articles, 3 articles are embodied by Engineering Index.

Qiang Fu, male, born in June 1973, professor, PhD supervisor, mainly engaged in agricultural soil and water resources system analysis, water-saving irrigation and agricultural systems engineering modeling and optimization technology. The number of presiding national and provincial projects is ten and the number of fulfilling actual production project is six. More than 150 academic articles and 6 monographs are published, in these articles, 50 articles are embodied by Engineering Index.

Fanxiang Meng, female, born in October 1984, master, mainly engaged in water-saving irrigation theory and new technology, The author obtained the bachelor degree of agriculture hydraulic project specialty in the Northeast Agricultural University in June 2007 and obtained the master degree of water and soil project in agriculture in the Northeast Agricultural University in June 2010. More than 10 academic articles are published, in these articles, 4 articles are embodied by Engineering Index.

A Novel Image Methodology for Interpretation of Gold Immunochromatographic Strip

Yurong Li

College of Electrical Engineering and Automation, Fuzhou University, Fuzhou ,China
Fujian Key Laboratory of Medical Instrumentation & Pharmaceutical Technology, Fuzhou ,China
Email: liyurong@fzu.edu.cn

Nianyin Zeng and Min Du

College of Electrical Engineering and Automation, Fuzhou University, Fuzhou ,China
Fujian Key Laboratory of Medical Instrumentation & Pharmaceutical Technology, Fuzhou ,China
Email: nianyin.zeng@gmail.com

Abstract—**Gold immunochromatographic strip assay is a rapid, simple, single-copy and on-site method. Quantitative Interpretation of the strip can provide more information than the traditional qualitative or semiquantitative strip assay. The paper aims to develop an image based assay method for quantitative determination of trace concentrations by gold immunochromatographic strip. The image of gold immunochromatographic strip is taken by CCD, and, after the proper filter and window cutting, the test line and control line is segmented by the genetic fast fuzzy *c*-means(FCM) clustering algorithm. In order to improve the measure property, based on Lambert-beer law, the relative reflective integral optical density(*RIOD*) is selected as the feature by which the interference in the test and control lines can be canceled out each other. The proposed method is applied to the quantitative detection of human chorionic gonadotropin (hCG) as a model. Firstly, the segmentation performance of the genetic fast FCM clustering algorithm is compared with threshold method and FCM clustering algorithm in terms of the peak signal-to-noise ratio (*PSNR*). Furthermore, the comparison of the blind experiment between the proposed method and commercial quantitative instrument swp-sc1 is carried out. This method is shown to deliver a result comparable and even superior to existing techniques.**

Index Terms—**gold immunochromatographic strip, quantitative interpretation, adaptive segmentation, genetic fast fuzzy c-means**

I. INTRODUCTION

An immunoassay is a biochemical test that measures the presence or concentration of a substance in solutions that frequently contain a complex mixture of substances. Analytes in biological liquids such as serum or urine are frequently assayed using immunoassay methods. Such assays are based on the unique ability of an antibody to bind with high specificity to one or a very limited group of molecules. In the field of biomedical diagnostics, the search for the increased detection sensitivity and the possibility of quantitative detection using simple inexpensive assays is an ongoing challenge.

Immunochromatographic assay, also called lateral-flow (LF) immunoassay or simple strip immunoassay[1], with benefits of low-cost, easy-to-use, rapid and sensitive detection of various analytes, has been developed for many years and mainly been used to detect drugs of abuse and for pregnancy testing at the early stage. Now, it has been surging in infectious disease diagnostics [2]. Gold immunochromatographic assay is a new lateral-flow immunoassay technique in which a cellulose membrane is used as the carrier and a colloidal gold-labeled antigen or antibody is used as the tracer. This technology has several advantages over traditional immunoassays, such as simple procedure, rapid operation and immediate results, low cost, no requirements for skilled technicians or expensive equipments and perhaps most important, well suitable for the on-site detection of antibodies or antigens[3,4]. It was first introduced to target human chorionic gonadotropin (hCG) for the detection of pregnancy. This method has been mainly applied to human diagnostics so far. It can also be applied to low molecular mass analyses, e.g., detection of drug abuse, prediction of steroid-based ovulation, and determination of progesterone, antibiotic in milk and pesticide residue according to competitive immunoassay protocols[1]. It is applied increasingly to various fields, such as clinical diagnosis, food safety testing and environment monitoring application [1,5-6].

However, unlike traditional quantitative immunoassay methods, most gold immunochromatographic assays can only get qualitative or semi-quantitative results observed directly with the naked eyes at present[7], which limit the applications of these assays. For example, the positive or negative result of female human chorionic gonadotropin (hCG) can only indicate the pregnancy status, but the

quantitative tracking result of serum hCG can provide more useful information in screening for fetal trisomy 18 also known as Edward's syndrome, open neural tube defects, and may also detect an increased risk of Turner syndrome, triploidy, trisomy 16 mosaicism, fetal death, Smith-Lemli-Opitz syndrome, and steroid sulfatase deficiency [8].

Recently, quantitative interpretation of gold immunochromatographic assay has been paid a great deal of research attention. This research mainly focuses on two directions: one is the material selection and the improvement of biochemical property of strip [9-11], and the other is the development of quantitative instrument [2,7,12-17]. It is obviously that only the progress in these two aspects can facilitate the performance of quantitative interpretation of gold immunochromatographic strip assay.

Most of the methods for developing quantitative instrument introduce the reflective optical fiber approach to acquire gold immunochromatographic strip signal, such as in [2,12-16]. It is unfortunate that the reflective optical fiber approach is based on mechanical scanning device, which causing slow operation and bulky style. The worse problem is that, the limited irradiated area of optical fiber will largely reduce the measure reliability.

Obviously, image based instrument, such as the scheme adopted in paper[7] and [17], is a good alternative, which can effectively avoid the problems emerged in the reflective optical fiber approach. For immunochromatographic strip image, the suitable image process technology is needed for anysis, among which the segmentation and feature selection are crucial.

Paper[7] and [17] provide two similar image based schemes, where the commercial optical scanner is used to obtain the strip image. The Otsu threshold selection method and fuzzy c-means (FCM) clustering algorithm were used to segment the control and test lines in paper[7] and [17], respectively. Three variables, gray intensity average of test and control, and the intensity difference between the test line and control line, are selected as features, which applied to be analyzed by the neural network model in paper [7]. In paper [17], the similar three dimensional features as in paper [7], namely, the average of RGB at control and test lines and the absolute of the difference between these averages, are the input features to the support vector regression.

The Otsu threshold selection method and FCM clustering algorithm are two common image segmentation techniques. However, in noisy situation and images of low concentration, there was difficult to infer a threshold as the range of intensities is very small, and perhaps part of the background intensity is larger than the signal, which gives rise to the main drawback of the threshold method. Although FCM clustering algorithm is widely used in the automated image segmentation[18,19], it is suffered from local optimum and is time-consuming. For the feature selection for measurement, considering only the gray intensities in test line and control lline is inadequate, because the background will provide some noise information of the strip, such as the influence of temperature, humidity and non-uniform penetration of liquid and colloidal gold.

In this paper, a novel method for interpretation of gold immunochromatographic strip, which is based on the image scheme, is investigated. The acquisition of gold immunochromatographic strip signal is taken by high gray level resolution area array Charge Coupled Device (CCD), which gets rid of the bulky mechanical scanning device. Then, we apply the averaging-based adaptive filter to reduce image noise, the canny edge detection operator, mathematical morphology open and close operation, and the hole filling algorithm to extract detection window, and finally the genetic fast FCM clustering algorithm to extract the test and control lines in the detection window. In order to improve the measure sensitivity and accuracy, then relative detection, consider the information in control and test lines and background, is chosen as the feature for quantitative interpretation of the strip based on Lambert-beer law. The property of the developed quantitative method is demonstrated by the detection of hCG and compared with our previous work. It is shown that the proposed technique has good linear response property and accuracy.

To deliver the proposed method and justify its advantages, the remainder of the paper is arranged as follows:1) characteristic of gold immunochromatographic strip, system architecture and structure of the image acquisition system are briefly discussed; 2) filter and window cutting, particularly genetic fast FCM clustering algorithm to extract test and control lines of gold immunochromatographic strip image, are described; 3) principle and feature selection of quantitative gold immunochromatographic strip based on Lambert-beer law are analyzed; and 4) the results from this method are validated on hCG immune standard sample and also compared with the commercially available quantitative instrument swp-sc1, which is our previous achievement.

II. SYSTEM DESIGN

A. Characteristic of Gold Immunochromatographic Strip

Gold immunochromatographic strip is based on colloidal gold labeled and chromatographic technology (Fig. 1). A typical immunochromatographic strip design utilizes antibody sandwich to detect antigen in the sample. In the design, one antibody is immobilized on the nitrocellulose membranes or other solid phase. The other
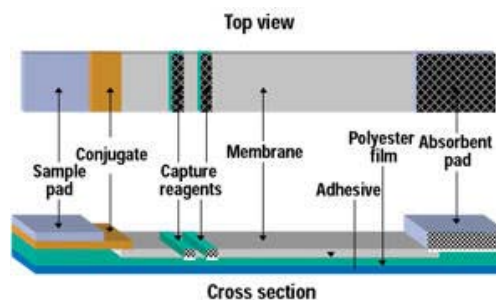

Figure 1. The schematic and structure of the gold immunochromatographic strip

antibody is labeled with colloidal gold. The labeled antibody is dried on a piece of low protein binding material in conjugate pad. When the specimen passes through and re-wet the material that is dried with the antibody conjugate, it will release the antibody conjugate. If the sample contains the antigen to be detected, the antigen will react with antibody conjugate and form an antigen-antibody conjugate compound. The compound continues to move by capillary action to the membrane where the capture antibody is embedded. The antibody will capture the antigen-antibody conjugate to form a sandwich type compound. This sandwich type compound will stay on the membrane and form a visible red or purple red color band or spot at the test and the control lines, otherwise the redundant compound goes forward to the absorbent pad.

After the antigen-antibody reactions, the red or purple red color caused by the accumulation of the colloidal gold at the test and the control lines will appear on the membrane. Monitoring the strength and area of the red color band provides the basis for the quantitative determination of the target molecule. Therefore, the concentration of the measured substance can be determined.

B. System Architecture

In the paper, gold immunochromatographic strip, and high gray level resolution area array CCD serve as the primary and secondary sensor, respectively. After the proper filter and window cutting, the test line and control line is segmented by the genetic fast fuzzy c-means(FCM) clustering algorithm from the CCD image. Finally, according to the gray intensities of control line, test line and background, the feature for the interpretation is calculated. The system block diagram is shown in Fig. 2.

C. Structure of image acquisition system

In the design of image acquisition system, the spectrum character of light source should be carefully selected according to absorption spectrum curve of the color band in order to enhance the sensitivity. The accumulation of the colloidal gold appears the red or purple red color, with absorption peak range from 510nm to 550nm, so the pure green LED with wavelength of 525nm which is complementary with red color, is selected as the light source.
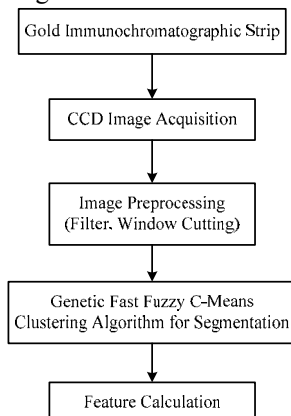
Figure 2. The system block diagram

The schematic diagram of the image acquisition system is shown in Fig. 3. In order to improve the light intensity and uniformity, sixteen LEDs are arranged in two rows, eight LEDs each row, lighting vertically on the strip.
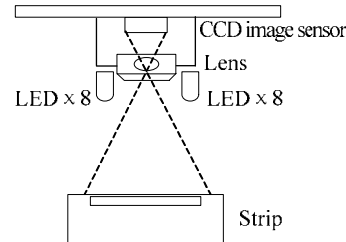
Figure 3. Schematic diagram of the image acquisition system

III. PROCESS OF GOLD IMMUNOCHROMATOGRAPHIC STRIP IMAGE

A. Filter and Window Cutting

The image taken from CCD is as Fig.4(a). Noise will be introduced to the strip image because of the limitation of device and transmission channel, which will degrade the measurement. So firstly the averaging-based adaptive filter is used[20], which can process the mixed noise composed of gauss and impulse noise, and refine the edge and detail.

Then the canny edge detection operator, mathematical morphology open and close operation, and the hole filling algorithm is applied to extract detection window, which includes the nitrocellulose membrane, test and control lines accumulated by the sandwich type compound in the nitrocellulose membrane, shown is Fig.4(b).

B. Genetic Fast Fuzzy C-Means(FCM) Clustering Algorithm to Extract Test and Control Lines

The succeeding task is the segmentation of the test line and control line from the detection window, which falls within the category of classification, that is, assigning pixels into test, control and background classes. However, the test line and control line may appear blurred and uncertain, mixed with background. According to the membership of every image pixel to the clustering centers, the FCM clustering algorithm divides the image pixels into c clustering classes through iteration [21]. Traditional FCM algorithm is an iterative optimization algorithm that minimizes the object function[22]:

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

Subjecting to $\left\{ u_{ij}\in[0,1], \sum_{i=1}^{c}u_{ij}=1, 0<\sum_{j=1}^{n}u_{ij}<n \right\}$,

where, $X=\{x_1, x_{1,...}x_n\}$ is the data set, n is the number of data points, c is the number of clusters, m $(1,\infty)$ is a weighting constant, $d_{ij}=\left\|x_j-v_i\right\|$ is the distance, which is usually Euclidian distance, between the i th cluster centre and the j th data point, u_{ij} is the membership of

vector x_j to i th cluster, $U=[u_{ij}]_{c \times n}$ is the fuzzy c-partition matrix, $V = \{v_i\}$ is the cluster centre matrix.

The FCM algorithm used in image segmentation focus on pixel set, resulting in large computation cost. By mapping from pixel space to gray histogram feature space, the clustering for pixels is mapped to the clustering for gray, reducing remarkably the computation cost and forming a fast FCM algorithm[23]. The fast FCM algorithm's formulas of object function, cluster centre and membership function are followed:

$$J_m(U,V) = \sum_{k=0}^{L_{max}-1} \sum_{i=1}^{c} (u_{ik})^m h(k) \|k - v_i\|^2 \qquad (2)$$

$$v_i = \frac{1}{\sum_{k=0}^{L_{max}-1} (u_{ik})^m h(k)} \sum_{k=1}^{L_{max}-1} (u_{ik})^m h(k)k \qquad (3)$$

$$u_{ik} = \left[\sum_{l=1}^{c} \left(\frac{\|k - v_i\|}{\|k - v_l\|} \right)^{\frac{2}{m-1}} \right]^{-1} \qquad (4)$$

where $k=\{0, 1, \cdots L_{max-1}\}$, L_{max} is the number of levels of k, $h(k)$ is the histogram of k.

Another problem is that, the FCM clustering algorithm is sensitive to initial value, easily run into local optimum. The genetic algorithm(GA) is a good global optimal tool, widely used in all kinds of field[24]. Combining GA with fast FCM algorithm, we get a hybrid algorithm which has good global search capability[25]. The procedure of genetic fast FCM algorithm, used for the gold immunochromatographic strip image segmentation in the paper, is as following:

(1)Specify the size of population and initialize the population of chromosomes randomly.

(2)Calculate the fitness for each individual in current population with the formula $f_m=1/(1+J_m)$. J_m is the formula (2).

(3)Carry out selection, crossover and mutation operations on current population.

(4)Perform the fast FCM on current generation. Firstly, calculate the new partition matrixes U of every chromosome with formula (4); Secondly, calculate the new chromosomes with the formula (3) using the new partition matrixes above.

(5)If the end condition is satisfied, return the best chromosome; otherwise, repeat step (2) through (4) until the end condition is satisfied.

Considering the image characteristic of gold immunochromatographic strip, the parameters of the segmentation algorithm in this paper are: The number of cluster $c=3$, the weighting constant $m=2$, the size of population is 30, selection operation uses the roulette wheel algorithm, crossover probability is 0.6, mutation probability is 0.02, evolution generation is 100. An example of the segmentation result is shown in Fig.4(c).
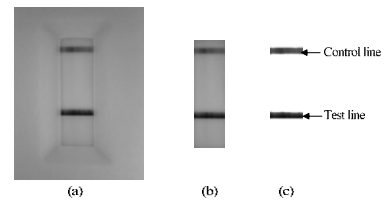
Figure 4. The extraction process of test and control lines from CCD sensor obtained GICS image.(a)Original image;(b)The result of filter and window cutting;(c)The result of using genetic fast FCM algorithm to extract the test and control lines.

IV. PRINCIPLE AND FEATURE SELECTION OF QUANTITATIVE GOLD IMMUNOCHROMATOGRAPHIC STRIP

A. *Mathimatical basis of the quantitative system*

For the parallel incident light with a given wavelength on the strip, according to Lambert-Beer Law [26]:

$$A = \alpha \cdot L \cdot C \qquad (5)$$

where A is the absorbency for the given wavelength, revealing the absorption degree of the substance to incident light. Coefficient α is the absorption coefficient, characteristic of the material. Considering a given wavelength light, it is a constant. The measured liquid is spread to a thin and semi lustrous surface, so L, thickness of absorbent substance, is also a constant. C is the concentration of the measured subject.

Consider the desirable situation, where no light transmission and scattering is occurring. Namely, besides the absorption, all the incident light is reflected. Then the absorbency A is equal to reflective integral optical density (*IOD*). Thus, if the thickness of absorbent substance L and the absorption coefficient α are known, and the reflective *IOD* is measured, the concentration of the substance can be calculated. The reflective *IOD* of strip image is defined as:

$$IOD = \sum_{i=1}^{N} OD^{(i)} = \sum_{i=1}^{N} \lg \frac{\phi_0}{\phi^{(i)}} \qquad (6)$$

where $OD^{(i)}$ is the reflective optical density of pixel i. When $C=0$, namely no absorption substance, the reflective light intensity is ϕ_0, equaling the intensity of the incident light. And $\phi^{(i)}$ is the reflective light intensity of pixel i after absorption. N is the amount of pixel. When no interference is occurring, strip background, the region outside the test and control lines, is white, fully reflects the incident light. Then the incident light intensity is equal to the reflective light intensity of background.

The output current of its photosensitive unit is direct proportional to its incident light intensity as a result of the linear photoelectric character of CCD with given wavelength. Moreover, gray level is also direct proportional to the current under uniform sampling and quantification. So the following equation is desired:

$$IOD = \sum_{i=1}^{N} OD^{(i)} = \sum_{i=1}^{N} \lg \frac{\phi_0}{\phi^{(i)}} = \sum_{i=1}^{N} \lg \frac{I_0}{I^{(i)}} = \sum_{i=1}^{N} \lg \frac{G_0}{G^{(i)}} \qquad (7)$$

where $I^{(i)}$ and $G^{(i)}$ are output current and gray level of pixel i, respectively, and I_0 and G_0, output current and gray level of background.

From (5)-(7), we can derived:

$$C = \frac{\sum_{i=1}^{N} \lg \frac{G_0}{G^{(i)}}}{\alpha \cdot L} \qquad (8)$$

It is found that in order to acquire the concentration of the measured substance, we only need to measure the gray level of every pixel at test line and background.

B. Selection of Feature Parameter

It should note that absolute values of integral optical density is depend on the physical properties of test strips and the image acquisition system used for detection and cannot necessarily be used to compare data from different experiments. The variability observed from strip to strip at the same concentration attributes to variations in membrane properties, humidity effects, batch to batch variability of the antibody conjugates, storage conditions, and other factors.

Also, the so-called Lambert-Beer deviation will appear in actual measurement. For example, the existence of transmission and scattering will decrease the reflective light intensity, and the non-uniform penetration of water, blood, and colloidal gold will cause the background to only partially reflect the incident light. And light path inference, unstable output of light source, time and temperature drifting of the photoelectric device, and the influence of strip temperature and humidity to chromatographic action will further disturb the measurement.

It is obviously that all the influence act equally on the test line and control line. So the relative integral optical density (RIOD) is introduced to cancel out the common influence. Supposed the reflective integral optical density of test and control lines are denoted by IOD_x and IOD_s respectively. And the concentration of test and control lines is C_x and C_s respectively. Then, according to Lambert-Beer Law,

$$IOD_x = \alpha \cdot L \cdot C_x, \qquad IOD_s = \alpha \cdot L \cdot C_s$$
$$RIOD = \frac{IOD_x}{IOD_s} = \frac{\alpha \cdot L \cdot C_x}{\alpha \cdot L \cdot C_s} = \frac{C_x}{C_s} \qquad (9)$$

We can set up the operating curve of $RIOD$ and C_x through several standard specimens. Then according to measured $RIOD$, we can get the needed C_x. According to (7) and (9), The selected feature, relative integral optical density (RIOD), is calculated from the test line, control line and background:

$$RIOD = \frac{IOD_x}{IOD_s} = \frac{\sum_{i=1}^{N} \lg \frac{G_0}{G_x^{(i)}}}{\sum_{i=1}^{M} \lg \frac{G_0}{G_s^{(i)}}} \qquad (10)$$

where $G_x^{(i)}$ and $G_s^{(i)}$ are gray level of pixel i at test and control lines, respectively, and G_0, gray level of background, which is calculated as the average value of background. N and M are the amount of pixels at the test and control lines respectively.

V. RESULTS AND DISCUSSION

A. Experiment design

In this study, we aim to develop a rapid and precise image system to quantitative interpretation of gold immunochromatographic strip. In this experiment, the quantification of hCG gold immunochromatographic strip is used as a model. The quantitative tracking result of hCG in serum or urine can provide more useful information in ectopic pregnancy differentiation and in fetal Down syndrome screening test.

In our experiment, all hCG specimens are diluted with hCG immune standard sample of National Institute for the Control of Pharmaceutical and Biological Products according to its specification. In every detection, the dosage of sample is strictly controlled to $90\mu l$. The reaction between antigen and antibody is dynamic, so the image in every detection is taken after 6 minutes, when the reaction is stable.

B. Performance of genetic fast FCM algorithm for image segmentation

Firstly, ten specimens with concentration of 0, 10, 35, 75, 100, 150, 200, 300, 400, 500 (mIU/ml) are selected. The gold immunochromatographic strip images taken from the ten specimens are shown in Fig. 5, and the test and control lines are extracted from image by the algorithms mentioned in Fig. 2.

In order to quantify the performance of different segmentatio methods, a quality measure is required in order to evaluate the validity of the test and control lines. For this purpose, an image quality measurement, known as the peak signal-to-noise ratio (PSNR) [27], is used and the rational is justified as follows.

The Mean Square Error (MSE) and the Peak Signal to Noise Ratio (PSNR) are the two error metrics requently used to compare image compression quality. The MSE represents the cumulative squared error between the compressed and the original image. The lower the value of MSE, the lower the error. The PSNR represents a measure of the peak error. The PSNR is most commonly used as a measure of quality of reconstruction of lossy

Figure 5. Images of gold immunochromatographic strip of different concentration(from the left side:0,10, 35, 75, 100, 150, 200, 300, 400, 500mIU/ml)

compression codes (e.g. for image compression). The signal in this case is the original data, and the noise is the error introduced by compression.

To compute the *PSNR*, the proposed algorithm first produces a binary mask that classifies the image pixels as belonging to either signal (the test and control lines, which is assigned 1) or background (which is assigned 0), then the mean-squared error using the following equation is calculated:

$$MSE = \frac{\sum_{M,N}\left[I_1(m,n) - I_2(m,n)\right]^2}{M \times N} \qquad (11)$$

where M and N are the numbers of rows and columns in the input images, respectively, I_1 is the grayscale normalized to [0 1] and I_2 is the mask image. Then, we obtain the *PSNR* using the following equation:

$$PSNR = 10\log\left[\frac{R^2}{MSE}\right] \qquad (12)$$

where R is the maximum fluctuation in the input image data type. For example, if the input image has a double-precision floating-point data type, then R is 1. If it has an 8-bit unsigned integer data type, R is 255,etc. In our situation, it is 1.

The *PSNR* returns the ratio between the maximum value in the signal and the magnitude of the signal's background noise. The output of the *PSNR* is decibel units (dB). An increase of 20dB corresponds to a ten-fold decrease in the MSE difference between two images. The higher the *PSNR* value the more strongly the binary spot mask fits with the raw image surface. The *PSNR* is a much better measure of the suitability of a spot mask to the raw image area containing the spot as it is more resilient against large intensity ranges within the spot area compared to the *MSE* [28,29].

Fig. 6 shows the *PSNR* comparison of segmenting images of 10 concentrations of specimens, where the three columns represent the peak signal to noise ratio (*PSNR*) for, respectively, the optimal threshold value algorithm, FCM algorithm and the genetic fast FCM algorithm. In the concentrations of 0,10, 35, 75, 100, 300mIU/ml, the threshold results are approximately equal

to the results of segmentation using FCM clustering algorithm. In the other four concentrations, the FCM clustering algorithm produce a better result than threshold method. Clearly, GA fast FCM clustering algorithm gives a much better segmentation outcome.

C. Construction of standard curve and blind experiment

Regardless of the method used, interpretation of the signal produced in an immunoassay requires reference to a calibrator that mimics the characteristics of the sample medium. For qualitative assays the calibrators may consist of a negative sample with no analyte and a positive sample having the lowest concentration of the analyte that is considered detectable. Quantitative assays require additional calibrators with known analyte concentrations. Comparison of the assay response of a real sample to the assay responses produced by the calibrators makes it possible to interpret the signal strength in terms of the presence or concentration of analyte in the sample.

So in our experiment, a total of ten specimens with concentration of 0, 10, 35, 75, 100, 150, 200, 300, 400, 500 (mIU/ml), are used, firstly, to test the linear property of the proposed quantitative method.

The results of *RIOD* of all the 10 specimens are calculated according to (10). The means of the calculated *RIOD* against specimen concentrations are plotted as scattered points in Fig. 7. The linearity of the quantitative method is examined by a straight line constructed using a least square method to all the data points. The line equation is:

$$RIOD = 0.00524C + 0.268 \qquad (13)$$

A linear fit to the data reveals an R^2 of 0.98995, indicating the good linearity of the response across the wide detection range.

To test the intra-assay variance, a series of 10 tests are repeated using the same batch of strips.The coefficient of variation(CV) ranged from 6.18% to 11.59% (n = 10) except the negative sample. The results in the form of mean plus standard deviation are shown in Tab.1

Then, another seven standard specimens with concentration of 50, 100, 200,250, 300, 400,450 (mIU/ml) are used to test the accuracy by blind
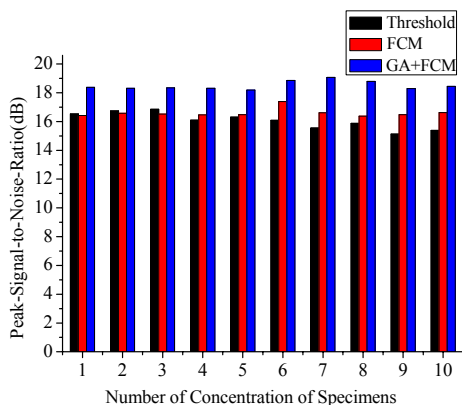

Figure 6. The PSNR comparison of segmenting images of 10 concentrations of specimens (from the left side:0,10, 35, 75, 100, 150, 200, 300, 400, 500mIU/ml)
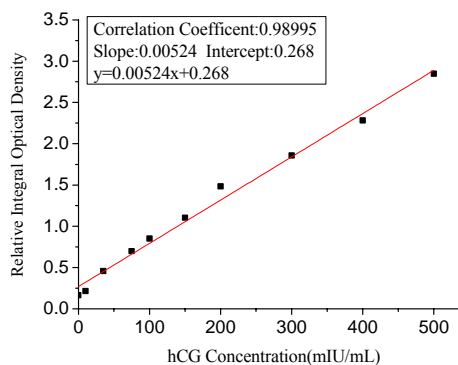

Figure 7. Correlation between the *RIOD* and hCG sample concentration, straight line constructed using least square method to all of the data points. Six repeated tests of each concentration were performed.

TABLE I.
THE *RIOD* OF DIFFERENT SPECIMEN CONCERTRATION

Sample	Mean *RIOD* ± S.D.
0 mIU/ml	0.1425 ± 0.04253
10 mIU/ml	0.2193 ± 0.01577
35 mIU/ml	0.4456 ± 0.02754
75 mIU/ml	0.7031 ± 0.07965
100 mIU/ml	0.8743 ± 0.09074
150 mIU/ml	1.0845 ± 0.10765
200 mIU/ml	1.4251 ± 0.15036
300 mIU/ml	1.7938 ± 0.19251
400 mIU/ml	2.1763 ± 0.25223
500 mIU/ml	2.8577 ± 0.26635

S.D.:standard deviation(n=10,same bach of test strips)

experiment. The detection operations are the same as above. For every specimen, the test and control lines are extracted from image, and then corresponding *RIOD* is calculated, which is used to acquire the measured hCG concentration by (13).

At the same time, these experiments are performed in parallel by the commercially available quantitative instrument swp-sc1 proposed in paper[12-14], which is our previous achievement, having acquired the *People Republic of China Registration Certificate for Medical Device*. The measured hCG concentrations of these two methods against standard hCG specimen concentrations are plotted in Fig. 8 with line of equality.
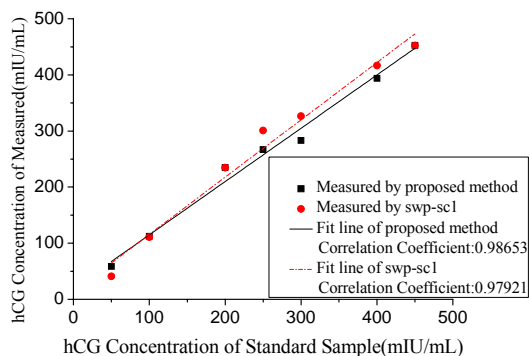

Figure 8. Comparison of hCG concentration of measured against standard specimen concentration by the blind experiment

We can find that these two methods are all performed well. The correlation of coefficient of the proposed method is 0.98653, higher than that of swp-sc1, which is 0.97921.

VI. CONCLUSION AND FUTURE WORK

To date, most of the immunochromatographic assays are qualitative or semi-quantitative. Although the intensity of the color band in test line on immunochromatographic strip correlates with concentration of the test target in particular range, the corresponding concentration level is difficult to be quantified directly. The image based quantitative method of immunochromatographic assay is a promising direction.

Based on Lambert-beer law, an image method was developed that is capable of quantitatively determining

gold immunochromatographic strip, where the test line and control line are segmented from detection window by GA fast FCM clustering algorithm, and the relative integral optical density (*RIOD*) is chosen as the measure parameter to cancel out the equal influence on test and control line by interference. The proposed quantitative method possesses linear response property and high accuracy.

The main contribution of this paper is the development of an image based method for quantitative interpretation of immunochromatographic strip. This method not only produces very impressive results that are very competitive with the results obtained using commercial instrument swp-sc1, but it also needs a more simpler structure than previous reflective optical fiber approaches. Clearly, this is a valuable addition to the area of quantitative interpretation of immunochromatographic strip.

A more detailed examination of the noise characteristic and corresponding filter algorithm is currently underway.

REFERENCES

[1] J. Zhu, W. Chen, Y. Lu, G. Cheng, "Development of an immunochromatographic assay for the rapid detection of bromoxynil in water," *Environmental Pollution*, vol. 156(1), pp. 136-142, November 2008.

[2] Zh. Yan, L. Zhou, Y. Zhao, J. Wang , L.H Huang, K.X Hu, "Rapid quantitative detection of Yersinia pestis by lateral-flow immunoassay and up-converting phosphor technology-based biosensor." *Sensors and Actuators B*.vol.119(2),pp.656–663, December 2006.

[3] H. Uto, A. Ido, K. Kusumoto, S. Hasuike, K. Nagata and et al."Development of a rapid semi-quantitative immunochromatographic assay for serum hepatocyte growth factor and its usefulness in acute liver failure," *Hepatology Research*. vol.33(4), pp. 272-276 , December 2005.

[4] D. Peng , S. Hu , Y. Hua , Y. Xiao Z. Li,X. Wang and et al. "Comparison of a new gold-immunochromatographic assay for the detection of antibodies against avian influenza virus with hemagglutination inhibition and agar gel immunodiffusion assays," *Veterinary Immunology and Immunopathology*, vol.117(1-2), pp. 17-25, May,2007.

[5] S. Huang, H. Wei, Y. Lee,"One-step immunochromatographic assay for the detection of Staphylococcus aureus,"*Food Control*, Vol. 18(8), pp. 893-897, 2007.

[6] J. Chandler, T.Gurmin T, and N.Robinson, "The place of gold in rapid tests," *IVD Technology*. vol. 6(2), pp.37-49, March,2000.

[7] L Chuang, J.Y Hwang, H.CH Chang, F.M Chang, and SH.B Jong, "Rapid and simple quantitative measurement of a-fetoprotein by combining immunochromatographic strip test and artificial neural network image analysis system," *Clinica Chimica Acta*. vol. 348, pp.87–93, October, 2004.

[8] PA Benn,"Advances in prenatal screening for Down syndrome: I. general principles and second trimester testing," *Clin. Chim. Acta*. vol.323 (1-2),pp: 1–16, September,2002.

[9] D. Li, S. Wei, H. Yang, Y. Li, and A. Deng, "A sensitive immunochromatographic assay using colloidal gold–antibody probe for rapid detection of pharmaceutical indomethacin in water samples," *Biosensors and Bioelectronics*. vol. 24(7) , pp. 2277-2280,March,2009.

[10] J. Kaur, K. Singh, R. Boro, K. Thampi, M. Raje, and G. Varshney, "Immunochromatographic dipstick assay format using gold nanoparticles labeled protein-hapten conjugate for the detection of atrazine," *Environmental Science and Technology*. vol. 41(14) , pp. 5028–5036,June,2007.

[11] R. Tanaka, T. Yuhi, N. Nagatani, T. Endo,K. Kerman, and Y. Takamura, "A novel enhancement assay for immunochromatographic test strips using gold nanoparticles," *Anal Bioanal Chem*. vol. 385(8) , pp. 1414–1420, July,2006.

[12] M. Du and ZH. Fang, "Research of the quantitative test instrument for colloidal-gold strips," *Chinese Journal of Scientific Instrument*. vol. 22(6) , pp. 626-629, December,2001.

[13] L. Huang, A. Zeng, Y. Zhang, B. Ren,K. Yu, and H. Huang, "Development of reflectance photom eter for gold-labeled test strip," *Chinese Journal of Scientific Instrument*. vol. 30(3) , pp. 663-667,March,2009.

[14] M. DU, F. Yang, and H. Fei, "Application of photoelectric sensor to quantitative determination of immunochro-matographic assay strip," *Chinese Journal of Scientific Instrument*. vol. 36(7) , pp. 671-673,July,2005.

[15] K.Faulstich, R.Gruler, M.Eberhard, and K.Haberstroh, "Developing rapid mobile POC systems. Part 1:Devices and applications for lateral-flow immunodiagnostics," *IVD Technology* .vol.13(6):47–53, July,2007.

[16] K.Faulstich, R.Gruler, M.Eberhard, and K.Haberstroh, "Developing rapid mobile POC systems. Part 2: Nucleic acid based testing platforms," *IVD Technology*. vol.13(7):47, September,2007.

[17] E. Sumonphan, S. Auephanwiriyakul, N. Theera-Umpon, "Interpretation of nevirapine concentration from immunochromatographic strip test using support vector regression," *Proceedings of 2008 IEEE International Conference on Mechatronics and Automation*,pp.633-637,August,2008.

[18] M. Forouzanfar, N. Forghani, M. Teshnehlab,"Parameter optimization of improved fuzzy c-means clustering algorithm for brain MR image segmentation," *Engineering Applications of Artificial Intelligence*. vol. 23(2), pp. 160-168, March,2010.

[19] Y. Li, Y. Shen,"An automatic fuzzy c-means algorithm for image segmentation," *Soft Computing*. vol. 14(2), pp. 123-128, January,2010.

[20] ZH. Zhang, Y. Zhang, ZH. Ma, and X. Wang, " An averaging-based adaptive filter," Journal of Image and Graphics. vol. 5(6) , pp. 530-534,June,2000.

[21] X. Zhou, Q. Shen, L. Liu, " An improved FCM method based on particle swarm optimization for color image segmentation," *Journal of Computational Information Systems*. vol. 4(5), pp. 2097-2102, October,2008.

[22] R. Hathaway., Y Hu, "Density-weighted fuzzy c-means clustering hathaway," *IEEE Transactions on Fuzzy Systems*, Vol.17(1),pp.243-252,February, 2009.

[23] ZH. Huang, Y. Xie, D. Liu, L. Hou, "Using Fuzzy C-means Cluster for Histogram-Based Color Image Segmentation," *2009 International Conference on Information Technology and Computer Science*, Vol.1 , pp.597-600,2009.

[24] H. Lau, T. Chan, W. Tsui, W. Pang, "Application of genetic algorithms to solve the multidepot vehicle routing problem," *IEEE Transactions on Automation Science and Engineering*, Vol.7(2), pp.383 -392,April,2010.

[25] C. De, B. Chakraborty, "Acoustic characterization of seafloor sediment employing a hybrid method of neural network architecture and fuzzy algorithm," *IEEE Geoscience and Remote Sensing Letters,* Vol. 6 (4),pp. 743 -747,October,2009.

[26] J. Ingle and S. Crouch, *Spectrochemical Analysis*. Prentice Hall, New Jersey, 1988.

[27] B. Zineddin, Z. Wang, K. Fraser, and X. Liu, "Investigation on filtering cDNA microarray image based multiview anaalysis," in the 14th International Conference on Automation & Computing, S. Zhang and D. Li, Eds. London, UK: Pacilantic International Ltd., 2008, pp. 201-206.

[28] K. Fraser, Z. Wang, and X. Liu, Microarray Image Analysis. Chapman & Hall/CRC, London, February ,2010.

[29] B. Zineddin, Z. Wang , Y. Li, M. Du, Y. Shi and X. Liu, "A novel neural network approach to cDNA microarray image segmentation," unpublished.

Yurong Li was born in Fujian Province, China, in 1973. She received her master degree in industry automation and PhD in control theory and control engineering from Zhejiang University, Zhejiang, Chian in 1997 and 2001, respectively.

Now she is an assistant professor at Fuzhou University. And since 2007, she is the member of Fujian Key Laboratory of Medical Instrumentation & Pharmaceutical Technology. Her research interests include biomedical instrument and intelligent information processing.
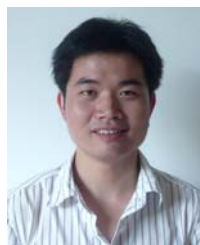
Nianyin Zeng was born in Fujian Province, China, in 1986. He received the B.S. degree in electrical engineering and automation from Fuzhou University, Fuzhou, China, in 2008, where he is currently working toward the Ph.D. degree in electrical engineering.

His current research interests include systems biology using control methods, computational biology, and bioinformatics.

Min Du was born in Fujian Province, China, in 1955. She received her PhD in electrical engineering from Fuzhou University, Fuzhou, China, in 2005.

Now she is a professor and doctorial supervisor at Fuzhou University. And since 2007, she is the associate director of Fujian Key Laboratory of Medical Instrumentation & Pharmaceutical Technology. Her research interests include smart instrument and photoelectrical system.

The Near-Optimal Preventive Maintenance Policies for a Repairable System with a Finite Life Time by Using Simulation Methods

Chun-Yuan Cheng

Department of Industrial Engineering and Management, Chaoyang University of Technology,
Taichung County, Taiwan
Email: cycheng@cyut.edu.tw

Abstract—For a repairable and deteriorating system, the theoretical preventive maintenance (PM) models are usually complicated and need numerical methods to obtain the optimal PM policies since the system's failure rate function is changed after each PM. It makes the application of the theoretical model not quite suitable for real cases. Moreover, the theoretical optimal PM solution is obtained by evaluating the expected cost rate of the system over an infinite time span. Yet, in reality, a deteriorating system always has a finite life time. Hence, searching for an optimal PM solution for a deteriorating system over an infinite time span might not be rational. Therefore, we consider using Monte Carlo simulation method to mimic the complicated failure process of a system with PM activity and then to obtain a range of the near-optimal PM policies. In this paper, the inverse transformation method and the rejection method are applied to generate the time-between-failures (TBF) random variates. The algorithms for generating the RVs are developed. The procedure of finding the near-optimal PM policies are also provided. Then, examples of using the proposed simulation method to obtain the near-optimal policies for the age-reduction PM model in a finite time span are presented and discussed.

Index Terms—Preventive maintenance, Simulation method, Time-between-failure random variates, Age reduction, Inverse Transformation Method, Rejection Method.

I. INTRODUCTION

For a deteriorating and repairable system, the system's failure rate function is changed after performing a preventive maintenance (PM) activity. Thus, the theoretical PM models are usually complicated and need numerical methods to obtain the optimal PM policies. Some theoretical PM models can be found in [1-4]. This limits the application of the theoretical model in real world. Furthermore, the theoretical optimal PM policies are obtained based on the long-term expectation of failure occurrences over the infinite time span [4-8]. Yet, in reality, the life time of a system is always finite. For systems with a finite life time, although there do exist some PM models in literature, however, the models are complicated and need numerical methods to find the optimal solution [9-12]. Hence, the optimal theoretical solution may not suitable for the real case of a single

system with finite life time. In practical, a near-optimal PM policy is good enough for applications which can be obtained by using Monte Carlo simulation method [13]. However, the research of using simulation method in solving the PM problems can not be commonly found in literature.

When using Monte Carlo simulation method to obtain a near-optimal PM policy, the critical step is to generate the time-between-failures (TBF) random variates (RV) to mimic the system's failure process which is affected by the PM activities. Percy and Kobbacy [14] investigate the scheduling of PM for repairable systems with renewals and minimal repairs models by using the simulation method. However, they did not present the RV generation method. Leemis and Schmeiser [15] describe algorithms based on the probability density function and hazard rate for generating a continuous non-negative random variates. Cheng and Liaw [16] applies the inverse transformation method to generate the RVs of the TBF for a PM model with age reduction effect. The algorithm developed by Cheng and Liaw [16] requires the derivation of the cumulative distribution function (CDF) of the TBF for each PM which is usually complicated. Cheng, Guo, and Liu [13] presented three RV generation methods for the TBF of a PM model and compared the accuracy among the three methods. They found that all three RV generation methods have high accuracy while the rejection method (or called acceptance-rejection method) is the simplest and easy-to-use method. Since the inverse transformation method and the rejection method are commonly applied in generating RVs [17, 18], in this paper, we apply these two methods to develop the algorithms of RV generation for the PM model with age reduction effect. Examples of finding the near-optimal PM policies are also provided and discussed.

II. THE THEORITICAL FEATURE OF THE PM MODEL WITH AGE REDUCTION

A. Nomenclature

L the finite life time span for the system or equipment

N the number of PM performed in the finite life time span (L)

T the time interval of each periodic PM where $T = L/(N+1)$

$x_{i,j}$ the generated time between the j-1st and the jth failures in the ith PM cycle, $j = 1, 2, \ldots, k_i$, $i = 1, 2, \ldots, N+1$.

$t_{i,j}$ the generated occurrence time of the jth failure in the ith PM cycle where $t_{i,j} = t_{i,j-1} + x_{ij}$

$u_{i,j}$

x_m the generated time between the m-1st and the mth failures, $m = 0, 1, \ldots$

t_m the generated occurrence time of the mth failures

γ the age reduced (restored) after each PM

$\lambda(t)$ The original failure rate function (before the 1st PM action)

$\lambda_i(t)$ The failure rate function at time t where t is in the ith PM cycle and $\lambda_0(t)=\lambda(t)$

$F(t)$ the cumulated distribution function (CDF) of the TBF at age t

$R(t)$ the reliability at age t

u_1 the random number for generating the RV with majorizing failure rage function, $\lambda(t)$, where $u_1 \sim Uniform(0,1)$

u_2 the random number for making the decision of rejecting the generated RV, where $u_2 \sim Uniform(0,1)$

C_{pm} Cost of each PM

C_{mr} Minimal repair cost of each failure

TC The total maintenance cost function in the finite life time span

B. Assumptions

- The system has a finite useful life time L.
- The system is deteriorating and repairable over time where the failure process follows the non-homogenous Poisson Process (NHPP) with increasing failure rate (IFR). Weibull distribution with failure rate function:

$$\lambda(t) = \left(\frac{\beta}{\theta}\right)\left(\frac{t}{\theta}\right)^{\beta-1}. \qquad (1)$$

is used to illustrate the examples in this paper, where β is the shape parameter and θ is the scale parameter.

- The periodic PM actions with constant interval (T) are performed over the finite time span L.
- The system's age can have a younger age (called the effective age) by reducing γ units of time after each PM. Hence, the failure rate function at time t in the ith PM cycle can be written as

$$\lambda_i(t) = \lambda(t - i\gamma). \qquad (2)$$

- Minimal repair is performed when failure occurs between each PM.
- The time required for performing PM, minimal repair, or replacement is negligible.

C. The PM Model and the Theoretical Optimal Policy

The PM model with age reduction is applied in this paper. Figure 1 illustrates the failure rate function of this PM model with 3 PM actions. In order to study the accuracy of near-optimal policies obtained from the proposed simulation method, we have to find the theoretical optimal policies for the age-reduction PM
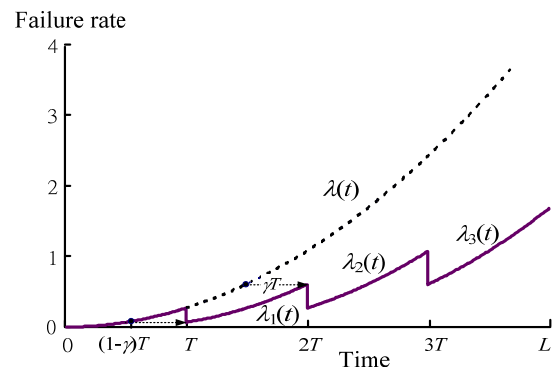
Figure 1. The failure rate function of the age-reduction PM Model with N=3.

model over a finite time span. The theoretical optimal policies are obtained based on Cheng, Liaw, and Wang [19] and Yeh and Chen [11] where the decision variables are the number of PM (N) and the restored age (γ) for each PM. We summarize the procedure for finding the theoretical optimal PM policy over a finite time span as follows.

The first step is to find the expected cost function for the PM model as shown below.

$$C(N,\gamma) = (N-1)C_{pm} + C_{pr} + C_{mr}\Lambda(N,\gamma), \qquad (3)$$

where $\Lambda(N,\gamma)$ is the expected number of failures occurred in the finite time span L and is defined as

$$\Lambda(N,\gamma) = \sum_{i=0}^{N-1} \int_{iT}^{(i+1)T} \lambda_i(t)dt = \sum_{i=0}^{N-1} \int_{iT}^{(i+1)T} \lambda(t-i\gamma)dt \qquad (4)$$

where $T = L/(N+1)$. Second step is to obtain the restored age of the PM effect (γ) as a function of N by taking the partial derivative of γ of the expected cost function shown in (3) and letting it equal to zero, i.e.,

$$\frac{\partial C(N,\gamma)}{\partial \gamma} = 0.$$

Third, since the cost function is a convex function, the optimal value γ^* and N^* of the theoretical PM policy can be obtained by numerically searching

$$\min_{N} C(N,\gamma), \quad N = 1, 2, \ldots$$

III. THE NEAR-OPIMAL PM POLICIES BY THE SIMULATION MEHTOD

A. The Concept of Random Variate Generation

There exist some useful methods for generating random variate with specific distribution, such as the inverse transformation method, the composition (linear combination) method, and the rejection method. The inverse transformation method is generally applicable and can be computationally efficient if the CDF can be analytically inverted, but may be complicated in computation for some probability distributions. The basic

algorithm of the inverse transformation method [15, 17, 18] is listed as follows.

(1) Invert the CDF $F(t)$.
(2) Generate $u \sim Uniform(0,1)$.
(3) Obtain t where $t = F^{-1}(1-u)$.

The composition method is typically used when the probability density function (pdf) can be written as a convex combination of n other pdf's. The rejection method is usually applied in cases where the form of probability density function (pdf), $f(x)$, makes the inverse transformation method difficult to use.

The rejection technique requires finding a majorizing function $f^*(t)$ which bounds the pdf $f(t)$, i.e., $f^*(t) > f(t) \; \forall \; t$. The majorizing function must integrate to a finite value so that it can be scaled to be a pdf, $g(t)$, i.e.,

$$g(t) = \frac{f^*(t)}{\int_0^\infty f^*(\tau)d\tau}.$$

The rejection method of using the majorizing function is illustrated in Figure 2. Values are generated from $g(t)$, then accepted or rejected so that the accepted random variates will have pdf $f(t)$.

The basic algorithm of the rejection method is shown below.

1. Generate t from $g(t)$ and u from $Uniform(0,1)$.
2. IF $u < f(t)/f^*(t)$,
      accept t as a realization of $f(t)$;
   else
      reject the value of t and repeat Step 1.

Instead of using a majorizing function, Leemis and Schmeiser [15] present a thinning algorithm by using a majorizing failure rate function for the rejection method which is applied in this paper and is shown as follows.

1. Find a majorizing hazard function $\lambda^*(t)$ such that $\lambda^*(t) \geq \lambda(t)$ for all $t > 0$.
2. Let $t = 0$.
3. Generate y from $\lambda^*(y)$; $t \leftarrow t + y$.
4. Generate u from $Uniform(0,1)$.
5. IF $u < \lambda(t)/\lambda^*(t)$,
      accept t as a realization of $\lambda(t)$;
   else
      reject the value of t and repeat Step 3.
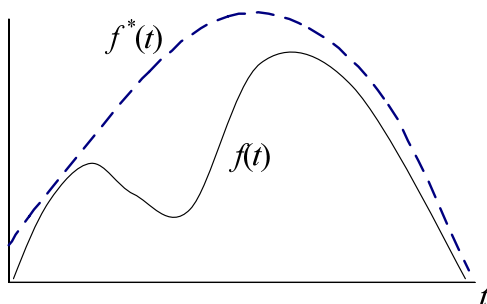
Figure 2.   The illustration of the rejection method by using the majorizing function

B. The Generation of Time-Between-Failure RVs

(1) The Inverse Transformation Method

In this paper, we apply the modified inverse transformation method presented by Cheng and Liaw [16] to develop an algorithm for generating the TBF random variates from a PM model with age reduction effect. Since the minimal repair is assumed for each failure in this age-reduction PM model, for the ith PM cycle, the CDF of the jth TBF ($x_{i,j}$) given that the $j\text{-}1^{st}$ failure occurred at $t_{i,j-1}$ can be obtained as

$$F(x_{i,j}|t_{i,j-1}) = 1 - R(x_{i,j}|t_{i,j-1}) = 1 - \exp\left\{-\int_{t_{i,j-1}}^{t_{i,j-1}+x_{i,j}} \lambda_i(t')dt'\right\} \quad (5)$$

$$\text{for } i = 0,1,...,N; \; j = 1,2,...,k_i.$$

Cheng and Liaw [16] originally assume that the last failure in the ith PM cycle is irrelative to the first failure in the $i\text{+}1^{st}$ PM cycle. However, this assumption may not be reasonable because the failure occurrence of the system follows the non-homogenous Poisson process (NHPP) and the PM is imperfect (i.e., the PM will not renew the system to zero failure rate). Therefore, The algorithm developed for the modified inverse transformation method assumes that the first failure in the $i\text{+}1^{st}$ PM cycle is affected by the last failure in the ith PM cycle. It turns out that the CDF of the first TBF in the $i\text{+}1^{st}$ PM cycle given that the k_i^{th} failure occurred at t_{i,k_i} can be obtained as

$$F(x_{i+1,1}|t_{i,k_i}) = 1 - R(x_{i+1,1}|t_{i,k_i})$$
$$= 1 - \exp\left\{-\left[\int_{t_{i,k_i}}^{(i+1)T} \lambda_i(t')dt' + \int_{(i+1)T}^{t_{i+1,1}} \lambda_{i+1}(t')dt'\right]\right\}. \quad (6)$$

For a Weibull failure distribution, providing that $u_{i,j} \sim Uniform(0,1)$, based on (5) and (6), the TBF $x_{i,j}$ can be generated as follows.

(1) For $i = 0$; $j = 1, 2, ..., k_1$ and $i = 1, 2, ..., N$; $j = 2, 3, ..., k_i$:

$$x_{i,j} = \left\{-\theta^\beta \ln(1-u_{ij}) + \left[t_{i,j-1} - i\gamma\right]^\beta\right\}^{1/\beta} - t_{i,j-1} + i\gamma, \quad (7)$$

where $t_{0,0} = 0$, $t_{i,j} = t_{i,j-1} + x_{i,j}$;

(2) for $i = 1, 2, ..., N$; $j = 1$:

$$x_{i,1} = \left\{ \begin{array}{l} \left[i(T-\gamma)\right]^\beta + \left[t_{i-1,k_{i-1}} - (i-1)\gamma\right]^\beta \\ -\left[iT-(i-1)\gamma\right]^\beta - \theta^\beta \ln(1-u_{i-1,1}) \end{array} \right\}^{1/\beta} - t_{i-1,k_{i-1}} + i\gamma, (8)$$

where $t_{i,1} = t_{i-1,k_{i-1}} + x_{i,1}$. The inverse transformation method of generating the TBF random variates for the age-reduction PM model is illustrated in Figure 3. The algorithm for the modified inverse transformation method is developed as follows.

1. Specify the values of the following parameters: β, θ, γ, N, L.
2. Let $i = 0$, $t_{0,0}=0$.
3. Let $j = 1$.
4. Generate $u_{i,j}$ from $Uniform(0,1)$.

Figure 3.   The illustration of the inverse transformation method of generating TBF random variates for the age-reduction PM model

5. Obtain the value of $x_{i,j}$ using (7);
  let $t_{ij} = t_{i,j-1} + x_{i,j}$.
6. If $t_{i,j} < iT$, let $j = j + 1$ and go back to Step 4, else go to Step 7.
7. If $t_{i,j} < L$,
   obtain the value of $x_{i+1,1}$ using (8);
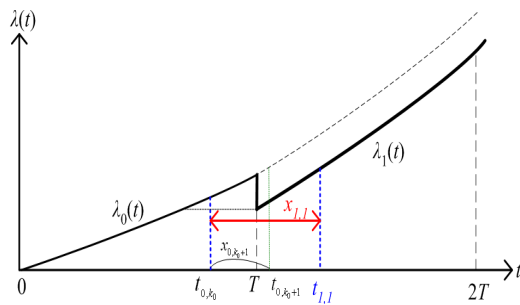   let $t_{i+1,1} = t_{i,k_i} + x_{i+1,1}$;
   let $i = i + 1$ and $j = 2$;
   go back to Step 4,
  else stop.

(2) The Rejection Method

In this paper, a thinning algorithm of the rejection method provided by Leemis and Schmeiser [18] is applied for generating the RVs from the non-homogeneous Poisson processes. A majorizing function $\lambda^*(t)$ must be found which bounds the failure rate function $\lambda(t)$. It can be seen from Figure 1 that the original failure rate function $\lambda(t)$ satisfies the condition $\lambda(t) \geq \lambda_i(t)$ for any $t \geq 0$ where $\lambda_i(t)$ is the failure rate function of the ith PM cycle as defined in (2). Thus, $\lambda(t)$ can be defined as the majorizing function, i.e., $\lambda^*(t) = \lambda(t)$.

When using the rejection method, two random numbers, u_1 and u_2, from $Uniform(0,1)$ are required for generating each RV from the PM model with age reduction effect. We use u_1 to generate random variates from the majorizing function $\lambda(t)$ by using the inverse transformation method. Basic on the concept of the inverse transformation method stated in previous and the equation of the CDF for the TBF which is shown in (5), we have

$$u_1 = F(x_m | t_{m-1}) = 1 - \exp\left\{-\int_{t_{m-1}}^{t_{m-1}+x_m} \lambda(\tau)d\tau\right\} \quad \text{for } m = 1,2,\ldots \quad (9)$$

where m is the number of failures generated, $F(x_m | t_{m-1})$ represents the conditional CDF at time $t_{m-1} + x_m$ given that the system is surviving after the minimal repair at time t_{m-1}.

When a system has Weibull failure distribution, based on (1) and (9), we can generate RVs from the majorizing function $\lambda(t)$ by the following equation.

$$x_m = \left[-\theta^\beta \ln(1-u_1) + (t_{m-1})^\beta\right]^{1/\beta} - t_{m-1}. \quad (10)$$

Then, x_m is accepted if $u_2 < \lambda_i(t)/\lambda(t)$. The algorithm for generating the time-to-failure random variates from the age-reduction PM model is shown as follows.

1. Specify the values of the following parameters: β, θ, γ, N, L.
2. Let $t_{0,0}=0$, $t_0=0$.
3. Let $m = 0$, $i = 0$, $j = 1$.
4. Generate random number u_1.
5. Obtain the value of x_m according to (10); let $t_m = t_{m-1}+x_m$.
6. If $t_m < iT$, go to Step 7 else go to Step 10.
7. Generate random number u_2.
8. Calculate $\lambda(t_m)$ and $\lambda_i(t_m)=\lambda(t_m-i\gamma)$.
9. If $u_2 \leq \lambda_i(t_m)/\lambda(t_m)$,
    let $t_{ij} = t_m$; $j = j+1$; $m = m+1$;
    go back to Step 4
   else
    $m = m+1$; go back to Step 4.
10. If $t_m < L$,
    let $i = i + 1$ and $j = 1$; go back to Step 7
   else stop.

C. The Comparision of theProposed RV Generation Methods Using Experiment Examples

To compare the accuracy of the two RV generation methods proposed in this paper, we assume the finite life time period (L) be 6 time units, the number of PM (N) be 5, PM interval (T) be 1. The PM restoration effect is set as $\gamma = 0.8$. The Weibull failure distribution with scale parameter $\theta = 0.4$ is assumed. Then, two experiments with $\beta = 2.2$ and 3.2 are constructed for each RV generating method. There are 30 runs for each experiment.

The accuracy of each proposed RV generation method

TABLE I.
THE COMPARISON OF THE PROPOSED RV GENERATION METHODS

(a) For $\beta = 2.2$

i	$E(k_i)$	$\bar{k}_{Ii}$	$\bar{k}_{Ri}$	p-value $(\bar{k}_{Ii}=\bar{k}_{Ri})$	MAD_{Ii}	MAD_{Ri}	p-value $(\bar{k}_{Ii} = E(k_i))$	p-value $(\bar{k}_{Ri} = E(k_i))$
0	7.5070	7.0333	7.0333	1	1.9676	1.9676	0.8457	0.8457
1	10.9939	11.6333	11.4333	0.8341	2.6350	3.2341	0.8483	0.9115
2	14.7380	15.000	16.2667	0.2580	3.3159	3.4731	0.9526	0.7235
3	18.6720	18.333	18.8333	0.6728	4.0656	3.5437	0.9448	0.9683
4	22.7621	22.8333	22.4667	0.7897	4.4175	4.1016	0.9892	0.9542
5	26.9862	29.2667	27.2333	0.1163	4.8740	3.7676	0.6931	0.9548
all	101.659	104.1000	103.2667	0.7490	7.3000	8.7561	0.8005	0.8782

(b) For $\beta = 3.2$

i	$E(k_i)$	$\bar{k}_{Ii}$	$\bar{k}_{Ri}$	p-value $(\bar{k}_{Ii}=\bar{k}_{Ri})$	MAD_{Ii}	MAD_{Ri}	p-value $(\bar{k}_{Ii} = E(k_i))$	p-value $(\bar{k}_{Ri} = E(k_i))$
0	18.7676	18.8333	18.8333	1	3.66432	3.6643	0.9883	0.9883
1	33.5259	33.7333	34.9333	0.4956	5.2632	5.6948	0.9744	0.8405
2	54.0830	56.000	54.13333	0.3105	5.4000	5.2111	0.7982	0.9939
3	80.7884	81.06667	78.43333	0.2623	7.2526	7.7910	0.9747	0.7993
4	113.917	113.5667	112.2333	0.6351	10.0389	7.4889	0.9767	0.8557
5	153.699	152.1333	154.3667	0.5484	10.8932	11.8068	0.9058	0.9645
all	454.7809	455.3333	452.9333	0.6708	17.9041	15.2375	0.9804	0.9277

is evaluated. In Tables I, we present the following values in the ith PM cycle for $\beta = 2.2$ and 3.2: the expected number of failures ($E(k_i)$), the average number of failures generated by the inverse transformation method and the rejection method ($\bar{k}_{Ii}$ and $\bar{k}_{Ri}$, respectively), the p-value obtained from the hypothesis test of H_o: $E(\bar{k}_{Ii}) = E(\bar{k}_{Ri})$, the mean absolute deviation of the generated number of failures from the expected number of failures for the inverse transformation method and the rejection method (MAD_{Ii} and MAD_{Ri}, respectively), and the p-values obtained from the hypothesis tests of H_o: $\bar{k}_{Ii} = E(k_i)$ and H_o: $\bar{k}_{Ri} = E(k_i)$.

From Table I, we can see that the two RV generation methods have the same results in the initial PM cycle ($i = 0$) because they produce the same failure observations before any PM action is performed. It can be found from the p-values of H_o: $E(\bar{k}_{Ii}) = E(\bar{k}_{Ri})$ for all the PM cycles that there are no significances between the results of the two RV generation methods for both $\beta = 2.2$ and 3.2. It can also be noted from the values of MAD and the p-values of H_o: $\bar{k}_{Ii} = E(k_i)$ and H_o: $\bar{k}_{Ri} = E(k_i)$ that the inverse transformation method and the rejection method have similarly high accuracy in generating the TBF random variates for the age-reduction PM model. Therefore, we apply the rejection method to find the near-optimal PM policies since the rejection method uses a majorizing failure rate function which is easier and simpler for generating the TBF random variates from the failure distributions with complicated formula.

D. Procedure of Finding the Near-Optimal PM Policies

Once the algorithm for generating the TBF random variates is developed, we can find the near-optimal PM policies by the following procedure.

1. Specify the failure distribution, the PM model, and the type of restoration.
2. Set the values for the required parameters, such as the parameters of the failure distribution ($f(t)$), repair cost (C_{mr}), PM cost (C_{pm}), life time (L), number of simulation runs, maximum number of PM to be simulated (N), .and the restoration effect (γ) if it has to be predetermined.
3. Develop the failure rate function, the CDF of the TBF, and the cost function (TC).
4. Apply the inverse transformation method or the rejection method to generate the TBF random variates for each simulation run.
5. Calculate the total cost for each simulation run.
6. Establish the tables of simulation results (similar to Tables II, III or IV, V).
7. Find the near-optimal policies from the tables established in Step 6.

IV. EXAMPLES AND DISCUSSION

In the examples, we assume the finite life time period (L) be 6 time units. The Weibull failure distribution with shape parameter $\beta = 3.2$ and scale parameter $\theta = 0.4$. The

minimal repair cost is set as 3.1036 per failure. The cost of each PM is assumed as function of the PM restoration effect, which is $C_{pm} = a+b\gamma = 5+100\gamma$. The near-optimal solutions are obtained by Monte Carlo simulation method with 30 runs. The theoretical optimal solution is also calculated using the algorithm provided by Yeh and Chen [11].

A. Example 1

In Example 1, for simplicity, the restoration effect γ for each N is predetermined using the theoretical method as shown in Table II. The 30-run simulation results for $N = 1$ to 6 using the rejection method are presented in Table III. The smallest (best) total maintenance cost (TC) of each run is highlighted with shadow background. It can be seen from Table III that, for each N, the average value of TC from the 30-run simulation is very close to the value obtained by using the theoretical method. Both methods (simulation and theoretical model) provides the same optimal policy of having the optimal number of PM $N^*=3$ and the optimal restoration effect $\gamma^*=0.4781$. It has

TABLE II.
THE VALUE OF γ CORRESPONDING TO EACH N FOR EXAMPLE 1

N	1	2	3	4	5	6
γ	1	0.6667	0.4781	0.3655	0.2957	0.2483

TABLE III.
THE RESULTS OF THE 30 SIMULATION RUNS FOR EXAMPLE 1

Run# N γ	1 — 1	2 — 0.6667	3 — 0.4781	4 — 0.3655	5 — 0.2957	6 — 0.2483
1	216.730	189.894	**180.155**	194.132	194.575	210.016
2	229.144	196.101	192.570	**187.925**	210.093	194.498
3	250.869	196.101	**189.466**	197.236	213.197	206.912
4	250.869	186.790	**180.155**	200.340	188.368	213.120
5	204.315	199.205	204.984	**187.925**	219.404	197.602
6	263.284	**189.894**	201.880	194.132	197.679	197.602
7	213.626	**165.065**	183.259	197.236	185.264	203.809
8	232.248	214.723	**192.570**	194.132	206.990	194.498
9	241.558	**177.480**	183.259	200.340	191.472	206.912
10	216.730	189.894	**180.155**	194.132	188.368	206.912
11	219.833	211.619	**204.984**	209.650	213.197	216.223
12	222.937	189.894	189.466	**181.718**	197.679	197.602
13	247.766	**177.480**	180.155	209.650	197.679	216.223
14	247.766	196.101	**180.155**	206.547	197.679	203.809
15	198.108	196.101	189.466	**181.718**	200.782	194.498
16	216.730	192.998	**183.259**	200.340	185.264	203.809
17	226.040	205.412	**189.466**	191.029	194.575	206.912
18	195.004	202.308	211.191	**191.029**	210.093	191.394
19	216.730	**183.687**	186.362	191.029	200.782	213.120
20	226.040	199.205	**183.259**	203.443	200.782	206.912
21	232.248	186.790	189.466	**184.822**	194.575	203.809
22	204.315	**196.101**	198.777	197.236	197.679	200.705
23	207.419	186.790	**180.155**	206.547	188.368	206.912
24	210.522	208.516	195.673	**187.925**	206.990	197.602
25	219.833	205.412	**195.673**	203.443	197.679	213.120
26	257.076	192.998	**173.948**	215.858	191.472	219.327
27	216.730	211.619	195.673	**172.407**	210.093	188.291
28	216.730	**189.894**	201.880	200.340	197.679	216.223
29	210.522	**168.169**	180.155	206.547	191.472	216.223
30	247.766	**186.790**	189.466	187.925	197.679	200.705
Average	225.316	193.101	**189.569**	195.891	198.92	204.843
Theoretical	221.495	191.076	**189.728**	192.850	197.222	202.051

demonstrated that the experimental results obtained by simulation method are consistent with those obtained by the theoretical model when large sample runs are generated.

It should be noted that the best solution of N, γ, and TC (marked with shadow) resulted from each simulation run are different from those obtained by the theoretical model. It is because the optimal solution of the theoretical model is obtained by taking the expectation result over the infinite time interval or over a large number of identical systems in a finite time interval. However, the simulation method can provide different policies for a single system having a finite life time. This is more close to the real situation where an organization will not possess a large number of identical systems nor will purchase an identical system in every replacement cycle for suiting the assumption of infinite time span.

It can be seen from Table III that the best solutions of each simulation run (marked with shadow) can be categorized into three near-optimal policies: (N=2, γ=0.6667), (N=3, γ=0.4781), and (N=4, γ=0.3655). Table IV lists the simulation runs in each category and presents the average, the smallest, and the largest values of the minimal TC for each category of the near-optimal policy.

Among these best solutions, the average of the minimal TC (184.1143) is smaller than the theoretical minimal TC (189.7280). The results demonstrate that the theoretical PM model over a finite time span might not be suitable for the problem of considering only a single system in a finite time span.

Therefore, in practical, when considering a single system to be preventively maintained in a finite time period, especially for short time period, more than one single near-optimal policy is suggested. In this example, either (N=2, γ=0.6667) or (N=3, γ=0.4781) or (N=4, γ=0.3655) may be chosen as the best (near-optimal) PM policy. Further examining the simulation results shown in Table IV, we can see that the minimal TCs marked with shadow are smaller than the theoretical optimal TC (189.7280) in which 6 out of 9 (67%) for Policy 1, 10 out of 13 (77%) for Policy 2, 7 out of 8 (88%) for Policy 3, and 23 out of 30 (77%) for overall policies are better than the theoretical optimal policy. Therefore, for a finite-time-span PM problem, these results show strong evidence that the simulation method not only flexibly provides near-optimal policies but also gives better-than-the theoretical optimal policies with high confidence.

B. Example 2

In the second example, we assume that the PM restoration value (γ) for each N is a decision variable and has to be determined by the simulation method. Therefore, different values of γ over the range of (0, 1) are used for each N where N is ranging from 1 to 6 in the simulation experiments. Only the experiments of $\gamma = 0.2T$, $0.5T$, $0.8T$, $0.985T$, and γ^* (the theoretical optimal value) are shown in this example. All other parameters are given with the same values as in Example 1. A 30-run simulation is performed for each combination of (N, γ) to find the smallest TC with the corresponding best value of

TABLE IV.
THE NEAR-OPTIMAL POLICIES FOR EXAMPLE 1

Policy 1 (N^*=2, γ^*=0.6667)		Policy 2 (N^*=3, γ^*=0.4781)		Policy 3 (N^*=4, γ^*=0.3655)	
Run#	Min. TC	Run#	Min. TC	Run#	Min. TC
6	189.894	1	180.155	2	187.925
7	165.065	3	189.466	5	187.925
9	177.480	4	180.155	12	181.718
13	177.480	8	192.570	15	181.718
19	183.687	10	180.155	18	191.029
22	196.101	11	204.984	21	184.822
28	189.894	14	180.155	24	187.925
29	168.169	16	183.259	27	172.407
30	186.790	17	189.466		
		20	183.259		
		23	180.155		
		25	195.673		
		26	173.948		
No of Runs	9		13		8
Runs Better than Theoretical	6 (67%)		10 (77%)		7 (88%)
Avg.	181.6177		185.6462		184.4337
Min.	165.0652		173.9480		172.4072
Max.	196.1012		204.9840		191.0288
Overall average of min. TC: 184.1143					
The Rate of better-than-theoretical solutions: 77%					

N and γ. The results are presented in Table V. Again, among these best solutions, the average of the minimal TC (182.5808) is smaller than the theoretical minimal TC (189.7280). This result shows that the best solution obtained from Monte Carlo simulation is better than the optimal theoretical solution.

Table VI lists the simulation runs in each category of the near-optimal policy and presents the average, the smallest, and the largest minimal TC of each near-optimal policy for Example 2. Likewise, the simulation method provides three near-optimal policies (with N = 2, 3, and 4). Since γ is a decision variable, each near-optimal policy has more than one value for γ. The values of γ are in the range of (0.6375, 0.6667) in Policy 1 (N = 2); it falls in the range of (0.400, 0.4925) in Policy 2 (N = 3); in Policy 3 (N =), γ has values in the range of (0.3655, 0.3940). From Table VI, Similarly, we can see that the minimal TCs marked with shadow are smaller than the theoretical optimal TC (189.7280) where 6 out of 8 (75%) for Policy 1, 10 out of 11 (91%) for Policy 2, 9 out of 11 (82%) for Policy 3, and 25 out of 30 (83%) for overall policies are better than the theoretical optimal policy.

Note that the value of γ specified for each policy of Example 1 (as shown in Table III) is covered in the range of the corresponding policy of Example 2. Moreover, it can also be found that although Examples 1 and 2 have similar results, the near-optimal policy of Example 2 has better solution (smaller average TC, narrower range of TC, and higher rate of better-than-theoretical solutions) than the corresponding policy in Example 1. it implies that the simulation method can obtain better solutions if γ is a decision variable than it is predetermined a value.

TABLE V.
THE RESULTS OF THE 30 SIMULATION RUNS FOR EXAMPLE 2

Run #	N	γ	TC	Run #	N	γ	TC
1	3	0.4781	180.1552	16	3	0.4925	181.3716
2	4	0.3655	187.9252	17	3	0.4925	187.5788
3	3	0.4925	187.5788	18	4	0.3655	191.0288
4	4	0.3825	176.1036	19	2	0.6567	181.6868
5	4	0.3655	187.9252	20	3	0.4781	183.2588
6	2	0.6667	189.8940	21	4	0.3655	184.8216
7	2	0.6667	165.0652	22	4	0.3940	190.0144
8	3	0.4000	187.7612	23	3	0.4000	175.3468
9	2	0.6567	175.4796	24	4	0.3655	187.9252
10	3	0.4000	178.4504	25	4	0.3825	191.6216
11	3	0.4925	196.8896	26	3	0.4781	173.9480
12	4	0.3655	181.7180	27	4	0.3655	172.4072
13	2	0.6667	177.4796	28	2	0.6667	189.8940
14	3	0.4781	180.1552	29	2	0.6667	168.1688
15	4	0.3655	181.7180	30	2	0.6375	184.0540
Average				3.1		0.4816	182.5808
Theoretical				3		0.4781	189.7280

TABLE VI.
THE NEAR-OPTIMAL POLICIES FOR EXAMPLE 2

Policy 1: $N=2$			Policy 2: $N=3$			Policy 3: $N=4$		
Run #	γ	TC	Run #	γ	TC	Run #	γ	TC
6	0.6667	189.8940	1	0.4781	180.1552	2	0.3655	187.9252
7	0.6667	165.0652	3	0.4925	187.5788	4	0.3825	176.1036
9	0.6567	175.4796	8	0.4000	187.7612	5	0.3655	187.9252
13	0.6667	177.4796	10	0.4000	178.4504	12	0.3655	181.718
19	0.6567	181.6868	11	0.4925	196.8896	15	0.3655	181.718
28	0.6667	189.8940	14	0.4781	180.1552	18	0.3655	191.0288
29	0.6667	168.1688	16	0.4925	181.3716	21	0.3655	184.8216
30	0.6375	184.0540	17	0.4925	187.5788	22	0.394	190.0144
			20	0.4781	183.2588	24	0.3655	187.9252
			23	0.4000	175.3468	25	0.3825	191.6216
			26	0.4781	173.9480	27	0.3655	172.4072
No of Runs	8			11			11	
Runs Better than Theoretical	6 (75%)			10 (91%)			9 (82%)	
Avg	0.6606	178.9653		0.4620	182.9540		0.3712	184.8372
Min	0.6375	165.0652		0.4000	173.9480		0.3655	172.4072
Max	0.6667	189.8940		0.4925	196.8896		0.3940	191.6216
Overall average of min. TC: 182.581								
The Rate of better-than-theoretical solutions: 83%								

V. CONCLUSIONS

In this paper, we apply the inverse transformation method and the rejection method to develop the algorithms of generating the time-between-failure random variates for the age-reduction PM model to find the near-optimal PM policies.

We conclude from the results of this paper that, for the age-reduction PM model in a finite time span, more than one near-optimal policy can be obtained by using Monte Carlo simulation method. Each of the near-optimal solution can be the best PM policy for any single system having a finite life time. The simulation results have demonstrated that the theoretical PM model might not be suitable for a single system in a finite time span. It can be

further concluded from this research that (1) the PM policies obtained from the proposed simulation methods can provide even better solution for an organization such as a car rental company who possesses many identical vehicles; (2) When considering the PM problem of a single system, the decision is more flexible because any of the near-optimal policies can work well.

The proposed simulation methods can be extended to solve more complicated real world situation, such as considering the random shocks in a PM model which is very difficult to be solved by the theoretical model.

REFERENCES

[1] H. Pham and H. Wang, "Invited review: imperfect maintenance", *European Journal of Operational Research*, Vol.94, pp425-438,1996.

[2] S. Martorell, A. Sanchez, and V. Serradell, "Age-dependent reliability model considering effects of maintenance and working conditions", *Reliability Engineering and System Safety*, Vol.64, pp.19-31, 1999.

[3] S.-H. Sheu and C.-C. Chang, "An Extended Periodic Imperfect Preventive Maintenance Model With Age-Dependent Failure Type", *IEEE Transactions on Reliability*, Vol.58, No. 2, pp397 – 405, 2009.

[4] M. Bartholomew-Biggs, M. J. Zuo, and X. Li, "Modeling and optimizing sequential imperfect preventive maintenance", *Reliability Engineering and System Safety*, Vol.94, pp53– 62, 2009.

[5] D. N. P. Murthy and D. G. Nguyen, "Optimal Age-Policy with Imperfect Preventive Maintenance," *IEEE Transactions on Reliability*, Vol.R-30, No.1, pp.80-81, 1981.

[6] T. Nakagawa, "Periodic and Sequential Preventive Maintenance Policies", *Journal of Applied Probability* R-23/2, pp536-542, 1986.

[7] D.H. Park, G.M. Jung, and J.K. Yum, "Cost Minimization for Periodic Maintenance Policy of a System Subject to Slow Degradation", *Reliability Engineering and System Safety*, Vol.68, pp105-112, 2000.

[8] C.-Y. Cheng and M. Chen, "The periodic maintenance policy for a Weibull life-time system with degradation rate reduction under reliability limit," *Asia-Pacific Journal of Operational Research*, Vol.25, Issue: 6, pp793-805, 2008.

[9] T. Nakagawa and S. Mizutani, "A summary of maintenance policies for a finite interval", *Reliability Engineering and System Safety*, Vol.94: pp89-96, 2009.

[10] J. Pongpech and D. N. P. Murthy, "Optimal Periodic Preventive Maintenance Policy for Leased Equipment," *Reliability Engineering & System Safety*, Vol.91, pp.772-777, 2006.

[11] R. H. Yeh and C. K. Chen, "Periodical Preventive-Maintenance Contract for a Leased Facility with Weibull Life-Time," *Quality & Quantity*, Vol.40, pp.303-313, 2006.

[12] C.-Y. Cheng and H.-H. Liu, "The Finite-Time-Period Preventive Maintenance Policies with Failure Rate Reduction under A Warranty Consideration," *Journal of the Chinese Institute of Industrial Engineers*, Vol.27, No.2, pp. 81-89, 2010.

[13] C.-Y. Cheng, R. Guo, and M.-L. Liu, 2009, "Random Variates Generating Methods of Time-Between-Failures for the Repairable Systems under Age-Reduction Preventive Maintenance," *Proceedings of Sixth International Conference on Informatics in Control, Automation and Robotics (ICINCO2009)*, July 2-5, 2009, Milan, Italy, pp325-330.

[14] D. F. Percy, K. A. H. Kobbacy, "Determining economical maintenance intervals," International Journal of Production Economics, Vol.67, pp87-94, 2000.

[15] L. Leemis and B. Schmeiser, "Random Variate Generation for Monte Carlo Experiments," *IEEE Transactions on Reliability*, Vol.R-34, No. 1, pp81-85, 1985.

[16] C.-Y. Cheng and C.-F. Liaw, "Statistical estimation on imperfectly maintained system," *European Safety & Reliability Conference 2005 (ESREL 2005)*, Jun. 27-30, 2005, Tri-City, Poland, pp 351-356, 2005.

[17] S. M. Ross, *Simulation*, Academic Press, San Diego, 1997.

[18] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, (second edition). New York: Springer-Verlag, 2004.

[19] C.-Y. Cheng, C.-F.Liaw, and M.Wang,. "Periodic preventive maintenance models for deteriorating systems with considering failure limit," *4th International Conference on Mathematical Methods in Reliability—Methodology and Practices*, Santa Fe, New Mexico, June, 2004.

Chun-Yuan Cheng received her M.S. and Ph.D. degrees in industrial engineering from Auburn University, Alabama, USA. She is currently an associate professor at Chaoyang University of Technology, Taiwan. Her research interests include modeling of reliability and maintainability, simulation, and quality engineering.

Solving DOPF in VSWGs Integrated Power System Using Improved Evolutionary Programming

Gonggui Chen[1,2]

[1]College of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan 430074, Hubei Province, China

[2]Dept. of Electrical Engineering, Hubei University for Nationalities, Enshi 445000, Hubei Province, China
chengonggui@yahoo.cn

Hangtian Lei[1] Haibing Fang[2] Jian Xu[2]

[1]College of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan 430074, Hubei Province, China

[2]Dept. of Electrical Engineering, Hubei University for Nationalities, Enshi 445000, Hubei Province, China

Abstract—**Wind turbine can be divided into two categories: fixed speed wind generators (FSWGs) and variable speed wind generators (VSWGs) . VSWGs's bus can be dealt with as PV-bus or PQ-bus in power flow calculation because reactive power compensation can be performed. Dynamic optimal power flow (DOPF) in VSWGs integrated power system is a typical complex multi-constrained non-convex non-linear programming problem when considering the valve-point effect of conventional generators. In this paper, an improved evolutionary programming (IEP) is proposed to solve DOPF in VSWGs integrated power system. In the methodology, the well-known evolutionary programming (EP) is used as a basic level search, which can give a good direction to the optimal global region. Then, a local search (LS) procedure is adopted as a fine tuning to determine the optimal solution. The modified IEEE 30-bus system is used to illustrate the effectiveness of the proposed method compared with those obtained from EP algorithm. In order to verify algorithm effectiveness in more complex power system, IEEE 39-bus system is used test system. It is shown that the proposed method is capable of yielding higher-quality solutions.**

Index Terms— **wind power generation, variable speed wind generators, dynamic optimal power flow, improved evolutionary programming**

I. INTRODUCTION

Wind energy is the world's fastest growing renewable energy source. With the increasing levels of wind generator penetration in modern power systems, one of major challenges in the present and coming years is the optimization control, such as optimal power flow including wind farms [1].

Wind turbine can be divided into two categories: fixed speed wind generators (FSWGs) and variable speed wind

generators (VSWGs) . FSWGs are still widely use in power system. The disadvantages of FSWGs are as follows.

When wind speed jump, huge wind force will pass through wind turbine blade, hit wind generator components including the main shaft, gearbox, engines, and so on. FSWGs mean that wind speed fluctuations are directly translated into electromechanical torque variations. This will bring in great mechanical stress and cause high fatigue damages on the components, and may result in swing oscillations between turbine and generator shaft. Also the periodical torque dips because of the tower shadow and shear effect are not damped by speed variations and result in higher flicker. Furthermore, the turbine speed cannot be adjusted with the wind speed to optimise the aerodynamic efficiency and wind energy utilization coefficient.

Compared with FSWGs, VSWGs turbine speed can be adjusted with wind speed. Mechanical stress is reduced, and gust energy can be absorbed by the means of inertia; wind energy utilization coefficient is improved, and reactive power compensation can be performed. So, VSWGs's bus can be dealt with as PV-bus or PQ-bus in power flow calculation.

In this paper, the problems of dynamic optimal power flow (DOPF) including VSWGs are researched. The expectation model of wind generators' active power outputs is adopted. DOPF is a typical complex multi-constrained non-convex non-linear programming problem in wind power integrated system when considering the valve-point effect of conventional generators. Both lambda-iterative and gradient technique methods in conventional approaches to the problems are calculus-based techniques and require a smooth and convex cost function and strict continuity of the search space.

In the field of global optimization, evolutionary programming (EP) was investigated and proved to be powerful in solving these problems in the last decades.

EP is a stochastic search technique with biological foundations. However, One disadvantage of EP in solving some of the multimodal optimization problems is its slow convergence to a good near-optimum[2, 3, 4].

In this paper, a new improved evolutionary programming (IEP) methodology is proposed for solving DOPF in VSWGs integrated power system; A simple EP [5] is applied as a basic level search, which can give a good direction to the optimal global region, and a local search (LS) procedure [6, 7] is used as a fine tuning to determine the optimal solution at the final. IEP methodology enhances the computational accuracy and accelerates convergence rate at the later period of the searching by adopting LS operator which is invoked if fitness evaluation improves.

The modified IEEE 30-bus system and IEEE 39-bus system are used to illustrate the effectiveness of the proposed method for solving the established DOPF model in VSWGs integrated power system compared with those obtained from EP algorithm. It is shown that the proposed method is capable of yielding higher-quality solutions.

II. DOPF MODEL IN VSWGS INTEGRATED POWER SYSTEM

Due to the random variation of the wind velocities and load demands, it is difficult to research the DOPF in the power system including wind farms. For simplifying this problem, the dividing-stage strategy is adopted in this paper. Wind power generated by wind turbines has intimate relationship with wind speed. Wind speed is converted into power through characteristic curve of a wind turbine. According to the wind velocity forecasting curves and the load forecasting curves in the planning horizon, the expectations of wind generators' power outputs and the load demands at dispatch interval can be calculated.

A. Constraints

Constraints include equality and inequality constraints. The equation constraint is the power flow formulation constraint while inequality constraints including generator power output, ramp rate and bus voltage are as in(1) -(3) . The constraints of real power generation limit and the ramp rate are taken into account as in (1) .

$$\begin{cases} \max\left(P_{i,\min}, P_i^{t-1} - D_{Ri}\Delta T\right) \le P_i^t \le \min\left(P_{i,\max}, P_i^{t-1} + U_{Ri}\Delta T\right) \\ i \in N_{gen} \end{cases}$$
(1)

$$Q_{Gi,\min} \le Q_{Gi}^t \le Q_{Gi,\max}, i \in N_{gen} .$$
(2)

$$V_{i,\min} \le V_i^t \le V_{i,\max}, i \in N .$$
(3)

where $P_{i,min}$ and $P_{i,max}$ are the maximum and minimum limits of the power generation of unit i, P_i^t is the real power output of unit i at the tth interval, $P_i^{t-1}$is the real power output of unit i at the t-1th interval; U_{Ri} is the up-ramp limit of the ith generator (in units of MW/time-period) ,and D_{Ri} is the down-ramp limit of the ith

generator (in units of MW/time-period) ; $\triangle T$ is time interval, N_{gen} is the number of conventional generating units, and N is the number of system buses (excluding slack bus) ; V_i^t is the voltage magnitude output of bus i at the tth interval; Q_{Gi}^t is the reactive power output of conventional generating unit i at the tth interval; max is the maximum value of the variable, min is the minimum value of the variable.

After calculating the power flow, the state variables, power loss and real power output of the slack bus generator corresponding to the current control variables are available. The real power output of the slack bus generator will be set to the limit if it violates the limit. After handling overlimit of the real power output of the slack bus generator, the system power balance constraints as in(4) must meet, otherwise adding (4) as penalty terms to the objective function to form a generalized objective function. Details of the generalized objective function used in this paper are given in section C.

$$\Delta P^t = \sum_{i=1}^{N_{gen-1}} P_i^t + P_{sl}^t + P_{w,av}^t - P_{ls}^t - P_{ld}^t = 0 .$$
(4)

where ΔP^t is the unbalance of the real power at the tth interval,N_{gen}-1 represents the number of conventional generating units excluding the slack bus, P_{sl}^t is the real power output of the slack bus generator after handling its overlimit at the tth interval, P_{ls}^t is the total power loss at the tth interval, P_{ld}^t is the total load expectation at the tth interval, $P_{w,av}^t$ is the expectation of wind generators' real power outputs at the tth interval.

B. Objective Function

Due to the fact that wind generation does not consume the fuel, the utility must purchase all the energy produced by wind generating units. Consequently, the objective is to minimize the following total incremental fuel cost function F associated to N_{gen} dispatchable units for T intervals in the given time horizon, subject to the above-mentioned equality and inequality constraints.

$$\min F = \sum_{t=1}^{T} \sum_{i=1}^{N_{gen}} F\left(P_i^t\right) .$$
(5)

The inclusion of valve-point loading effects makes the modeling of the fuel cost function of the unit more practical. This increases the non-linearity and local optima in the solution space. Also the solution procedure can easily trap in the local optima in the vicinity of optimal value. The fuel cost function of the ith unit $F(P_i^t)$ with valve-point loadings are represented as follows

$$F\left(P_i^t\right) = a_i + b_i P_i^t + c_i P_i^{t2} + \left|e_i \sin(f_i(P_{i,\min} - P_i^t)\right| .$$
(6)

where a_i, b_i, and c_i are cost coefficients and e_i, f_i are constants from the valve-point effect of the ith generating unit.

C. Evaluation Function

We must define the evaluation function for evaluating the fitness of each individual in the population. In the

most of the nonlinear optimization problems, the constraints are considered by generalizing the objective function using penalty terms.

To sum up, the above problems are generalized as follows

$$
\min\Big\{ \sum_{t=1}^{T}\sum_{i=1}^{N_{gen}} F\big(P_i^t\big) + K_V \sum_{t=1}^{T}\sum_{i\in N_{PQ}} (V_i^t - V_i^{\lim})^2 + \cdots
$$
$$
+ K_Q \sum_{t=1}^{T}\sum_{i\in N_{gen}} (Q_{Gi}^t - Q_{Gi}^{\lim})^2 + K_D \sum_{t=1}^{T}\big(\Delta P^t\big)^2 \Big\} \qquad (7)
$$

where K_V, K_Q and K_D are variable overlimit penalty coefficients, V_i^t is the voltage magnitude of bus i at the tth interval (excluding the slack bus and PV bus) ; Q_{Gi}^t is the reactive power output of generator i at the tth interval; V_i^{lim} and Q_{Gi}^{lim} denote the violated upper or lower limits.

In this paper, K_V, K_Q are set to 1, 1 respectively. Because the unbalance of the real power ΔP is hard to meet, an adaptive penalty function to handle penalty coefficient K_D is adopted, $K_D = k\sqrt{k}\ \gamma\beta^\alpha$, where k is the algorithm's current iteration number; β is a relative violated value of the constraints, γ is a multi-stage assignment value, a is the power of the penalty value.

Meanwhile, several experiments have been done in order to obtain the penalty parameters. In this study, if $\beta\le1$ then $a=1$, otherwise $a=2$. Furthermore, if $\beta\le0.001$, then $\gamma=1$, else, if $\beta\le0.01$ then $\gamma=10$, else, if $\beta\le0.1$ then $\gamma=30$, else, if $\beta\le1$ then $\gamma=100$, otherwise $\gamma=300$.

III. IEP AND IMPLEMENTS

A. Evolutionary Programming

EP is a powerful global optimization technique, has proved itself effective to handle complex optimization problems [2,3,4].EP starts with a population of randomly generated candidate solutions and evolves towards the better solutions over a number of iterations. It uses probabilistic rules to explore the complex search space. Hence, it is more suitable to effectively handle complex optimization problems. The main stages of EP include initialization, mutation, and competition and selection. The generalized mapping procedure of the EP technique is as follows

1) Representation and initialization

For DOPF problem including VSWG, there are T dispatches by N_{gen-1} conventional generating units. An individual array of control variable arrays is

$$
\boldsymbol{P} = \begin{pmatrix}
P_1^1 & P_1^2 & \cdots & P_1^t & \cdots & P_1^T \\
P_2^1 & P_2^2 & \cdots & P_2^t & \cdots & P_2^T \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
P_{n-1}^1 & P_{n-1}^2 & \cdots & P_{n-1}^t & \cdots & P_{n-1}^T \\
P_n^1 & P_n^2 & \cdots & P_n^t & \cdots & P_n^T
\end{pmatrix}. \qquad (8)
$$

$$
\boldsymbol{P} = 1,2,\cdots g \ . \qquad (9)
$$

where $\boldsymbol{P}$ is individual vector, g is the number of population individuals, P_n^t is the real power output of nth

generating unit at the tth interval.

For the complete g population individuals, the candidate solution of each individual is randomly initialized within the feasible range in such a way that it should satisfy the constraint given by (1) .

2) Power flow and fitness calculation

Through the power flow calculation including wind farms, the state variables, power loss and real power output of the slack bus generator corresponding to the current control variables have been able to get. The real power output of the slack bus generator will be set to the limit if it violates the limit. After handling overlimit of the real power output of the slack bus generator, the system power balance constraints as in(4) must meet, otherwise adding (4) as penalty terms to the objective function to form a generalized objective function. In this paper, (7) is used as the fitness or evaluation function. This is a generalized fitness function used to evaluate the fitness of the candidate solution of each individual. Also, record the individual's position with the global best fitness as gBest, record the current position of each individual as its current pBest. Set the iteration count $k=0$.

3) Creation of offspring

The value of each decision variable in the individuals of the offspring population is obtained by perturbing the corresponding variable $p_{i,j}$ in the individuals of the parents population according to

$$
p'_{i,j}[k] = p_{i,j}[k] + \sigma_{i,j}[k]\cdot N_{i,j}(0,1)[k] \ . \qquad (10)
$$

where $\sigma_{i,j}[k]$ denotes the corresponding strategy parameter of $p_{i,j}[k]$ and $N_{i,j}(0,1)\,[k]$ is a Gaussian random value generated anew at each time of mutation. If the $p'_{i,j}[k]$ is outside the range, it is fixed to the boundaries.

$$
\sigma_{i,j}[k] = \beta[k]\frac{F_i[k]}{F_{\max}[k]}(p'_{j,\max} - p'_{j,\min}) \ . \qquad (11)
$$

where $F_i[k]$ denotes the fitness value of the ith individual in the kth generation; $F_{max}[k]$ and $F_{min}[k]$ denote the maximum and minimum fitness in the kth generation. where β is a scaling factor, which can be tuned during the process of search for optimum. The value of β used here was suggested by [3,4].

4) Selection and Competitio*n*

The q-tournament selection scheme is adopted in this paper. Each individual is assigned a score s_i according to

$$
s_i = \sum_{l=1}^{q} s_{i,l} \ . \qquad (12)
$$

$$
s_{il} = \begin{cases} 1 & , \ if \ \ F_i < F_l \\ 0 & , \ if \ \ F_i > F_l \end{cases} . \qquad (13)
$$

where F_i is the fitness of individual i, F_l is the fitness of an opponent individual randomly selected from the whole $2N$ individuals, and q is called the tournament size. The N individuals with higher scores are selected to form the parent population of the next generation. Tournament size q is set to $0.9N$ in this paper.

B. LS Subroutine

EP is a powerful global optimization technique, has proved itself effective to handle complex optimization problems. However, the standard EP convergence rate is very slow [2,3,4]. Consequently, the IEP of blending the standard EP with the following LS is proposed.

The LS procedure is outlined below [6,7].

The initial search point is taken as P_G^0, $P_G^0 = [x_{g1}^0 \ x_{g2}^0, \ldots, x_{Gd}^0]^T$ and the evaluation function value at P_G^0 is F_{gbest}^0. where D is the number of dimension.

Step 1) The initial LS range is selected around P_G^0 as follows

$$Y^{\min} = P'_{\min} + (P_G^0 - P'_{\min}) \times \beta \ . \qquad (14)$$

$$Y^{\max} = P'_{\max} - (P'_{\max} - P_G^0) \times \beta \ . \qquad (15)$$

$$R^0 = Y^{\max} - Y^{\min} = (P'_{\max} - P'_{\min})(1 - \beta) \ . \qquad (16)$$

where Y^{min} and Y^{max} are the lower and upper boundaries of the local search region; β is the local area parameter which is set to 0.4; $P'_{\max}$ and $P'_{\min}$ are the vectors of decision variables limits; and R^0 is the initial LS range. P_{Gbest}^0 (best search point at the beginning of LS) and P_{opt} (optimum search point) are set to P_G^0.

Step 2) The N_L LS points are randomly generated as follows

$$P_n^m = P_{Gbest}^{m-1} + R^{m-1} \times \mathbf{r}(D,1), \quad n = 1,2,\cdots,N_L \qquad . \qquad (17)$$

where r$(D,1)$ is a random number vector of length D, whose elements are randomly generated between -1 and 1 in this paper. If any LS point violates the limits, it is forced within the boundaries. N_L is the number of LS points which is set to 5.

Step 3) For each LS point, the evaluation function values are calculated. Then the minimum evaluation function among all is taken as F_{gbest}^m, and the corresponding P_G is taken as P_{Gbest}^m. The optimum values are updated as follows

IF $F_{gbest}^m < F_{gbest}^{m-1}$ then $F_{opt} = F_{gbest}^m$ and $P_{opt} = P_{Gbest}^m$ Otherwise $F_{opt} = F_{gbest}^{m-1}$ and $P_{opt} = P_{Gbest}^{m-1}$.

Step 4) The search range is reduced as

$$R^m = R^{m-1} \times (1 - \eta) \ . \qquad (18)$$

where η is the range reduction parameter which is set to 0.05.

Step 5) $m = m+1$, if maximum iteration for LS is not reached, the iteration count is incremented by one and the above procedure is repeated from *step 2*, otherwise, F_{opt} and P_{opt} are taken as the optimum results found by the LS algorithm.

C. IEP and Computational Procedure

The overall procedure of the proposed solution methodology can be summarized as follows

1) Get the initial data;

2) Initialize randomly the initial population in the feasible range and iteration count $k=0$; Evaluate the initial population and identify the $F_{min}(0)$ and the best initial individual;

3) $k=k+1$, creation of new population by mutation, competition and selection;

4) Evaluate the fitness score for each individual. Identify the $F_{min}(k)$ and the best individual of the current iteration k;

5) If $F_{min}(k) < F_{min}(k-1)$

6) Solve the DOPF in VSWGs integrated power system using the LS subroutine with the individual of $F_{min}(k)$ of the EP as starting point;

7) Replace $F_{min}(k)$ of the EP with the final solution obtained using the LS;

8) Repeat for generations until the terminal conditions $k_{\max}=150$ being satisfied.

So, it is beneficial not only for global optimization in the early evolution but also for the computational accuracy and convergence rate in the later period of the searching.

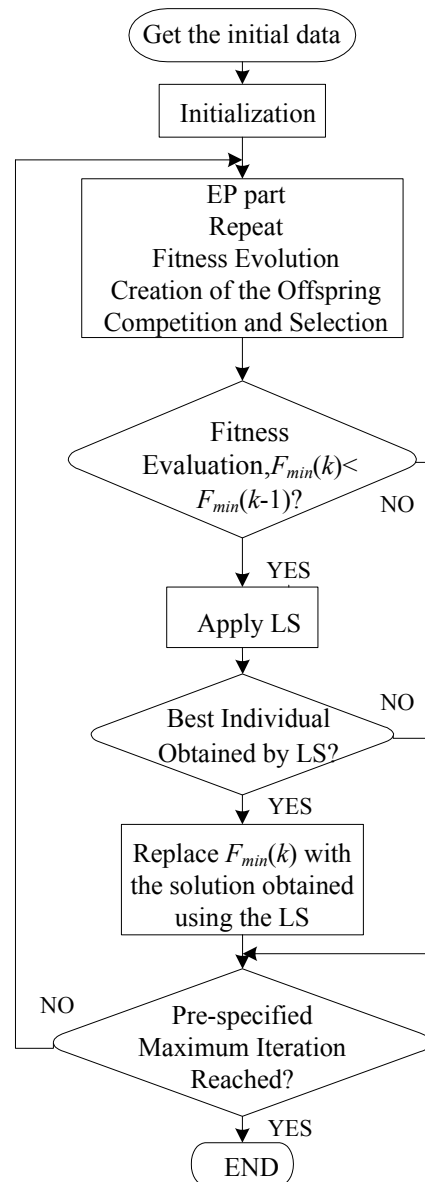The above strategies are clearly illustrated in Fig. 1.

Fig. 1. Flow chart for the proposed method

IV. NUMERICAL RESULTS

To verify the effectiveness and efficiency of the adopted IEP for DOPF problems including VSWGs, The modified IEEE 30-bus system (see Fig.2) and IEEE 39-bus system (see Fig.3) are used as the test systems. The procedure has been implemented in Matlab 7.0 programming language and numerical tests are carried on a Pentium 4 2.4G computer. The wind farm including 60 wind generators with the same type, the rating power of which reaches 36MW, are connected to the system at the bus 6 for the modified IEEE 30-bus system and at the bus 22 for IEEE 39-bus system respectively. For simplifying the analysis, the load size is considered invariable in the planning horizon. The planning horizon is divided into 9 intervals for the modified IEEE 30-bus system and 12 intervals for IEEE 39-bus system respectively, and every interval is 1hr. The wind generators' outputs are shown in Tab. I for the modified IEEE 30-bus system and Tab. II for IEEE 39-bus system respectively. The modified IEEE 30-bus system data are given in [8]. The modified IEEE 30-bus system parameters of the conventional generating units are shown in Tab. III and Tab. IV [8, 9]. IEEE 39-bus system data are given in[8]. The IEEE 39-bus system parameters of the conventional generating units are shown in Tab. V and Tab. VI [8, 10].
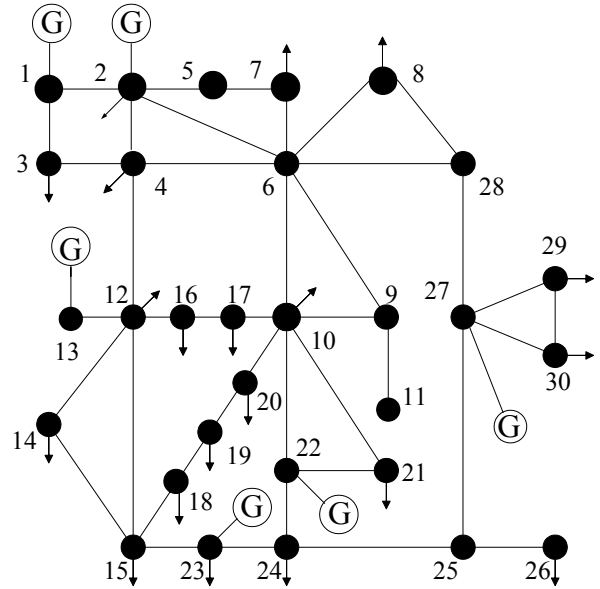
TABLE I

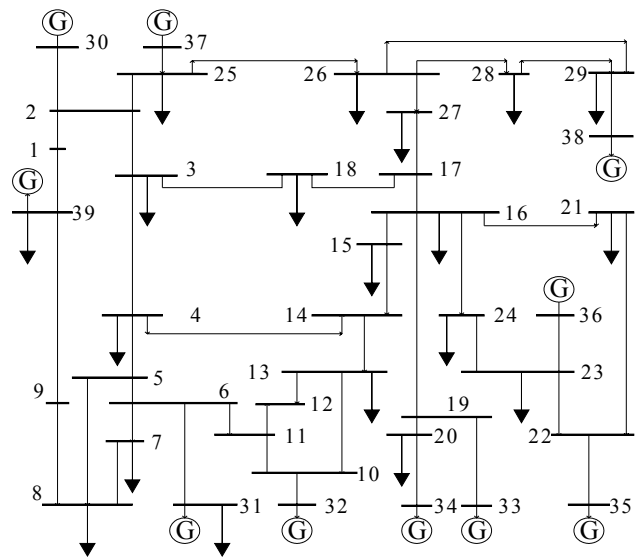THE WIND FARM DATA IN DIFFERENT PERIODS IN THE MODIFIED IEEE 30-BUS SYSTEM

Stage	1	2	3	4	5	6	7	8	9
$P^t_{w,av}$(MW)	0	4.5	9	13.5	18	22.5	27	31.5	36

TABLE II

THE WIND FARM DATA IN DIFFERENT PERIODS IN THE IEEE 39-BUS SYSTEM

Stage	1	2	3	4	5	6	7	8	9	10	11	12
$P^t_{w,av}$(MW)	9	12	15	18	36	36	36	36	0	10	14	19

TABLE III

THE PARAMETERS OF CONVENTIONAL GENERATING UNITS IN THE MODIFIED IEEE 30-BUS SYSTEM

Generator	a_i ($/h)	b_i ($/MWh)	c_i ($/MW^{2}h)	D_{Ri} (MW/h)	U_{Ri} (MW/h)	P_i^0 (MW)	e_i ($/h)	f_i (rad/MW)
G_1	0.0	2.00	0.0200	21.6	21.6	23.54	300	0.2
G_2	0.0	1.75	0.0175	18	18	60.97	200	0.22
G_{22}	0.0	1.00	0.0625	14.4	14.4	21.59	150	0.42
G_{27}	0.0	3.25	0.00834	10.8	10.8	26.91	100	0.3
G_{23}	0.0	3.00	0.0250	14.4	14.4	19.2	200	0.35
G_{13}	0.0	3.00	0.0250	18	18	37	200	0.35

TABLE IV

THE LIMITS OF CONVENTIONAL GENERATING UNITS IN THE MODIFIED IEEE 30-BUS SYSTEM

Generator	$Q_{i,max}$ (MVAr)	$Q_{i,min}$ (MVAr)	$V_{i,max}$ (p.u.)	$V_{i,min}$ (p.u.)	$P_{i,max}$ (MW)	$P_{i,min}$ (MW)
G_1	150	-20	1.05	0.95	80	0
G_2	60	-20	1.05	0.95	80	0
G_{22}	62.5	-15	1.05	0.95	50	0
G_{27}	48.7	-15	1.05	0.95	55	0
G_{23}	40	-10	1.05	0.95	30	0
G_{13}	44.7	-15	1.05	0.95	40	0

Fig.2. The modified IEEE 30-bus system

Fig.3. The 39-bus, 10-generator, IEEE system

TABLE V

THE PARAMETERS OF CONVENTIONAL GENERATING UNITS IN THE IEEE 39-BUS SYSTEM

Generator	a_i ($/h)	b_i ($/MWh)	c_i ($/MW^{2}h)	D_{Ri} (MW/h)	U_{Ri} (MW/h)	P_i^0 (MW)	e_i ($/h)	f_i (rad/MW)
G_{30}	0.2	0.3	0.01	80	80	250	450	0.041
G_{31}	0.2	0.3	0.01	80	80	572.9	600	0.036
G_{32}	0.2	0.3	0.01	80	80	650	320	0.028
G_{33}	0.2	0.3	0.01	50	50	632	260	0.025
G_{34}	0.2	0.3	0.01	50	50	508	280	0.063
G_{35}	0.2	0.3	0.01	50	50	650	310	0.048
G_{36}	0.2	0.3	0.01	30	30	560	300	0.086
G_{37}	0.2	0.3	0.01	30	30	540	340	0.082
G_{38}	0.2	0.3	0.006	30	30	830	270	0.098
G_{39}	0.2	0.3	0.006	30	30	1000	380	0.094

TABLE VI
THE PARAMETERS AND LIMITS OF CONVENTIONAL GENERATING UNITS
IN THE IEEE 39-BUS SYSTEM

Generator	$Q_{i,max}$ (MVAr)	$Q_{i,min}$ (MVAr)	$V_{i,max}$ (p.u.)	$V_{i,min}$ (p.u.)	$P_{i,max}$ (MW)	$P_{i,min}$ (MW)
G_{30}	9999	-9999	1.06	0.94	350	0
G_{31}	9999	-9999	1.06	0.94	1145.55	0
G_{32}	9999	-9999	1.06	0.94	750	0
G_{33}	9999	-9999	1.06	0.94	732	0
G_{34}	9999	-9999	1.06	0.94	608	0
G_{35}	9999	-9999	1.06	0.94	750	0
G_{36}	9999	-9999	1.06	0.94	660	0
G_{37}	9999	-9999	1.06	0.94	640	0
G_{38}	9999	-9999	1.06	0.94	930	0
G_{39}	9999	-9999	1.06	0.94	1100	0

To demonstrate the superiority of the proposed approach for DOPF problems, simulation results have been compared with the EP method. Owing to the randomness in intelligent algorithms, two algorithms are executed 20 times when applied to the test system.

For DOPF problem including VSWGs, Tab. VII and Tab. VIII list the best control variables found by IEP and EP algorithm for the modified IEEE 30-bus system respectively. In Tab. VII, it is clearly shown that, by using IEP, the total production cost savings of 28.3677$/h is obtained compared with EP algorithm. Hence, it is justified that IEP approach gives the exact

minimum dispatch solution. From Tab. XI, the best, worst and average cost values are 9080.5735$/h, 9200.4325$/h, 9132.5035$/h and 9108.9412$/h, 9250.1672$/h, 9188.56373$/h respectively with IEP and EP after 20 independent trials. From the results, the superiority of IEP strategies over EP can be noticed. The difference between the best and worst solutions are 119.859$/h with IEP. At the same time, the difference between the best and worst solutions is 141.226 $/h with EP. Moreover, the best and worst solutions obtained by IEP are very close to the average value, which proves that IEP is more robust and consistent. In conclusion, it is clearly shown that IEP is the most accurate and gives the exact minimum dispatch solution.

In order to verify algorithm effectiveness in more complex power system, IEEE 39-bus system is used test system. Tab. IX and Tab. X list the best control variables found by IEP and EP algorithm respectively. In Tab. IX, it is clearly shown that, by using IEP, the total production cost savings of 815.457$/h is obtained compared with EP algorithm. Hence, The same results are obtained. From Tab. XIII, the best, worst and average cost values are 460571.8601$/h, 461226.8524$/h, 460886.3652$/h and 461387.3171$/h, 462537.4354$/h, 461952.4763$/h respectively with IEP and EP after 20 independent trials. The difference between the best and worst solutions are 654.9923$/h with IEP. At the same time, the difference between the best and worst solutions is 1150.1183 $/h with EP.

TABLE VII
BEST SOLUTION OBTAINED USING IEP METHOD IN THE MODIFIED IEEE 30-BUS SYSTEM

Stage	1	2	3	4	5	6	7	8	9
P_{G1}(MW)	32.1628	26.881	30.61201	34.51722	29.78709	31.29296	31.07789	32.61336	16.72349
P_{G2}(MW)	57.00394	61.08434	55.17584	45.11469	52.62017	45.44802	44.2772	43.06717	57.77425
P_{G22}(MW)	23.05492	28.29231	24.03166	30.68287	29.96296	37.48977	29.89494	32.55328	22.52746
P_{G27}(MW)	25.79418	26.01045	30.52419	29.68614	28.98494	30.16027	26.72748	25.80592	20.62617
P_{G23}(MW)	21.3959	18.5878	15.68668	16.15716	9.585935	6.024802	13.87003	15.62889	18.70101
P_{G13}(MW)	32.11083	25.97463	26.30325	21.53214	22.25072	18.29059	18.17907	9.930773	18.55814

Total production cost: 9080.5735 $/h

TABLE VIII
BEST SOLUTION OBTAINED USING EP METHOD IN THE MODIFIED IEEE 30-BUS SYSTEM

Stage	1	2	3	4	5	6	7	8	9
P_{G1}(MW)	36.65525	43.1043	30.6769	17.91149	32.04969	15.60642	16.76836	13.58982	7.69469
P_{G2}(MW)	59.36403	58.5255	54.9398	51.61109	56.61143	47.39638	53.91842	60.12006	55.17884
P_{G22}(MW)	21.47479	18.14985	27.65491	30.40768	23.19058	29.44758	29.83708	36.22827	30.44646
P_{G27}(MW)	24.23683	19.73205	20.8879	31.6879	28.29225	31.76615	23.96418	21.04326	26.11672
P_{G23}(MW)	23.43493	17.99541	19.81188	18.72996	15.15849	17.75405	21.13465	10.16315	18.56167
P_{G13}(MW)	26.531	29.76962	28.29633	27.1601	18.00986	26.39931	18.31043	18.31929	16.81538

Total production cost: 9108.9412 $/h

TABLE IX
BEST SOLUTION OBTAINED USING IEP METHOD IN THE IEEE 39-BUS SYSTEM

Stage	1	2	3	4	5	6	7	8	9	10	11	12
P_{G30}(MW)	297.845	306.2852	302.6746	299.8669	274.7264	335.2846	301.1622	294.8944	308.8122	316.6097	302.4517	276.8535
P_{G31}(MW)	550.5877	600.1873	671.5014	605.8921	606.2954	538.8119	564.3921	584.1886	608.1819	607.222	600.5313	586.509
P_{G32}(MW)	626.5221	640.2318	569.7617	622.8446	616.9007	560.5899	594.6618	569.9685	600.4495	646.3009	639.911	624.6706
P_{G33}(MW)	634.3726	610.493	620.8039	614.5901	634.593	653.5109	654.2763	639.7166	623.09	573.336	598.1811	590.8908
P_{G34}(MW)	511.1227	476.2915	443.175	493.175	477.3373	510.5703	460.5703	472.3293	483.933	465.6822	487.3279	491.51
P_{G35}(MW)	610.7836	594.3052	627.8122	590.7639	587.3861	623.3719	618.3914	591.8316	585.7375	566.0117	556.7142	590.8262
P_{G36}MW)	568.042	555.3821	549.7241	570.1911	550.7421	579.9209	586.7932	616.7932	586.7932	587.416	567.623	579.9771
P_{G37}(MW)	527.2084	544.5856	545.6563	542.6359	555.7967	540.9971	550.4644	548.0077	553.644	562.9881	562.9605	539.0908
P_{G38}(MW)	849.1404	850.5165	832.8275	820.6171	850.6171	826.3819	827.0456	840.5518	829.1908	843.1028	819.0072	849.0072
P_{G39}(MW)	1008.702	1001.955	1011.994	1012.711	1002.851	988.7539	999.6285	999.6381	1012.127	1013.031	1042.002	1043.645

Total production cost : 460571.8601 $/h

TABLE X

BEST SOLUTION OBTAINED USING EP METHOD IN THE IEEE 39-BUS SYSTEM

Stage	1	2	3	4	5	6	7	8	9	10	11	12
P_{G30}(MW)	282.7469	238.4208	282.6066	338.1247	341.7531	333.3748	327.2719	306.0325	319.6207	322.2673	342.0072	305.8343
P_{G31}(MW)	594.6711	538.4092	504.6806	435.1558	475.9911	527.6847	564.3274	544.4797	607.5929	661.6504	646.2157	598.6778
P_{G32}(MW)	660.1812	689.5568	631.677	657.2323	622.9245	609.8651	565.3495	554.2467	531.8468	509.444	493.9961	502.6156
P_{G33}(MW)	633.0862	647.1537	657.152	695.4439	656.121	613.5017	634.5771	624.303	650.9625	654.4778	639.7837	643.3003
P_{G34}(MW)	497.4557	495.9905	523.4377	506.0668	491.8216	456.7792	470.8762	449.5858	423.7924	451.8563	445.8726	460.6822
P_{G35}(MW)	605.0152	616.4778	602.7038	583.1142	612.878	641.2923	592.1213	642.1213	650.9918	608.5196	650.9933	665.4863
P_{G36}MW)	544.6999	547.3016	570.5253	579.3933	565.1149	593.0327	567.4237	569.0068	574.9087	564.7232	548.5948	554.9131
P_{G37}(MW)	546.2254	571.3317	582.452	561.9026	563.9268	545.5942	575.5942	574.6134	573.9448	547.9221	550.402	529.3439
P_{G38}(MW)	817.0216	835.4741	852.515	841.1737	846.6516	871.43	864.4906	894.4906	880.61	898.6867	872.7176	902.7176
P_{G39}(MW)	1001.559	1002.257	974.3036	980.9517	982.4731	966.97	996.97	1002.012	981.4998	965.9979	989.3619	1013.062

Total production cost: 461387.3171 $/h

The average execution time taken to complete the fixed number of iterations (T_{fix}) and the average execution time taken to converge into the lower solution range (T_{low}) for 20 trials are shown in Tab. XII for the modified IEEE 30-bus system and Tab. XIV for IEEE 39-bus system respectively.

For the modified IEEE 30-bus system, EP takes an average execution time of 1800.23 sec. to complete 150 iterations. EP converges faster than IEP by reason of the small sub-memeplex generation number of IEP. In comparison to EP, IEP has additional components, i.e., the LS procedure. This extra burdens increase the execution time of IEP. IEP takes 1898.34 sec. more than EP to complete 150 iterations. Nevertheless, IEP takes only 1020.33 sec. to converge into the lower solution range (9080–9098$/h) , EP are not able to converge into the lower solution range.

For IEEE 39-bus system, EP takes an average execution time of 4247.25 sec. to complete 150 iterations. EP converges faster than IEP by reason of the small sub-memeplex generation number of IEP. In comparison to EP, IEP has additional components, i.e., the LS procedure. This extra burdens increase the execution time of IEP. IEP takes 4347.92 sec. more than EP to complete 150 iterations. Nevertheless, IEP takes only 2243.82 sec. to converge into the lower solution range (460571–

460671$/h) , EP are not able to converge into the lower solution range.

TABLE XI
COMPARISON OF BEST, WORST AND AVERAGE COST VALUES IN THE MODIFIED IEEE 30-BUS SYSTEM

Algorithms	Best ($/h)	Worst ($/h)	Average ($/h)
IEP	9080.5735	9200.4325	9132.5035
EP	9108.9412	9250.1672	9188.5637

TABLE XII
AVERAGE EXECUTION TIME COMPARISON IN THE MODIFIED IEEE 30-BUS SYSTEM

Methods	Average execution time (sec.)	
	T_{fix}	T_{low}
EP	1800.23	——
IEP	1898.34	1020.33

TABLE XIII
COMPARISON OF BEST, WORST AND AVERAGE COST VALUES IN THE IEEE 39-BUS SYSTEM

Algorithms	Best ($/h)	Worst ($/h)	Average ($/h)
IEP	460571.860	461226.8524	460886.365
EP	461387.317	462537.4354	461952.4763

TABLE XIV
AVERAGE EXECUTION TIME COMPARISON IN THE IEEE 39-BUS
SYSTEM

Methods	Average execution time (sec.)	
	T_{fix}	T_{low}
EP	4247.25	——
IEP	4347.92	2243.82

V. CONCLUSION

Considering the valve-point effect and ramp rate limits of conventional generators including VSWGs, DOPF model, which takes the all conventional units cost minimum as the objective function and takes the whole time and the inherent relations of different stages into account in wind power integrated system, is established. The PV-bus model of VSWGs bus is adopted in power flow calculation in this paper. A novel IEP is proposed for solving the established DOPF model and the detailed methods of the algorithm are given. The modified IEEE 30-bus system is used to illustrate the effectiveness of the proposed method compared with those obtained from EP algorithm. In order to verify algorithm effectiveness in more complex power system, IEEE 39-bus system is used test system. It is shown that the proposed method is capable of yielding higher-quality solutions.

ACKNOWLEDGMENTS

REFERENCES

[1] H. Li and z. Chen, "Overview of different wind generator systems and their comparisons," IET Renewable Power Generation, vol. 2,no.2, pp. 123-138,2008.

[2] X. Yao, y. Liu and G. Lin, "Evolutionary programming made faster," IEEE Trans. Evol. Comput., vol 3, no. 2, pp. 82-102, 1999.

[3] C. H. Liang, C. Y. Chung, and k. P. Wong etc., "Comparison and improvement of evolutionary programming techniques for power system optimal reactive power flow," IEE Proceedings-Generation, Transmission & Distribution, vol.153, no.2, pp.228-235, 2006.

[4] J. T. Ma and L. L. Lai, "Evolutionary programming approach to reactive power planning," IEE Proceedings-Generation, Transmission & Distribution,vol.143, no.4, pp.365-370, 1996.

[5] D. B. Fogel, "An introduction to simulated evolutionary optimization," IEEE Trans. Neural Networks, vol. 5, no.1, pp. 3–14, 1994

[6] R. Luss and T. H. I. Jaakola, "Optimization by direct search and systematic reduction of the size of the search region," AICHE J., vol.19, no.4, pp.760–766, 1973.

A. I. Selvakumar and K. Thanushkodi, "A new particle swarm optimization solution to nonconvex economic dispatch problems," IEEE Transactions on Power Systems,vol.22, no.1, pp.42-51, 2007.

[7] R. D. Zimmierman, C. E. M. Sanchez, and D. Gan :Matpower. a matlab power system simulation package. [online].available:
http://www.pserc.cornell.edu/matpower.

[8] R. Yokoyama, S.H. Bae, T.Morita, H. Sasaki, "Multi-objective optimal generation dispatch based on probability security criteria," IEEE Transactions on Power Systems, vol. 3,no. 1, pp. 317–324, 1988.

[9] T. A. A. Victoire and A. E. Jeyakumar, "A modified hybrid ep–sqp approach for dynamic dispatch with valve-point effect," International Journal of Electrical Power And Energy Systems, vol. 27,no. 8, pp. 594–601, 2005.

Gonggui Chen was born in Hubei, China in 1964. He received his B.S. degree in physics from Huazhong Normal University, and M. Eng. degree in computer technology from Huazhong University of Science and Technology(HUST) , in 1987 and 2004 respectively. He is at present a professor with the Dept. of electrical engineering, Hubei University for Nationalities, and currently pursuing the PH.D. degree in electrical engineering in HUST. His research interests include power system analysis and operation, distributed generation and application of artificial intelligence.

Time-Varying Sliding Mode Adaptive Control for Rotary Drilling System

Lin Li

Key Laboratory of Drilling Rigs Controlling Technique, Xi'an Shiyou University, Xi'an, China
E-mail: kjclilin@ xsyu.edu.cn

Qi-zhi Zhang and Nurzat Rasol

Key Laboratory of Drilling Rigs Controlling Technique, Xi'an Shiyou University, Xi'an, China
E-mail: zhangqzqz@ gmail.com
E-mail: altiy@126.com

Abstract — **This paper presents a time-varying sliding mode adaptive controller in order to handle the stick-slip oscillation of nonlinear rotary drilling system. The time-varying sliding mode controller with strong robust has two time-varying sliding surfaces, one of them induced time-varying integral sliding mode control can control the transient stage of the rotary drilling system and ensure the system remains the sliding condition whatever in usual or existing the parameter changes and disturbances to arrive at a controller capable of global stability. The herein developed controller is, a time-varying sliding mode adaptive controller has tracking performance and identification of drilling parameters. Lyapunov principles have been carried out to verify the stability and robustness of system. The simulation results show that the controller has faster dynamic responses and suppress stick-slip in oil well drill string, can achieve global stability of rotary drilling system.**

Index term — **time-varying sliding mode control, adaptive control, stick-slip, rotary drilling system, nonlinear system**

I.   INTRODUCTION

The complexity of drilling process, the uncertainties of rock formation and the operation characteristics of drilling rig result in unstable behavior and drill string component failures. Stick-slip phenomenon appearing at the bottom-hole assembly (BHA) is particularly harmful for the bit. When drillstring rotation begins, the drillpipe stores torsional energy until the applied torque exceeds the total static frictional torque on the BHA. The BHA then begins to rotate, and because the static friction is higher than the dynamic friction, the stored energy in the drillpipe is transferred to inertial energy in the BHA. It then can accelerate to a speed faster than steady-state rotational speed [1].

The great practical significance of oilwell drillstrings has interested some researchers. Some researchers hold the structure of bit is a major cause of the stick-slip oscillation, so they study to the mechanical structure [2]-[3] and the size [4]-[5] of the bit and analyze the stress of the bit in the drilling process[6]-[7]. At the same time, various solutions have been proposed in the

literature for controlling rotary system oscillation and to manipulate this problem of instability. For example, classical controller as PID [8], Backstepping control [9], H∞ technique on the local linearized model [10], and finally, sliding-mode control has been effectively used in many practical control problems [11]-[12]. Now the proposed sliding control is a traditional sliding control way, although it has robustness、 good dynamic and static characters, it can only achieve in the sliding surface, the transient stage of system has not any robustness during existing large errors or disturbances.

The paper presents time-varying sliding mode controller based on the nonlinear equation of the rotary drilling system for the stick-slip phenomena and the drawback of traditional sliding mode controller. It can achieve the global stability and robustness through the second order integral sliding surface during existing large error or disturbance and the numerical simulations have been carried out to verify the idea.

But we assume that the range of the parameters is known in the designed controller, however, it is difficult to accurately defined in the practical system, so the paper presents the way of combined the adaptive control with the time-varying sliding mode system and designs a two-layer sliding mode adaptive controller for rotary drilling system with adopting parameter adaptive method which can real time adjust controller parameters and provide the highlight advantages of the controller.

II.   DRILLING WELL COMPONENTS

Deep wells for the exploration and production of oil and gas are drilled with a rotary drilling system. The basic components of a rotary drilling rig are the derrick and hoist, swivel, kelly, turntable, drill pipes, bit, and pump as shown in Fig.1. The torque driving the bit is generated at the surface by a motor with a mechanical transmission box or the top-drive. The medium to transport the energy from the surface to the bit is a drillstring, mainly consisting of drill pipes, drill collars and bit. The drillstring can be up to 8km long. The bottom end of the drillstring is the bottom-hole-assembly (BHA) consisting of drill collars and the bit, which

provides weight on the bit (WOB) required to generate accurate cutting force. During the process drilling, the drilling fluid (mud) is continuously circulated to the bottom of the hole and back to surface to remove cuttings from the bottom of the hole, to cool and lubricate the bit, and to control downhole pressures.
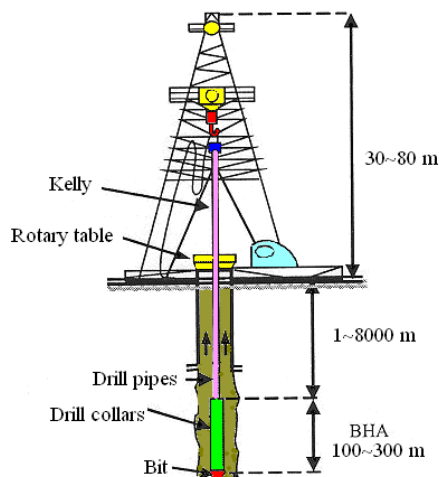

Figure 1.   The rotary drilling rig

III. TORSIONAL MODEL OF A DRILLSTRING

The rotary drilling rig is an essential part of oil drilling which provides enough torque and rotary speed for the bit and the drilling devices. The basic components of a rotary drilling rig are the derrick and hoist, swivel, kelly, turntable, drill pipes, bit, and pump.

Fig.2 depicts a simplified torsional model of the drill-string. The model essentially, consists of two damped inertias mechanically coupled by an elastic intertia less shaft (drillstring).
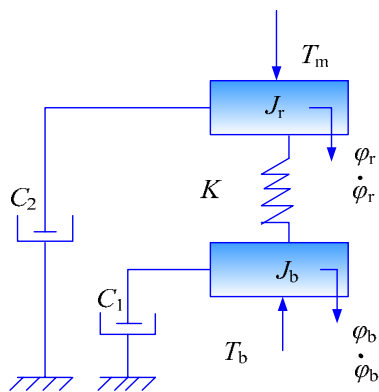
Figure 2.   Drilling rotary system model

Some assumptions are made, such as: (a) the drillstring is homogenous along its entire length and simply considered as a single linear torsional spring with stiffness coefficient k. (b) the borehole and the drillstring are both vertical and straight, (c) no lateral hit motion is present, (d) the friction in the pipe connections and between the pipes and the borehole are neglected, (e) the drilling mud is simplified by a viscous-type friction element at the hit, (f) the drilling mud fluids orbital motion is considered to be laminar, i.e., without

turbulences [13]. Then, the equations of motion are given in [12] as

$$\begin{cases} J_b\ddot{\varphi}_b + C_1\dot{\varphi}_b - k(\varphi_r - \varphi_b) = -T_{tob}(\dot{\varphi}_b) \\ J_r\ddot{\varphi}_r + C_2\dot{\varphi}_r + k(\varphi_r - \varphi_b) = T_m \\ T_m = C_2\dot{\varphi}_{ref} + u \end{cases} \quad (1)$$

where φ_b is the angular displacement of the bit, φ_r is the angular displacement of the rotary table or the top-drive, and $\dot{\varphi}_{ref}$ is the desired velocity of the bit. J_b is the equivalent of moment of the inertia of the collars and the drillpipes, and J_r represents the inertia of the rotary table. C_1 is the equivalent viscous damping coefficient of BHA, and C_2 is the viscous damping coefficient of the rotary table. T_{tob} is a nonlinear function which will be referred to be the torque on-bit, and T_m is the torque delivered by the motor to the system.

In the oil drillstring, the stick-slip oscillations are driven by nonlinear friction T_{tob} at near-zero bit velocities. T_{tob} represents the combined effects of reactive torque on the bit and nonlinear frictional forces along the BHA. Fig.3 shows the excitation of torsional vibrations leading to the phenomena of stick-slip, by nonlinear friction torque between the drill bit and the rock formation. The friction torque T_{tob} as a function of the bit speed is given by the following nonlinear function:

$$T_{tob}(\dot{\varphi}_b) = T_{tobdyn}\frac{2}{\pi}(\alpha_1\dot{\varphi}_b e^{-\alpha_2|\dot{\varphi}_b|} + \arctan(\alpha_3\dot{\varphi}_b)) \quad (2)$$

Where T_{tobdyn}=0.5kNm, α_1=9.5, α_2=2.2, and α_3=35.
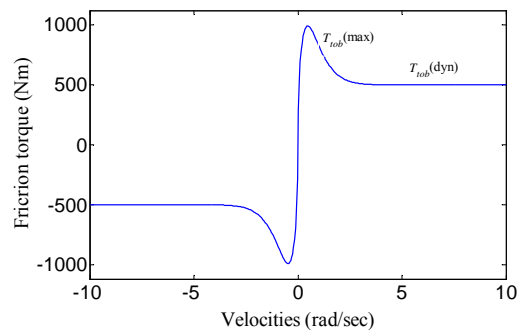

Figure 3.   Torque as a function of the angular velocity at the bit

IV. CONTROLLER DESIGN

In this section, the main task is to design a controller to avoid stick-slip oscillations and optimize drilling process. The controller design procedure consists of two steps: first, in order to compensate the nonlinearity in the drilling caused by the oscillation of stick-slip and simplify the design of sliding surface, a description of the input-state linearization controller is derived; second, the two-layer time-varying sliding mode controller based on the linear equation of rotary drilling system is designed.

A. Input-State Linearization Controller

Consider the following nonlinear equation of the rotary drilling system given in (1)-(2):

$$\begin{cases} \dot{x}(t) = f(x(t)) + g(x)u(t) \\ y(t) = h(x(t)) \end{cases} \qquad (3)$$

where $x(t) = [\dot{\varphi}_b, \varphi_r - \varphi_b, \dot{\varphi}_r]^T \in R^3$ is the state vector, $y(t) \in R$ is the measured output variable,

$$f(x) = \begin{bmatrix} -a_1\dot{\varphi}_b + b_1(\varphi_r - \varphi_b) + c_1 T_{tob}(\dot{\varphi}_b) \\ \dot{\varphi}_r - \dot{\varphi}_b \\ -a_2\dot{\varphi}_r - b_2(\varphi_r - \varphi_b) + c_2\dot{\varphi}_{ref} \end{bmatrix}$$

$$g(x) = \begin{bmatrix} 0 & 0 & c_2 \end{bmatrix}^T$$

$$h(x) = \dot{\varphi}_b$$

where

$$a_1 = \frac{C_1}{J_1}, \ b_1 = \frac{k}{J_1}, \ c_1 = \frac{1}{J_1}, \ a_2 = \frac{C_2}{J_2}, \ b_2 = \frac{k}{J_2}, \ c_2 = \frac{1}{J_2}$$

For reducing calculated amount the angular velocity of the bit is acquired by means of the input state. To find out the state transformation z, in case

$$\begin{cases} z_1 = \dot{\varphi}_b \\ z_2 = \dot{z}_1 = -a_1\dot{\varphi}_b + b_1(\varphi_r - \varphi_b) + c_1 T_{tob}(\dot{\varphi}_b) \\ z_3 = \dot{z}_2 = Hz_2 + b_1(\dot{\varphi}_r - \dot{\varphi}_b) \end{cases} \qquad (4)$$

where for $\dot{\varphi}_b$

$$H = -a_1 + c_1\left[\frac{-2T_{tobdyn}}{\pi}(\alpha_1 e^{-\alpha_2 z_1} - \alpha_1\alpha_2 z_1 e^{-\alpha_2 z_1} + \frac{\alpha_3}{1+\alpha_3^2 z_1^2})\right] T$$

hen the new state equations are in the canonical form

$$\begin{cases} \dot{z}_1 = z_2 \\ \dot{z}_2 = z_3 \\ \dot{z}_3 = Mz^2{}_2 + H^2 z_2 + H(z_3 - Hz_2) + b_1 Q + b_1 c_2 u \end{cases} \qquad (5)$$

where

$$Q = -\frac{a_2}{b_1}(z_3 - Hz_2) + (a_1 - a_2)z_1 - (b_1 + b_2)\varphi + a_2\dot{\varphi}_{ref} - c_1 T_{tob}(z_1)$$

$$M = \frac{2c_1 T_{tobdyn}}{\pi}\left[e^{-\alpha_2 z_1}(2\alpha_1\alpha_2 - \alpha_1\alpha_2^2 z_1) + \frac{2\alpha_3^2 z_1}{(1+\alpha_3^2 z_1^2)^2}\right]$$

$$\varphi = \varphi_r - \varphi_b$$

Now, it is easy to observe that the input state vector

$$u_{lin} = \frac{1}{b_1 c_2}\left[v - (M + H^2)z_2 - H(z_3 - Hz_2) - b_1 Q\right] \qquad (6)$$

B. Sliding Mode

The sliding mode control approach (see for example) [14]-[16] leads to a controller which can be stabilized over a wide range of operating conditions and is robust with respect to parameters variations. To track the bit angular velocity to the target, the error e is simply defined as

$$e = \dot{\varphi}_b - \dot{\varphi}_{ref} \qquad (7)$$

where $\dot{\varphi}_{ref}$ is the desired state of the system.

Therefore, the sliding surfaces can be chosen as

$$s = \lambda e \qquad (8)$$

The input u is becoming

$$u = u_{lin} + u_{sm} \qquad (9)$$

where u_{lin} is the input-state linearization controller defined in (9), u_{sm} is the sliding mode controller.

Usually, we can design sliding mode controller based on various reaching laws:

a)   Constant rate reaching law

$$u_{sm} = -k \cdot \text{sgn}(s)$$

b)   The index number reaching law

$$u_s = -k \cdot \text{sgn}(s) - \lambda_1 s$$

c)   The exponent reaching law

$$u_s = -k|s|^\alpha \cdot \text{sgn}(s) - \lambda_1 \cdot s$$

C. Two-Layer Time-Varying Sliding Mode Controller

Two-layer time-varying sliding mode controller is presented for the drilling rotary system. Two-layer time-varying sliding mode controller has two time-varying sliding surfaces. One of them is stationary surface, as main surface; the other surface is a time-varying surface with integral sliding mode control. The initial state of drilling rotary system runs in the second order sliding surface that can control the reaching stage of main sliding surface, and the transient process of system can remain insensitive to parameter variations and other disturbances; at the same time, the reaching stage of the second order sliding surface can be cancelled because of the initiate state runs in there. The drilling rotary system is in the sliding condition all the time, therefore, the system with strong robustness can resist the stick-sliding oscillation of drill bit and ensure the global stability of system [17].

Appropriate time-varying sliding surfaces are most important to the design of sliding mode controller. The error e in the drilling rotary system is defined as:

$$e_1 = z_1 - \Omega_{ref} \qquad (10)$$

According to (5), these errors can be considered:

$$\begin{aligned} e_2 &= \dot{e}_1 = \dot{z}_1 = z_2 \\ e_3 &= \ddot{e}_1 = \ddot{z}_1 = z_3 \end{aligned} \qquad (11)$$

Then main sliding surface is defined as:

$$s_1(t) = \lambda e \qquad (12)$$

where:

$$\lambda = [\lambda_1 \ \lambda_2 \ 1], \quad e = [e_1 \ e_2 \ e_3]^T$$

The second sliding surface induced integral sliding mode control in order to increase the accuracy and robustness [18]-[19], and it is defined as:

$$s_2(t) = s_1(t) + A\int s_1(\tau)d\tau + Q(t) \qquad (13)$$

where

$$Q(t) = \begin{cases} Bt^2 + Ct + D & t \leq T \\ 0 & t > T \end{cases}$$

A, B, C, D are real constants, and A>0.
When the initial state $t=0$, the $s_2(x, 0)=0$, hence

$$D = -\lambda_1(z_1 - \Omega_{ref}) - \lambda_2 z_2 - \lambda_3 z_3 \qquad (14)$$

s_2 and its derivative is continuous, so

$$BT^2 + CT + D = 0$$

$$2BT + C = 0$$

Thus

$$B = \frac{D}{T^2}, \quad C = -\frac{2D}{T}$$

After this, we design the controller that not only ensures the sliding motion existence in the two-layer surface under the certain condition, but also the main sliding surface and the tracking error approximate to zero during the limited time.

The derivative of second order sliding surface is

$$\dot{s}_2(t) = \dot{s}_1(t) + As_1(t) + \frac{dQ}{dt} \qquad (15)$$

$$= \lambda_1 z_2 + \lambda_2 z_3 + \dot{z}_3 + As_1(t) + \frac{dQ}{dt}$$

The state vector can be chosen according to the input state linearization control law [20]:

$$u = \frac{1}{b_1 c_2}[-Mz_2^2 - H^2 z_2 - H(z_3 - Hz_2) - b_1 Q \qquad (16)$$

$$- \lambda_1 z_2 - \lambda_2 z_3 - As_1(t) - \frac{dQ}{dt} - ks_2(t)]$$

V. THE STABILITY ANALYSIS OF DRILLING ROTARY SYSTEM

Oilwell drillstrings are mechanical system which undergo complex dynamical phenomena, often involving non-desired oscillations. These oscillations are a source of failures which reduce penetration rates and increase drilling operation costs. So it is an advisable decision to choose sliding mode controller for drilling rotary system. However, the traditional sliding mode controller for the transient state has not any robustness, when there are a large error or disturbance in the system, it can not remain its good performance, but the time-varying sliding mode control exception. There is the stability analysis of rotary drilling system as follows [21].

We can choose a positive Lyapunov function based on the error dynamics of the system as:

$$V_2 = \frac{1}{2}s_2^2(t) \qquad (17)$$

and its time derivative is

$$\dot{V}_2 = s_2(t)\dot{s}_2(t) \qquad (18)$$

Using (16), relation (15) takes following forms

$$\dot{s}_2(t) = -ks_2(t) \qquad (19)$$

so

$$\dot{V}_2 = -ks_2^2(t) \leq 0 \qquad (20)$$

Equation (20) indicates the sliding motion in the second order surface is existent and stable.

Also, we can choose another positive Lyapunov function based on the error dynamics of the system as:

$$V_1 = \frac{1}{2}s_1^2(t) \qquad (21)$$

and the time derivative of (21) as

$$\dot{V}_1 = s_1(t)\dot{s}_1(t) \qquad (22)$$

There are $s_2(t)=0$ in the second surface, then

$$s_1(t) + A\int s_1(\tau)d\tau + Q(t) = 0 \qquad (23)$$

and the time derivative of (23) as

$$\dot{s}_1(t) = -As_1(t) - \begin{cases} 2Bt + C & t \leq T \\ 0 & t > T \end{cases} \qquad (24)$$

so

$$\dot{V}_1 = -As_1^2(t) - s_1(t)\begin{cases} 2Bt + C & t \leq T \\ 0 & t > T \end{cases} \qquad (25)$$

When $t>T$, then $\dot{V}_1 = -As_1^2(t) < 0$, there has the sliding motion that is stable according to Lyapunov principle. For $t \leq T$

$$s_1(t) = \left[s_1(0) + \frac{C}{A} - \frac{2B}{A^2}\right]e^{(-At)} - \frac{2B}{A}t - \frac{C}{A} + \frac{2B}{A^2} \qquad (26)$$

For $t > T$

$$s_1(t) = s_1(T)e^{[-A(t-T)]} \qquad (27)$$

When $t \to \infty \ (t > T)$, $s_1 \to 0$ and consider with (26)-(27), the reaching stage of main sliding surface is the index number reaching, Previously, that has not any robustness in the traditional sliding mode control, but the reaching stage of $s_1(t)=0$ runs in the second order sliding surface of the time-varying sliding mode controller.

In the above, the drilling rotary system with the time-varying sliding mode controller has strong robustness to parameter changes and disturbances from beginning to end, as well as the system is in the sliding condition along the $s_2(t)=0$ until the s_1 approximate to zero, then it also is in sliding condition along the $s_1(t)=0$, so the whole system can achieve global stability and has strong robustness.

With $s_1(t) = 0$:

$$\sum_{i=1}^{n-1} k_i e^{(i-1)} + e^{(n-1)} = 0$$

where $k_1 + k_2 p + \ldots + k_{n-1}p^{n-2} + p^{n-1}$ is Hurwitz polynomials, so the error $e = y - y_d$ will approximate to zero.

VI. TIME-VARYING SLIDING MODE ADAPTIVE CONTROL

There presents a sliding mode adaptive control for in the oil drilling system with large parameter changes and uncertainties which combines the adaptive control with sliding mode control. The sliding mode adaptive control can improve the control characteristic through adopting parameter adaptive control、on-line identification and tacking technology [17].

If M, H is uncertain, using estimated parameter, relation (16) takes the following form:

$$u = \frac{1}{b_1 c_2}[-\hat{M}z_2^2 - \hat{H}^2 z_2 - \hat{H}(z_3 - \hat{H}z_2) - b_1 Q \qquad (28)$$

$$- \lambda_1 z_2 - \lambda_2 z_3 - As_1(t) - \frac{dQ}{dt} - ks_2(t)]$$

where $\hat{M}, \hat{H}$ are estimate of M, H.

We can choose a positive Lyapunov function V as:

$$V = \frac{1}{2}s_2^2 + \frac{1}{2}\rho_1\tilde{M}^2 + \frac{1}{2}\rho_2\tilde{H}^2 \qquad (29)$$

where $\tilde{M} = M - \hat{M}$, $\tilde{H} = H - \hat{H}$。

The time derivative of (29) as:

$$\dot{V} = s_2\dot{s}_2 + \rho_1\tilde{M}\dot{\hat{M}} + \rho_2\tilde{H}\dot{\hat{H}} \qquad (30)$$

Using (28), relation (15) takes following forms:

$$\dot{s}_2(t) = \dot{s}_1(t) + As_1(t) + \frac{dQ}{dt} \qquad (31)$$

$$= \tilde{M}z_2^2 + \tilde{H}^2 z_3 + \tilde{H}(z_3 - \tilde{H}z_3) - ks_2(t)$$

so

$$\dot{V} = \tilde{M}(s_2 z_2^2 - \rho_1\dot{\hat{M}}) + \tilde{H}(s_2 z_3 - \rho_2\dot{\hat{H}}) - ks_2^2 \qquad (32)$$

We can choose following adaptive low according to the Lyapunov principle:

$$\begin{cases} \dot{\hat{M}} = \dfrac{s_2 z_2^2}{\rho_1} \\[3mm] \dot{\hat{H}} = \dfrac{s_2 z_3}{\rho_2} \end{cases} \qquad (33)$$

Where $\dot{V} \leq 0$, semi-positive, and V is positive, so the system is stable according to Lyapunov method.

Considering the characteristic of the practical drilling, combined with the controller given (28) with adaptive low given (33) is an advisable way for the controlling rotary drilling system with many uncertainties. The rotary drilling system with time-varying sliding mode adaptive controller with strong robustness has the characteristic of identification and tracking, it can achieve global stability.

VII. SIMULATION RESULTS AND DISCUSSION

The time-varying sliding mode control has high quality for the system. This section presents the simulation results of the drilling rotary system with time-varying sliding mode controller.

The parameters, used for the simulation are taken from [12]. In addition these parameters are typical in oil well drilling operations.

Fig.4 shows the step response of the bit angular velocity without the controller. The effects of stick slip oscillations can be seen at the bit speed, which take more than 100 second to vanish.
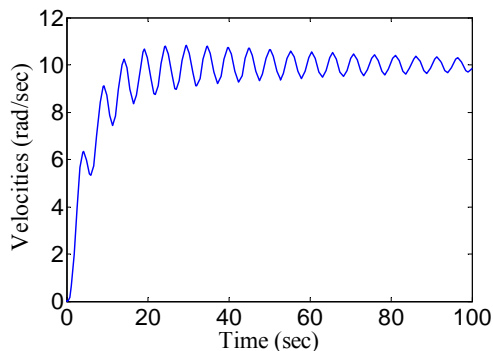
Figure 4.   The step response of the bit angular velocity without the controller

Previously, we have designed the sliding mode PID controller with different reaching law and adaptive PID controller for rotary drilling system. Now we compare the characteristics of the time-varying sliding mode adaptive controller with previous work.

The Fig.5 shows the step response of bit angular velocity with sliding mode PID controller with different reaching law. From figure can be seen, the reaching law can affect the efficiency of suppressing stick-slip oscillation, where the exponent reaching law has good character for fast reaching spend and the constant rate reaching law although can suppress the stick-slip oscillation, the response curve of system is not smoothing [22].
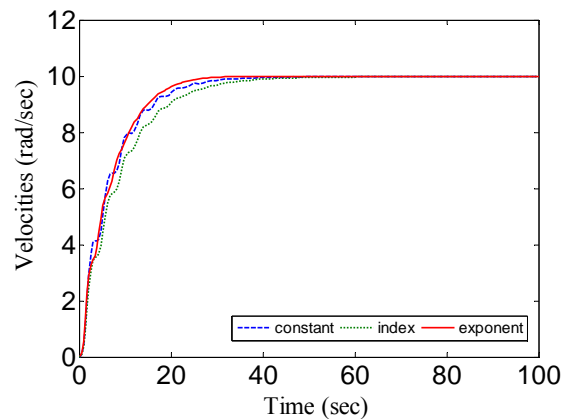

Figure 5.   The step response of the bit angular velocity

The drawbacks of tradition sliding controller exist in the sliding controller with different reaching law. However, the time-varying sliding control can guarantee the stability of main sliding surface sliding motion and reaching stage.

Fig.6 shows the step response the drilling rotary system with adaptive PID controller. As seen the figure, although the adaptive PID controller has certain robustness, the drillingstring length changes bring some oscillation to system but the time-varying sliding controller exception, which is insensitive to drilling length changes
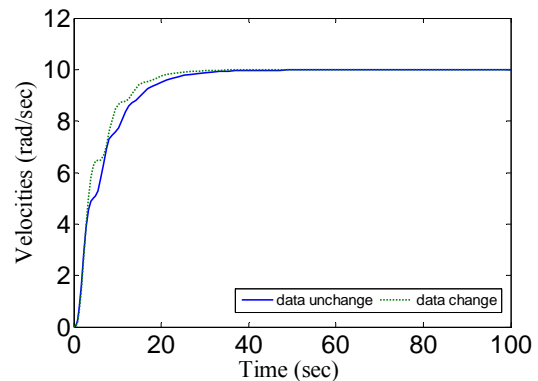

Figure 6.   The step response of the bit angular velocity in adaptive PID controller

Fig.7 shows the step response of the bit angular velocity with the time-varying sliding mode controller. In

simulation sliding mode controller parameter λ=[0.24 0.6 1]，A=0.29，k=0.0002，T=0.1s.

As seen in the Fig.7, rise time $t_r\approx12s$, settling time $t_s\approx18s$, overshoot $\sigma\%$=0, steady-state error e_{ss}=0. Because of the second order sliding surface with integral sliding mode control, there are no more stick-slip oscillations as arising large errors or disturbances and increase the accuracy of system. All in all, the drilling rotary system with time-varying controller has faster response, dynamic and static characters.
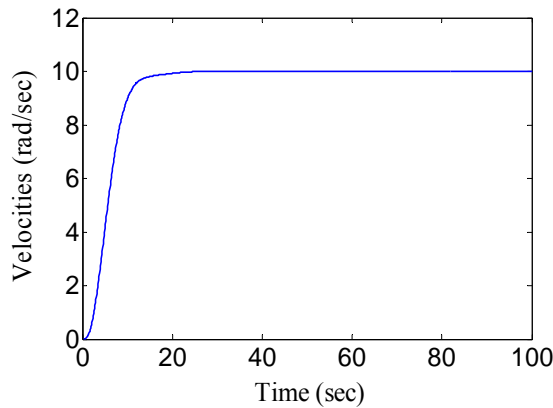
Figure 7.   The step response of the bit angular velocity

If M, H is uncertain, we can get similar results through the tracking performance and identification ability of adaptive controller.

VIII. CONCLUSION

In this paper, we have proposed a time-varying sliding mode adaptive controller for handling the stick-slip oscillation and the drawbacks of traditional sliding mode control, from comparing the simulation results, we can get as follows:

   a) Time-varying sliding mode controller can ensure the stability and the robustness of transient stage.
   b) Time-varying sliding mode controller can suppress stick-slip in oil well drill string under existing large error or disturbance, achieve the global stability of rotary drilling system.
   c) Integral sliding mode control induced by time-varying sliding mode controller can increase the robustness and stable accuracy of the controller.
   d) The rotary drilling system has tracking performance and identification of drilling parameters from combined the sliding control with adaptive control.

ACKNOWLEDGMENT

APPENDIX   PARAMETER VALUES

PARAMETERS USED IN THE SIMULATION

Symbol	Description	Values
J_b	Drill bit inertia	374 kgm^2
J_r	Rotary table + motor inertia	2122 kgm^2
C_1	BHA damping	0.5 Nms/rad
C_2	Rotary table damping	425 Nms/rad
K	Drillstring stiffness	473 Nms/rad
Ω_{ref}	Drill bit reference velocity	10rad/s

REFERENCES

[1] J.F.Brett. The Genesis of Torsional Drillstring Vibrations. SPE Drilling Engineering, vol. September, 1992: 168-174.
[2] G. Mensa-Wilmot, M. Booth and A. Mottram. New PDC Bit Technology and Improved Operational Practices. Saves IM in Central North Sea Drilling Program. in the SPE/IADC Drilling Conference. New Orleans, 2000. SPE/IADC 59108.
[3] Peng Ye. The Theory and Testing Study of Two-stage Bit[D]. China University of Petroleum Doctoral Dissertation, 2008.
[4] Zhang Shaoping. Study and Application of Bit Selection for Zhongyuan Oil Field[D]. China University of Petroleum Doctoral Dissertation, 2007.
[5] Pan Qifeng, Gao Deli, Sun Shuzhen, Sun Xiangcheng. A New Method for PDC Selection[J]. Acta Petrolei Sinica, 2005, 26(3): 123-126.
[6] Guo Jian, Sun Wenlei. The Stress Analysis PDC Drill Bits in the Process of Drilling[J]. Machine Tool and Hydraulics, 2008, 36(12): 25-27.
[7] J.D. Macpherson and PN. Jogi and J.E.E. kingman. Application and Analysis of Simultaneous near Bit and Surface Dynamics Measurements. SPE Drilling and Completion, 2001, 16(4): 230-238.
[8] F.Abbassian and V.A.Dunayevsky. Application of Stability Approach to Torsional and Lateral Bit Dynamics. SPE Drilling and Completion, 1998, 13(2): 99-107.
[9] F. Abdulgalil and H. Siguerdidjane. Backstepping Design for Controlling Rotary Drilling System. Proceedings of the 2005 IEEE Conference on Control Applications Toronto, Canada, August, 2005: 120-124.
[10] A. F. A. Serrarens. H∞ control as applied to torsional drillstring dynamics. Msc. Thesis, Eindhoven University of Technology, 2002.
[11] Eva M. Navarro-López and Domingo Cortés. Sliding-mode control of a multi-DOF oilwell drillstring with stick-slip oscillations. Proceedings of the 2007 American Control Conference New York City, USA, July, 2007.
[12] F. Abdulgalil and H. Siguerdidjane. PID Based on Sliding Mode Control for Rotary Drilling System. Serbia & Montenegro, Belgrade, November, 2005.
[13] Eva María Navarro-López and Rodolfo Suárez, "Practical Approach to Modelling and Controlling Stick-slip oscillations in oilwell drillstrings", Proceedings of the 2004 IEEE International Conference on Coutrol Applications Taipei. Taiwan, September 2-4, 2004, pp.1454-1460.
[14] A.S.Yigit and A.P.Christoforou, "Coupled Torsional and Bending Vibrations of Actively Controlled Drillstring

Journal of Sound and Vibration, vol. 234, no.1, 2000, pp. 67-83.

[15] N.Mihajlovic, Torsional and Lateral, "Vibrations in Flexible Rotor Systems with Friction", Technische Universiteit Eindhoven, 2005.

[16] N.Mihajlovic, N.van de Wouw, R.C.J.N.Rosille and H.Nijmeijer, "Interaction between torsional and lateral vibrations in flexible rotor systems with discontinuous friction", Nonlinear Dynamics, vol. 50, no. 3, November 2007, pp 679-699.

[17] Guan Cheng. Sliding Mode Adaptive Control of Nonlinear System and Application to Electro-Hydraulic Control System[D]. Zhejiang University Doctoral Dissertation, 2005.

[18] Hu Qinglei, Ma Guangfu, Jiang Ye, Liu Yaqiu. Variable structure control with time-varying sliding mode and vibration control for flexible satellite[J]. Control Theory & Application, 2009, 26(2): 122-126.

[19] Guan Cheng, Zhu Shan-an. Dervative and Integral Sliding Mode Adaptive Control for A Class of Nonlinear System and Its Application to An Electro-Hydraulic Servo System[J]. Proceedings of the CSEE, 2005, 25(4): 103-108.

[20] F. Abdulgalil, H. Siguerdidjane. Nonlinear control design for suppressing stick slip oscillations in oil well drillstrings, 5th ASCC, Melbourne, July 2004.

[21] Xu Wenlin, Wu Ronghui. Lyapunov's Indirect Method for Stability Analysis of Fuzzy Control System[J]. Journal of Hunan University, 2004, 31(3): 86-89.

[22] ZHANG Qi-zhi, HE Yu-yao, LI Lin. Sliding Mode Control of Rotary Drilling System With Stick Slip Oscillation. The 2nd International Workshop on Intelligent Systems and Applications (ISA2010), Wuhan, China, May, 2010: 30-33.

Lin Li was born in 1963, and received his M.S degree in 1991 with auto-mearment technique from Xi'an Jiaotong University.

Since 1993 he has been a professor at the Electrical Engineering Department and the Chief of Science and Technology Division in the Xi'an Shiyou University, published *Automaic Technology for Long Distance Pipeline*, Petroleum Industry Press, 2005, Beijin; *Automaic Technology for The Electronic Drilling Rig*, Petroleum Industry Press, 2009, Beijin. His main interests at moment are the automatic control of electronic drilling rig.

Mr .Li is a membership of China Petroleum Institute and Deputy Chairman of Shaanxi High-Tech Association, has worn second place in Science and Technology Prize of Shaanxi Province and first prize in Science and Technology Prize of Shaanxi Advanced School.

Qizhi Zhang was born in 1965, and received the B.S. degrees in electronic control technical in 1987 from Beijing University of Aeronautics &Astronautics and M.S. degrees in control theory and engineering in 1992 from college of automatic control, Northwestern Polytechnical University. She is currently a professor of control science and engineering in the Key Laboratory of Drilling Rig Controlling Technique, Xi'an Shiyou University. Her main interests at the moment are the automatic control of electronic drilling rig.

Nurzat.Rasol was born in 1984, and received her B.S. degree in control theory and engineering from Xi'an Shiyou University in 2002, now a M.S. candidate at the Key Laboratory of Drilling Rigs Controlling Technique. Her research interests are learning real-time drilling monitoring and control system.

Delay-dependent Robust H∞ Control for T-S Fuzzy Descriptor Networked Control System with Multiple State Delays

Baoping Ma

School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing, China
Email: mabaoping@njnu.edu.cn

Qingmin Meng

Institute of Signal and Transmission, Nanjing University of Posts and Telecommunications, Nanjing, China
Email: mengqm@njupt.edu.cn

Abstract—The problem of delay-dependent robust H_∞ control for T-S fuzzy descriptor networked control systems with multiple state time-delays is investigated. First, in terms of linear matrix inequality (LMI) approach, an improved delay-dependent criterion is established to ensure that the descriptor system is regular, impulse free and stable with H_∞ performance. Without using model transformation and bounding technique for cross terms, the criterion is less conservative. With the proposed criteria, we are able to obtain maximum allowable delay bound (MADB) of NCS. Numerical example shows that the proposed criteria are effective and less conservative.

Index Terms—delay-dependent, fuzzy descriptor system, networked control system, linear matrix inequality

I. Introduction

In a Networked Control Systems (NCS) all the components of the control system such as sensors, controllers, and actuators along with the plant (to be controlled) are all connected via a real-time network, and the information (reference input, plant output, control effort etc.) between all these devices are all exchanged through communication channels[1]. Such NCS have received increasing attentions in recent years due to their low cost, simple installation and maintenance, and high reliability. In spite of these advantages, insertion of a communication network into the feedback path makes the analysis and design of NCS quite difficult. The study of NCS is an interdisciplinary research area, combining both network and control theories. Hence, in order to guarantee the stability and performance of an NCS, analysis and design tools based on both network and control parameters are required. The stability problem of NCS has been well investigated in [1]–[3].

Consider a typical NCS with network-induced delays

as shown in Fig. 1 where $\tau_{sc}(t)$ is the delay from the sensor to the controller and $\tau_{ca}(t)$ is the delay from the controller to the actuator; $\tau_c(t)$ is the delay taken for processing the control signal in the controller i.e.
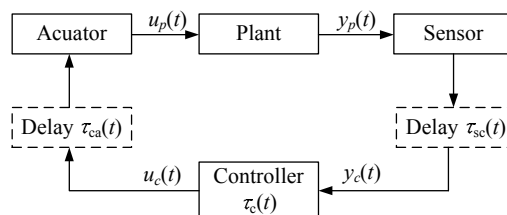

Figure 1. Networked Control Systems

computational delay of the processor. This delay is usually small when compared to network-induced delays since processor coded with the control algorithm works at a higher speed; hence, for the sake of simplicity, it is usually ignored in the analysis of NCS. But, if the processor speed is slow, then the effects of the computational delay in the controller cannot be neglected.

Three main issues raised in NCS are network-induced delay, data transmission dropout, and bandwidth and packet size constraints. As is known, network induced delay can significantly degrade the performance and even lead to instability of NCS, so the maximum allowable size of network-induced delay are an important indexes in the sense of guaranteeing system stability. Many researchers have studied stability analysis and controller design for NCS with linear controlled plant (namely, linear NCS) in the presence of network-induced delay[2][3][4]. Without loss of generality, a model of closed-loop MIMO NCS with structured uncertainties and multiple state time-delays is given [5].

Fuzzy control is a useful approach to solve the control problems of nonlinear systems. Takagi-Sugeno (T-S) fuzzy system proposed by [6] is a popular and convenient tool to approximate nonlinear systems because of its simple structure with local dynamics. Descriptor system, which are also referred to as singular systems, implicit systems , generalized state-space systems, differential-

algebraic systems, have been extensively studied for many years as descriptor systems are more general and natural in describing the practical dynamical systems than standard state-space systems. In [7], a fuzzy model in the descriptor form is introduced, and stability and stabilization problems for the system are addressed. Recently, more and more attention has been paid to the study of fuzzy descriptor systems [8-10]. Therefore, it is meaningful to employ fuzzy descriptor model in NCS systems design. However, to the best of our knowledge, the fuzzy descriptor NCS with time-delay has not yet been fully investigated.

Since 2001, both delay-dependent [11][12] and delay-independent [13][14] stability conditions for singular time-delay systems have been derived by using the time domain method. Generally speaking, delay dependent conditions are less conservative than the delay-independent ones, especially when the size of delay is small. Moreover, in engineering practice, information on the delay range is generally available.

To obtain delay-dependent conditions, many efforts have been made in the literature, among which the model transformation technique and bounding technique on cross product terms are often used. While the applying of inequality on bounding some cross terms still led to relative conservatism. The delay-dependent conditions in [12] was derived without using model transformation and bounding technique, while the authors only proposed the criterion for E-exponential stability. Delay-dependent robust stability criteria for time-delay systems with norm-bounded is investigated in [15]. On the other hand, considerable attention has been paid to the problems of robust stabilization and H_∞ control for time-delays systems [16-19]. For continuous descriptor time-delay systems, some results on the problem of delay-dependent robust stability/stabilization are given [20-23], the delay-dependent robust H_∞ control problem for descriptor time-delay system is discussed [24-25].

In this article, we mainly study the problem of robust stability of T-S fuzzy descriptor NCS with multiple time delays in state. First, an improved delay-dependent stability criterion for the T-S fuzzy descriptor NCS with multiple time delays is established in terms of LMIs without using any model transformation and bounding technique. Then based on this criterion, we are able to obtain maximum allowable delay bound (MADB) of NCS. Numerical example is given to show the effectiveness of the presented results.

The paper is organized as follows. Section II formulates the problem and gives a continuous model for fuzzy descriptor NCS with multiple time delay. Some preliminaries also included. The main results are presented in Section III. An example is included to demonstrate the power of this method in Section IV. Finally, Section V concludes this paper.

Notations. Throughout this paper, $\boldsymbol{R}^n$ denotes the n-dimensional Euclidean space, whereas $\boldsymbol{R}^{n\times m}$ refers to the set of all $n\times m$ real matrices. For real symmetric matrix X, the notation $X \geq 0$ ($X > 0$) means that the matrix X is positive-semidefinite (positive-definite). The superscript

T represents the transpose. The symbol * will be used in some matrix expressions to induce a symmetric structure.

II. PROBLEM FORMULATION AND PRELIMINARIES

Consider a model of fuzzy descriptor NCS with multiple state time-delays as

Rule i: If $\xi_1(t)$ is M_{i1} and $\cdots$ and $\xi_p(t)$ is M_{ip}, Then

$$Ex(t) = A_i x(t) + \sum_{i=1}^{N} A_{dik} x(t - \tau_k) + B_i u(t) + B_{wi} w(t)$$

$$z(t) = C_i x(t) + D_i u(t)$$

$$x(t) = \phi(t), \quad t \in [-\bar{\tau}, 0] \qquad i=1,2,\ldots,s \qquad (1)$$

where $\xi_1(t) \sim \xi_n(t)$ are the premise variables, which are assumed to be independent of the input variable, M_{ik} ($k=1,2,\cdots,p$) is the fuzzy set and r is the number of *if-then* rules. $x \in \boldsymbol{R}^n$ is the state, $u \in \boldsymbol{R}^r$ is the control input, $w \in \boldsymbol{R}^p$ is the disturbance input which belongs to $L_2[0,\infty)$, $z \in \boldsymbol{R}^m$ is the control output, τ_k ($k=1,2,\cdots,N$) are constant bounded delays in the state satisfying $0 < \tau_k < \bar{\tau}$. $E \in \boldsymbol{R}^{n\times n}$ is singular and assume that $0 < rank E = q < n$. A_i, A_{dik}, and C_i are real constant matrices of appropriate dimensions.

The system disturbance is assumed to belong to $L_2[0,\infty)$, that is

$$\int_0^\infty w^T(t)w(t)dt < \infty$$

This implies that the disturbance has finite energy.

In view of multi-input and multi-output NCS with many independent sensors and actuators, [15] described the distributed time delays, such as the delays from sensor to controller and the delays from controller to actuator and controller process delay, as multiple state time-delays.

By using center-average defuzzier, product inference and singleton fuzzifier, the dynamic system (1) can be expressed by the following global model:

$$E\dot{x}(t) = A(t)\boldsymbol{x}(t) + \sum_{k=1}^{N} A_{dk}(t)x(t - \tau_k) + B(t)u(t) + B_w(t)w(t)$$

$$z(t) = C(t)x(t) + D(t)u(t) \qquad (4)$$

where

$$A(t) = \sum_{i=1}^{s} h_i(\xi(t))A_i$$

$$A_{dk}(t) = \sum_{i=1}^{s} h_i(\xi(t))A_{dik}$$

$$B(t) = \sum_{i=1}^{s} h_i(\xi(t))B_i$$

$$B_w(t) = \sum_{i=1}^{s} h_i(\xi(t))B_{wi}$$

$$C(t) = \sum_{i=1}^{s} h_i(\xi(t))C_i$$

$$D(t) = \sum_{i=1}^{s} h_i(\xi(t)) D_i$$

$h_i(\xi(t))$ denotes the normalized membership function which satisfies

$$h_i(\xi(t)) = w_i(\xi_i(t)) / \sum_{i=1}^{s} w_i(\xi_i(t))$$

$$w_i(\xi(t)) = \prod_{l=1}^{p} M_{il}(\xi_l(t))$$

$M_{il}(\xi_l(t))$ is the grade of membership of $\xi_l(t)$ in the fuzzy set M_{il}, then

$$h_i(\xi(t)) \geq 0, \quad \sum_{i=1}^{s} h_i(\xi(t)) = 1$$

Consider the nominal unforced descriptor system with w(t)=0 described by

$$E\dot{x}(t) = Ax(t) + \sum_{k=1}^{N} A_{dk} x(t - \tau_k) \tag{5}$$

To facilitate the discussion, we introduce the following definitions.

Definition 1. [26]

1. The pair $(\boldsymbol{E}, \boldsymbol{A})$ is said to be regular if det(s$\boldsymbol{E}$-$\boldsymbol{A}$) is not identically zero.

2. The pair $(\boldsymbol{E}, \boldsymbol{A})$ is said to be impulse-free if deg(det(s$\boldsymbol{E}$-$\boldsymbol{A}$)) = rank $\boldsymbol{E}$.

Definition 2. [26]

For a given scalar $\bar{\tau} > 0$, the nominal descriptor delay system (5) is said to be regular and impulse-free for all constant time delays τ_k satisfying $0 < \tau_k < \bar{\tau}$, if the pair $(\boldsymbol{E}, \boldsymbol{A})$ is regular and impulse-free.

Definition 3.

The uncertain fuzzy descriptor NCS with multiple state time delays system (1) is said to be robustly stable, if the system with $\boldsymbol{u}(t)$=0 and $\boldsymbol{w}(t)$=0 is regular, impulse-free and stable for all admissible uncertainties.

Definition 4.

For a given scalar The uncertain fuzzy descriptor NCS with multiple state time delays system (1) is said to be robustly stable, if the

Definition 4.

For given scalars $\bar{\tau} > 0$ and $\gamma > 0$, the uncertain fuzzy descriptor NCS with time delay (1) is said to be robustly stable with H$_\infty$ performance . if it is robustly stable in the sense of Definition 3 and under zero initial condition, $\|z\|_2 < \gamma \|w\|_2$ for any non-zero $\omega(t) \in L_2[0,\infty)$.

Without loss of generality, we can assume the matrices in (2) have the forms

$$E = \begin{bmatrix} I_q & 0 \\ 0 & 0 \end{bmatrix}, \qquad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

$$A_{dk} = \begin{bmatrix} A_{dk}^{11} & A_{dk}^{12} \\ A_{dk}^{21} & A_{dk}^{22} \end{bmatrix}, \quad B_w = \begin{bmatrix} B_{w1} \\ B_{w2} \end{bmatrix}$$

and denote

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

with $\boldsymbol{A}_{11} \in \boldsymbol{R}^{q \times q}$, $\boldsymbol{A}_{dk}^{11} \in \boldsymbol{R}^{q \times q}$ and $\boldsymbol{x}_1 \in \boldsymbol{R}^q$

Considering the network action, our main purpose is to design a state feedback controller

$$u(t) = \sum_{j=1}^{s} h_j(\xi(t)) K_j x(t)$$
$$= K(t)x(t) \tag{6}$$

where $\qquad K(t) = \sum_{j=1}^{s} h_j(\xi(t)) K_j$.

For simplicity, we will use *A, A_{dk}, B, B_w, C, D* and *K* instead of $A(t), A_{dk}(t), B(t), B_w(t), C(t), D(t)$ in (2) and $K(t)$ in (6).

Such that the following resultant closed-loop system

$$E\dot{x}(t) = (A + BK)\boldsymbol{x}(t) + \sum_{k=1}^{N} A_{dk} x(t - \tau_k) + B_w w(t)$$

$$z(t) = (C + DK)x(t) \tag{7}$$

is regular, impulse free and stable with H$_\infty$ performance for any constant time delays τ_k satisfying $0 < \tau_k < \bar{\tau}$.

III. MAIN RESULT

In this section we will give a solution to the delay-dependent robust H$_\infty$ control problem formulated previously by using LMI technique. Initially, we present the following theorem for the nominal unforced system

$$E\dot{x}(t) = Ax(t) + \sum_{k=1}^{N} A_{dk} x(t - \tau_k) + B_w w(t)$$
$$z(t) = Cx(t) \tag{8}$$
$$x(t) = \phi(t), \quad t \in [-\bar{\tau}, 0]$$

to be regular, impulse free and stable, which will play a key role in solving the aforementioned problem

Theorem 1: For a prescribed scalar $\bar{\tau} > 0$, if there exist matrices

$$P = \begin{bmatrix} P_{11} & 0 \\ P_{21} & P_{22} \end{bmatrix}, \quad P_{11} > 0, \quad Q_k > 0, \quad Z_k = \begin{bmatrix} Z_k^{11} & Z_k^{12} \\ * & Z_k^{22} \end{bmatrix}$$

$$Y_k = \begin{bmatrix} Y_k^{11} & 0 \\ Y_k^{21} & 0 \end{bmatrix}, \quad Y_{k1} = \begin{bmatrix} Y_k^{11} \\ Y_k^{21} \end{bmatrix}, \quad W_k = \begin{bmatrix} W_k^{11} & 0 \\ W_k^{21} & 0 \end{bmatrix}, \quad W_{k1} = \begin{bmatrix} W_k^{11} \\ W_k^{21} \end{bmatrix}$$

with appropriate dimensions such that the following LMIs hold

$$E^T P = P^T E \geq 0 \tag{9}$$

$$P^T A + A^T P < 0 \tag{10}$$

$$\Gamma^k = \begin{bmatrix} \Gamma_{k1} + \bar{\tau} A^T Z_k A & \Gamma_{k2} + \bar{\tau} A^T Z_k A_d & \Gamma_{k4} + \bar{\tau} A^T Z_k B_w & -\bar{\tau} Y_{k1} \\ * & \Gamma_{k3} + \bar{\tau} A_d^T Z_k A_d & \bar{\tau} A_d^T Z_k B_w & -\bar{\tau} W_{k1} \\ * & * & -\gamma^2 I + \bar{\tau} B_w^T Z_k A_d & 0 \\ * & * & * & -\bar{\tau} Z_k^{11} \end{bmatrix} < 0$$

$$k=1,2,\cdots,N \quad (11)$$

where

$$\Gamma_{k1} = Y_k + Y_k^T + C^T C + Q_k,$$

$$\Gamma_{k2} = P^T A_{dk} - Y_k + W_k^T$$

$$\Gamma_{k3} = -Q_k - W_k - W_k^T$$

$$\Gamma_{k4} = P^T B_w$$

$$A_d = \sum_{k=1}^{N} A_{dk}$$

Then the fuzzy descriptor NCS system (6) is regular, impulse free and asymptotically stable for any constant time delays τ_k satisfying $0 < \tau_k < \bar{\tau}$.

Proof ： From (10), it is easy to see that $A_{22}^T P_{22} + P_{22}^T A_{22} < 0$, this implies that A_{22} is nonsingular, and thus the pair (E, A) is regular and impulse free. Therefore, system (8) is regular and impulse free.

Construct Lyapunov-Krasovskii functional for system (8) as

$$V(x_t) = V_1(x_t) + V_2(x_t) + V_3(x_t)$$

where

$$V_1(x_t) = x^T(t)E^T P x(t)$$

$$V_2(x_t) = \sum_{k=1}^{N} \int_{t-\tau_k}^{t} x^T(\alpha) Q_k x(\alpha) \mathrm{d}\alpha$$

$$V_3(x_t) = \sum_{k=1}^{N} \int_{-\tau_k}^{0} \int_{t+\beta}^{t} \dot{x}^T(\alpha) E^T Z_k E \dot{x}(\alpha) \mathrm{d}\alpha \mathrm{d}\beta$$

where $E^T P = P^T E \geq 0$, Z_k, Q_k are symmetric positive-definite matrices to be determined.

Taking the derivative of $V(x_t)$ along with the solution of (8), we have

$$\dot{V}_1(x_t) = \dot{x}^T(t)E^T P x(t) + x^T(t)E^T P \dot{x}(t)$$

$$= 2x^T(t)P^T E\dot{x}(t)$$

$$= 2x^T(t)P^T \left(\begin{bmatrix} A_{11} + \Omega_{11} \\ A_{21} + \Omega_{21} \end{bmatrix} x_1(t) + \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} x_2(t) - S_1 + S_2 \right.$$

$$\left. + B_w w(t) \right)$$

$$= 2x^T(t)P^T \left(\begin{bmatrix} A_{11} + \Omega_{11} \\ A_{21} + \Omega_{21} \end{bmatrix} x_1(t) + \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix} x_2(t) \right)$$

$$+ 2x^T(t)\left(S_3 - P^T S_1\right) + 2S_4 - \left(2x^T(t)S_3 + 2S_4\right)$$

$$+ 2x^T(t)P^T S_2 + 2x(t)^T P^T B_w w(t)$$

$$= 2x^T(t)P^T A x(t) + \sum_{k=1}^{N} \frac{1}{\tau_k} \int_{t-\tau_k}^{t} \{2x^T(t)Y_k x(t)$$

$$+ 2x^T(t)(P^T A_{dk} - Y_k + W_k^T)x(t - \tau_k) - 2x^T(t - \tau_k)W_k x(t - \tau_k)$$

$$- 2x^T(t)\tau_k Y_{k1}\dot{x}_1(\alpha) - 2x^T(t - \tau_k)\tau_k W_{k1}\dot{x}_1(\alpha)$$

$$+ 2x^T(t)P^T B_w w(t)\} \mathrm{d}\alpha$$

$$(12)$$

where

$$\Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{N} A_{dk}^{11} & \sum_{k=1}^{N} A_{dk}^{12} \\ \sum_{k=1}^{N} A_{dk}^{21} & \sum_{k=1}^{N} A_{dk}^{22} \end{bmatrix}$$

$$S_1 = \sum_{k=1}^{N} \int_{t-\tau_k}^{t} \begin{bmatrix} A_{dk}^{11} \\ A_{dk}^{21} \end{bmatrix} \dot{x}_1(\alpha) \mathrm{d}\alpha ,$$

$$S_2 = \sum_{k=1}^{N} \begin{bmatrix} A_{dk}^{12} \\ A_{dk}^{22} \end{bmatrix} x_2(t - \tau_k)$$

$$S_3 = \sum_{k=1}^{N} \int_{t-\tau_k}^{t} \begin{bmatrix} Y_k^{11} \\ Y_k^{21} \end{bmatrix} \dot{x}_1(\alpha) \mathrm{d}\alpha$$

$$S_4 = \sum_{k=1}^{N} x^T(t - \tau_k) \int_{t-\tau_k}^{t} \begin{bmatrix} W_k^{11} \\ W_k^{21} \end{bmatrix} \dot{x}_1(\alpha) \mathrm{d}\alpha$$

$$\dot{V}_2(x_t) = \sum_{k=1}^{N} \left(x^T(t)Q_k x(t) - x^T(t - \tau_k)Q_k x(t - \tau_k) \right)$$

$$= \sum_{k=1}^{N} \frac{1}{\tau_k} \int_{t-\tau_k}^{t} \left(x^T(t)Q_k x(t) - x^T(t - \tau_k)Q_k x(t - \tau_k) \right) \mathrm{d}\alpha$$

$$(13)$$

$$\dot{V}_3(x_t) = \sum_{k=1}^{N} \left\{ \tau_k \left(\dot{x}^T(t)E^T Z_k E\dot{x}(t) \right) - \int_{t-\tau_k}^{t} \dot{x}^T(t)E^T Z_k E\dot{x}(t) \mathrm{d}\alpha \right\}$$

$$= \sum_{k=1}^{N} \frac{1}{\tau_k} \int_{t-\tau_k}^{t} \left\{ \tau_k \left(Ax(t) + \sum_{k=1}^{N} A_{dk}x(t - \tau_k) + B_w w(t) \right)^T \right.$$

$$\times Z_k \left(Ax(t) + \sum_{k=1}^{N} A_{dk}x(t - \tau_k) + B_w w(t) \right)$$

$$\left. - \dot{x}^T(\alpha)E^T Z_k E\dot{x}(\alpha) \right\} \mathrm{d}\alpha \qquad (14)$$

Noting that

$$z^T(t)z(t) - \gamma^2 w^T(t)w(t)$$

$$= x^T(t)C^T C x(t) - \gamma^2 w^T(t)w(t)$$

Then we have

$$\dot{V}(x_t) + z^T(t)z(t) - \gamma^2 w^T(t)w(t)$$

$$= \dot{V}_1(x_t) + \dot{V}_2(x_t) + \dot{V}_3(x_t) + x^T(t)C^T C x(t)$$

$$- \gamma^2 w^T(t)w(t)$$

$$= x^T(t)\Delta x(t) + \sum_{k=1}^{N} \frac{1}{\tau_k} \int_{t-\tau_k}^{t} \zeta^T(t,\alpha)\Psi^k \zeta(t,\alpha)\mathrm{d}\alpha$$

where

$$\Delta = P^T A + A^T P$$

$$\zeta(t,\alpha) = \left[x^T(t), x^T(t - \tau_k), w^T(t), \dot{x}_1^T(\alpha) \right]^T, k=1, \cdots, N$$

$$\Psi^k = \begin{bmatrix} \Psi_{11}^k & \Psi_{12}^k & \Psi_{13}^k & \Psi_{14}^k \\ * & \Psi_{22}^k & \Psi_{23}^k & \Psi_{24}^k \\ * & * & \Psi_{33}^k & \Psi_{34}^k \\ * & * & * & \Psi_{44}^k \end{bmatrix} \qquad (15)$$

$$\Psi_{11}^k = \Gamma_{k1} + \tau_k A^T Z_k A$$

$$\Psi_{12}^k = \Gamma_{k2} + \tau_k A^T Z_k A_{dk}$$

$$\Psi_{13}^k = \Gamma_{k4} + \tau_k A^T Z_k B_w$$

$$\Psi_{14}^k = -\tau_k Y_{k1}$$

$$\Psi_{22}^k = \Gamma_{k3} + \tau_k A_{dk}^T Z_k A_{dk}$$

$$\Psi_{23}^k = \tau_k A_{dk}^T Z_k B_w$$

$$\Psi_{24}^k = -\tau_k W_{k1}$$
$$\Psi_{33}^k = -\gamma^2 I + \tau_k B_w^T Z_k B_w$$
$$\Psi_{34}^k = 0$$
$$\Psi_{44}^k = -\tau_k Z_k^{11}$$

Therefore, in the light of Lyapunov-Krasovskii stability theorem, if $\Delta < 0$ and $\Psi^k \le 0$, then $\dot{V}(x_t) < 0$ for any $\zeta(t, \alpha) \ne 0$, thus the nominal unforced descriptor system (6) is asymptotically stable. Next, We consider the following index

$$J_{zw} = \int_0^\infty \left[z(t)z(t) - \gamma^2 w^T(t)w(t) \right] dt \qquad (16)$$

Noting that we have

$$J_{zw} = \int_0^\infty \left[z(t)z(t) - \gamma^2 w^T(t)w(t) + \dot{V}(x_t) \right] dt - \int_0^\infty \dot{V}(x_t) dt$$
$$= \int_0^\infty \left[z(t)z(t) - \gamma^2 w^T(t)w(t) + \dot{V}(x_t) \right] dt - \lim_{t \to \infty} V(x_t) + V(x_0)$$
$$= \int_0^\infty \left(x(t)\Delta x(t) + \sum_{k=1}^k \frac{1}{\tau_k} \int_{t-\tau_k}^t \zeta^T(t,\alpha) \Psi^k \zeta(t,\alpha) \right) d\alpha$$
$$- \lim_{t \to \infty} V(x_t) + V(x_0) \qquad (17)$$

Applying Schur complement argument to (15), it follows from (11) and Schur complement argument that for any τ_k satisfying $0 < \tau_k < \bar{\tau}$

$$\begin{bmatrix} \Gamma_{k1} & \Gamma_{k2} & \Gamma_{k4} \\ * & \Gamma_{k3} & 0 \\ * & * & -\gamma^2 I \end{bmatrix} + \tau_k \begin{bmatrix} A^T & -Y_{k1} \\ A_{dk}^T & -W_{k1} \\ B_w^T & 0 \end{bmatrix} \begin{bmatrix} Z_k & 0 \\ 0 & (Z_k^{11})^{-1} \end{bmatrix}$$
$$\times \begin{bmatrix} A & A_d & B_w \\ -Y_{k1}^T & -W_{k1}^T & 0 \end{bmatrix}$$
$$\le \begin{bmatrix} \Gamma_{k1} & \Gamma_{k2} & \Gamma_{k4} \\ * & \Gamma_{k3} & 0 \\ * & * & -\gamma^2 I \end{bmatrix} + \bar{\tau} \begin{bmatrix} A^T & -Y_{k1} \\ A_{dk}^T & -W_{k1} \\ B_w^T & 0 \end{bmatrix} \begin{bmatrix} Z_k & 0 \\ 0 & (Z_k^{11})^{-1} \end{bmatrix}$$
$$\times \begin{bmatrix} A & A_d & B_w \\ -Y_{k1}^T & -W_{k1}^T & 0 \end{bmatrix}$$
$$< 0$$

Since $\Delta < 0$, $\Psi^k < 0$, $V(x_0) = 0$ under zero initial condition and $\lim_{t \to \infty} V(x_t) \ge 0$, then the following LMI holds:

$$J_{zw} < 0$$

This gives the desired results

$$\|z\|_2 < \gamma^2 \|w\|_2$$

which implies that system (6) has H$_\infty$ performance γ.
So for any $0 < \tau_k < \bar{\tau}$, we have $\Psi^k < 0$. This completes the proof. □

Remark 1: Theorem 1 provides the delay-dependent asymptotic stability criteria for nominal NCS (6). Based on the theorem 1, we can obtain a less conservative MADB $\bar{\tau}$ such that NCS are asymptotically stable by solving the following optimization problems,

$$\begin{cases} \min 1/\bar{\tau} \\ \text{s.t.} P, Q_k, Z_k, Y_k, W_k, (9), (10), (11) \end{cases} \qquad (18)$$

Equation (18) is a convex optimization problem and can be obtained efficiently using the MATLAB LMI Toolbox.

Now, we concentrate on the design of a state feedback controller of the form (6) which guarantees the resultant closed-loop system (7) to be regular. Based on Theorem 1, we obtain the following theorem.

Theorem 2. For prescribed scalars $\bar{\tau} > 0$ and $\gamma > 0$, there exists a state feedback controller (6) such that the closed-loop fuzzy descriptor system (7) is regular, impulse-free and stable with H$_\infty$ performance γ for any constant time delays τ_k satisfying $0 < \tau_k \le \bar{\tau}$ if there exist nonsingular matrix P, a

$$E^T P = P^T E \ge 0 \qquad (19)$$
$$A\tilde{P} + \tilde{P}^T A^T + BN + N^T B^T < 0 \qquad (20)$$

$$\Xi^k = \begin{bmatrix} \tilde{\Xi}_{k1} & \tilde{\Xi}_{k2} & \tilde{\Xi}_{k4} & 0 & \bar{\tau}\tilde{P}^T A^T \\ * & \tilde{\Xi}_{k3} & 0 & 0 & \bar{\tau}\tilde{P}^T A_d^T \\ * & * & -\gamma^2 I & 0 & \bar{\tau}N^T B^T \\ * & * & * & -\bar{\tau}Z_k^{11} & \bar{\tau}B_w\tilde{P} \\ * & * & * & * & -Z_k \end{bmatrix} < 0$$
$$k=1,2,\cdots,N \qquad (21)$$

where

$$\tilde{\Xi}_{k1} = \tilde{Y}_k + \tilde{Y}_k^T + \tilde{P}^T C^T C\tilde{P} + Q_k + \tilde{P}^T C^T DN\tilde{P} + N^T D^T C$$
$$\tilde{\Xi}_{k2} = A_{dk} - \tilde{Y}_k + \tilde{W}_k^T$$
$$\tilde{\Xi}_{k3} = -\tilde{Q}_k - \tilde{W}_k - \tilde{W}_k^T$$
$$\tilde{\Xi}_{k4} = B_w\tilde{P}$$

In this case, a desired state feedback controller is given by

$$u(t) = N\tilde{P}^{-1}x(t) \qquad (22)$$

Proof

Using inequality (10), we have

$$P^T(A + BK) + (A + BK)^T P < 0$$

Then pre-, post-multiplying both sides with P^{-T}, P^{-1}, and defining $\tilde{P} = P^{-1}$, $N = K\tilde{P}$, we can obtain (20).

Pre-, post-multiplying both sides of inequality (11) with diag$\{ P^{-T}, P^{-T}, P^{-T}, I \}$ and its transpose, and defining

$$\tilde{P} = P^{-1}$$
$$N = K\tilde{P}$$
$$\tilde{Y}_k = \tilde{P}^T Y_k \tilde{P}$$
$$\tilde{W}_k = \tilde{P}^T W_k \tilde{P}$$

we have

$$\Xi^k = diag\{P^{-T}, P^{-T}, P^{-T}, I\} \Gamma^k diag\{P^{-1}, P^{-1}, P^{-1}, I\}$$

$$= \begin{bmatrix} \Xi_{k1} & \Xi_{k2} & & \Xi_{k4} \\ * & \Xi_{k3} & & \bar{\tau}\widetilde{P}^T A_d^T Z_k B_w \widetilde{P} \\ * & * & & -\gamma^2 I + \bar{\tau}\widetilde{P}^T B_w^T Z_k A_d \widetilde{P} \\ * & * & & * \end{bmatrix}$$

$$\begin{bmatrix} -\bar{\tau}\widetilde{Y}_k \\ -\bar{\tau}\widetilde{W}_k \\ 0 \\ -\bar{\tau}\widetilde{Z}_k \end{bmatrix} < 0$$

where

$$\Xi_{k1} = \widetilde{Y}_k + \widetilde{Y}_k^T + \widetilde{P}^T C^T C \widetilde{P} + Q_k + \widetilde{P}^T C^T D N \widetilde{P}$$
$$+ N^T D^T C + \bar{\tau}\left(\widetilde{P} A Z_k A \widetilde{P} + \widetilde{P}^T A^T Z_k B N \right.$$
$$\left. + N^T B^T Z_k A \widetilde{P} + N^T B^T Z_k \widetilde{P}\right)$$

$$\Xi_{k2} = A_{dk} - \widetilde{Y}_k + \widetilde{W}_k^T + \bar{\tau}\left(\widetilde{P}^T A^T Z_k A_d \widetilde{P} + N^T B^T Z_k A_d \widetilde{P}\right)$$
$$\Xi_{k3} = -\widetilde{Q}_k - \widetilde{W}_k - \widetilde{W}_k^T + \bar{\tau}\widetilde{P}^T A_d^T Z_k A_d \widetilde{P}$$
$$\Xi_{k4} = B_w \widetilde{P} + \bar{\tau}\left(\widetilde{P}^T A^T Z_k B_w \widetilde{P} + N^T B^T Z_k B_w \widetilde{P}\right)$$

By Schur complement, we can obtain $\Xi^k < 0$ is the same as (21).

Therefore, we can see the closed-loop fuzzy descriptor system (7) is regular, impulse free and stable with H_∞ performance γ for all constant time delays τ_k satisfying $0 < \tau_k \le \bar{\tau}$, and a desired controller can be obtained by $u(t) = Kx(t)$ with $K = N\widetilde{P}^{-1}$. The proof is completed.

□

Remark 2: In the proof of these heorems , it is noted that neither model transformation nor bounding technique for cross terms, which are usually used in the existing results, is required. Hence, the derivation procedure is simpler and these theorems are less conservative than the existing ones.

Remark 3: If $E=I$, Theorems in this paper can be regarded as an extension of standard state-space fuzzy systems with multiple time-delay.

V. ILLUSTRATIVE EXAMPLES

In order to demonstrate the validity of the proposed methods, a numerical example is involved to illustrate the effectiveness of the proposed criteria. A T-S fuzzy descriptor system with multiple time delay is described as follows. It is supposed that x_1 is measurable online.

If x_1 is P, then

$$E\dot{x}(t) = A_1 x(t) + \sum_{k=1}^{2} A_{d1k} x(t - \tau_k)$$
$$+ B_1 u(t) + B_{w1} w(t)$$

If x_1 is N, then

$$E\dot{x}(t) = A_2 x(t) + \sum_{k=1}^{2} A_{d2k} x(t - \tau_k)$$
$$+ B_2 u(t) + B_{w2} w(t) \qquad (23)$$

Here the membership functions of P and N are given as follows:

$$h_1(x_1) = 1 - \frac{1}{1 + e^{-2x_1}}$$

$$h_2(x_1) = \frac{1}{1 + e^{-2x_1}}$$

$$E = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad A_1 = \begin{bmatrix} -1.1 & 0.3 \\ 0 & 0.4 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} -0.4 & 0.1 \\ 0.4 & -0.8 \end{bmatrix}, \quad A_{d11} = \begin{bmatrix} 0.1 & -0.2 \\ 0 & 0 \end{bmatrix}$$

$$A_{d12} = \begin{bmatrix} 0.2 & -0.1 \\ 0 & 0 \end{bmatrix}, \quad A_{d21} = \begin{bmatrix} 0.2 & 0 \\ 0.1 & 0.3 \end{bmatrix}$$

$$A_{d22} = \begin{bmatrix} 0.1 & 0 \\ 0.2 & 0.1 \end{bmatrix}$$

H_∞ performance $\gamma = 2$. The disturbance input $w(t)$ is as follows:

$$w(t) = \begin{cases} 1 - t, & 0 < t \le 1s \\ 0, & \text{otherwise} \end{cases}$$
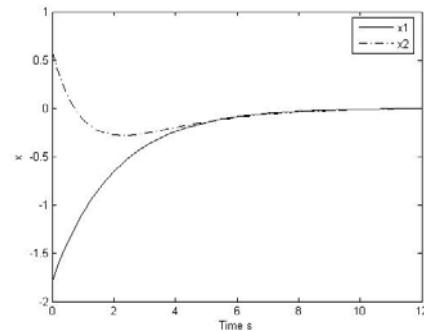

Figure 2. Simulation results of x_1 and x_2.

According to Theorem 2 and using MATLAB LMI Toolbox, it is found that the maxium bound of time-delay is $\bar{\tau} = 0.7832$. For example, when $\tau_1 = 0.73$, $\tau_2 = 0.61$, the corresponding state feedback controller is

$$u(t) = \begin{bmatrix} -24.6357 & -31.7942 \end{bmatrix} x(t)$$

Fig. 1 gives the simulation results of x_1 and x_2 when the initial function $\phi(t) = \begin{bmatrix} -1.8 & 0.6 \end{bmatrix}, t \in [-\bar{\tau}, 0]$. From Fig 1, we can see that the states x_1 and x_2 asymptotically converge to zero.

VI. CONCLUSION

In this paper, we firstly model fuzzy descriptor NCS with network induced delays for a class of fuzzy descriptor systems with multiple time delays in state. Based on the model, utilizing the Lyapunov stability theories combined with LMIs techniques, a new stability criteria in terms of LMIs is proposed. The criterion is obtained without using any model transformation and bounding technique. Applying the delay-dependent robust stability criteria proposed here to solve the robust control problem for fuzzy descriptor NCS , such as H∞

control, guaranteed cost control, variable structure control and so on, will be interesting topics for further research.

REFERENCES

[1] G.C. Walsh, O. Beldiman, L.G. Bushnell, Stability analysis of networked control system, Proceedings of American Control Conference, San Diego, pp.2876–2880, 1999.

[2] W. Zhang, M. S. Branicky, S. M. Phillips, Stability of networked control systems, IEEE Control System Magzine, vol. 21, pp.84–99, 2001.

[3] M. S. Branicky, S. M. Phillips, W. Zhang, Stability of networked control systems: Explicit analysis of delay, Proceedings of American Control Conference, Chicago, pp.2352–2357, 2000.

[4] B.G. Marieke, Nathan van de Wouw, Stability of networked control systems with uncertain time-varying delays, IEEE Trans. on Automatic Control, vol.54, pp.1575–1580, 2009.

[5] H. Yan, X. Huang, M. Wang, H. Zhang, Delay-dependent stability criteria for a class of networked control systems with multi-input and multi-output, Chaos, Solitons and Fractals, vol. 34, pp.997–1005, 2007.

[6] T. Takagi, M. Sugeno, Fuzzy identification of systems and its application to modeling and control, IEEE Trans. on Systems, Man and Cybernetics,vol.15, pp.116–132, 1985.

[7] K. Tanaka, M. Sano, A robust stabilization problem of fuzzy control systems and its application to backing up control of a truck-trailer, IEEE Trans. Fuzzy Systems, vol. 2, pp.119–134, 1994.

[8] Z. Hongbin, S. Yanyan, F. Gang, Delay-dependent stability and H∞ control for a class of fuzzy descriptor systems with time-delay, Fuzzy Sets and Systems, vol.160, pp.1689–1707, 2009.

[9] Y. Jun, Z. Shouming, X. Lianglin, A descriptor system approach to non-fragile H∞ control for uncertain fuzzy neutral systems, Fuzzy Sets and Systems, vol. 160, pp.423–438, 2009.

[10] T. Weihua, Z. Huaguang, Optimal guaranteed cost control for fuzzy descriptor systems with time-varying delay, Journal of Systems Engineering and Electronics, vol. 19, pp.584–591, 2008.

[11] E. Fridman, Effects of small delays on stability of singular perturbed systems, Automatica, vol. 38, pp.897–902, 2002.

[12] D. Yue, Q. Han, A delay-dependent stability criterion of neutral systems and its application to partial element equivalent circuit model, IEEE Trans. Circuits System II, vol. 51, pp.685–689, 2004.

[13] S. Xu, P. Van Dooren, R. Stefan, and J. Lam, Robust stability and stabilization for singular systems with state delay and parameter uncertainty, IEEE Trans. Autom. Contro, vol. 47, pp.1122–1128, 2002.

[14] J. Feng, S. Zhu, and Z. Cheng, Guaranteed cost control of linear uncertain singular time-delay systems, Proceedings of 41st IEEE Conference Decision Control, pp.1802–1807, 2002.

[15] Z. Shuqian, Z. Chenghui, C. Zhaolin, and F. June, Delay-dependent robust stability criteria for two classes of uncertain singular time-delay systems, IEEE Trans. on Automatic Control, vol. 52, pp.880–885, 2007.

[16] Q.C. Chen, L. Yu, W.A. Zhang. Delay-dependent output feedback guaranteed cost control for uncertain discrete-time systems with multiple time-varying delays. IET Control Theory and Applications 2007; 1(1):97–103.

[17] Z. Gao, T. Breikin, H. Wang. Reliable observer-based control against sensor failures for systems with time delays in both state and input. IEEE Transactions on Systems, Man and Cybernetics, Part A 2008; 38(5):1018–1029.

[18] H.C., Yan , X.H., Huang, M. Wang, et al. New delay-dependent stability criteria of uncertain linear systems with multiple time-varying state delays. Chaos, Solitons and Fractals 2008, 37(1):157–165.

[19] C.Y., Lu, Tsai JSH, T.J Su. On improved delay-dependent robust stability criteria for uncertain systems with multiple state delays. IEEE Transactions on Circuits and Systems 2002,49(2):253–256.

[20] E.K. Boukas, Delay-dependent stabilization of singular linear systems with delays. International Journal of Innovative Computing, Information and Control 2006, 2(2):283–291.

[21] H.J, Wang, A.K, Xue, R.Q.Lu, Delay-dependent robust stability and stabilization for uncertain singular system with time-varying delay. Proceedings of the 2008 American Control Conference, Seattle, WA, U.S.A., 2008, 3626–3631.

[22] S. Zhu, Z. Li, Z. Cheng. Delay-dependent robust resilient H∞ control for uncertain singular time-delay systems. Proceedings IEEE Conference on Decision and Control, and European Control Conference. Seville: Spain, 2005, 1373–1378.

[23] R. Zhong, Z. Yang, Delay-dependent robust control of descriptor systems with time delay. Asian Journal of Control 2006; 8(1):36–44.

[24] Z. Wu, W. Zhou, Delay-dependent robust H∞ control for uncertain singular time-delay systems. IET Control Theory and Applications 2007; 1(5):1234–1241.

[25] EK., Boukas, Singular linear systems with delay: H∞ stabilization. Optimal Control Applications and Methods 2007;28(4):259–274.

[26] Z. Wu, W. Zhou, Delay-dependent robust H∞ control for uncertain singular time-delay systems, IET Control Theory Appl. vol. 1, pp.1234–1241, 2007.

Baoping Ma was born in Shanxi, China, in 1973. She received the B.Eng. degree in automation and M.S. degree in control theory and control engineering from Taiyuan University of Technology, Taiyuan, in 1994 and 1997, respectively. She received Ph.D. Degree in Thermal Engineering from Southeast University in 2001.

She is currently a teacher of Automation in School of Electrical and Automation Engineering, Nanjing Normal University, Nanjing, China. Her current research interests are mainly in Networked Control System, fuzzy descriptor system and robust control.

Numerical Simulation of Air Flow Properties around High Speed Train in Very Long Tunnel

Yanfeng Zhu

Guangdong University of Technology, Guangzhou, China
Email: zhuyf@gdut.edu.cn

Xiangyun Liu

Guangdong University of Technology, Guangzhou, China
Email: forlxy@163.com

Zhanlin Cui

Harbin Railway Technical l College, Harbin, China
Email: zhanlin-cui@163.com

Abstract—**Numerical simulation of air flow disturbed around a train body with different angle of train nose and different velocity of high speed train in a very long tunnel have been made. Three-dimensional flow theory is capable of modeling the effects of interest in the tunnel environment. The scale of train is 2500mm$\times$ 320mm$\times$ 427mm length$\times$ width$\times$ heigh, and the scale of tunnel is 150m$\times$ 10.4m$\times$ 8.2m length$\times$ width$\times$ heigh. Train speed is 432km/h, 500km/h and 600km/h respectively. The distributions of pressure, velocities streamline and second flow around high speed train in a very long tunnel space can be obtained by simulating. Results show that the train nose angles and train speed affects the air flow around the train strongly. The pressure drag increases with rising of the nose angles and train speed, and it reaches to the highest value of 23073 Pa when the angle is 60^0 and 32466Pa as the velocity is 600km/h. The air velocity in the annular space is quite enormous and the maximize value is 178m/s with train nose 60^0 and 247m/s with train speed 600km/h. The simulation results can be taken to help us for study the high-speed train energy-consuming, maximum velocity, safety and stabilization when it running in a very long tunnel.**

Index Terms—**-high-speed train; nose angle; velocity; pressure drag; velocities streamline; second flow**

I. INTRODUCTION

High-speed rail transportation is quickly becoming a popular form of mass transit throughout the world. With the speed-up of train, many engineering problems which have been neglected at low train speeds are being raised. Especially more and more deep buried long tunnel has appeared in the construction of high-speed railway engineering. When a train travels through a long tunnel, the compressed air will not flow timely and swimmingly along the train side and upside as in the opening because of the air viscosity and friction resistance distribution in the 'smooth' tunnel surface and 'rough' train surface. When the train velocity and angle of train nose changes, aerodynamic phenomena gain in complexity and

importance. As the train speed or train nose angle increased, air pressure will be increased with it; wind which is induced by train will be amplified; aerodynamic drag will be also magnified several times and disturbing of aerodynamic noise will appear in long tunnel.

The first studies of train aerodynamics in tunnels were carried out in 1927 by von Tollmien [1], who searched for an analytical expression of train drag by using a quasi-static incompressible model. In the 1960s and 1970s, a very significant contribution to the subject was made by Japanese researchers [2, 3]. The present study concerns the analysis of aerodynamic phenomena in long tunnels. The approach described consists in a numerical analysis using a quasi one-dimensional finite volume model of air flow and train motion in the tunnel. W.A.Woods[4] investigated a generalized one-dimensional flow prediction method for calculating the flow generated by a train in a single-track tunnel. Raghu S. Raghunathana, H.-D. Kimb, and T. Setoguchi [5] indicated the aerodynamics of high-speed railway train, Pierre Riccoa,b, Arturo Baronb, Paolo Molteni [6] investigated the nature of pressure waves induced by a high-speed train traveling through a tunnel. Yanfeng Zhu and Yapping Wu [7-9] studied the temperature rise by high speed train traveling through a long tunnel with different speed and different blockage ratio, result show that the temperature field in long tunnel is indispensably. Arturo Baron, Michele Mossi and Stefano Sibilla [10] calculated the alleviation of the aerodynamic drag and wave effects of high-speed train in very long tunnels. However, the air flow properties around a train body influence the safety and stability of the train-tunnel system and traction capacity of high-speed train. Comprehensively considered the heating of auxiliary equipment, train equipment, train braking, pantograph electric arc etc and the energy loss of heat conduction induced by tunnel wall, this paper carried simulation to study the air flow properties around high speed train with different train nose angles and different train speed.

Pressure distributing, velocity streamlines and second flow behind train tail have been obtained.

II. GOVERNING EQUATIONS AND NUMERICAL MODEL

A. Governing Equations

The unsteady three-dimensional flow of a viscous, compressible fluid is governed by the Navier-Stokes equations, namely by a set of partial differential equations which express the physical principles of conservation of mass, momentum and energy. The quality force can be assumed to be zero because of the density of air is very small. The governing equations shown as follows:

$$\frac{\partial \boldsymbol{r}}{\partial t} + \nabla \cdot (\boldsymbol{r}V) = 0 \tag{1}$$

$$\frac{\partial(\boldsymbol{r}u)}{\partial t} + \nabla \cdot (\boldsymbol{r}uV) = -\frac{\partial p}{\partial x} + \frac{\partial \boldsymbol{t}_{xx}}{\partial x} + \frac{\partial \boldsymbol{t}_{yx}}{\partial y} + \frac{\partial \boldsymbol{t}_{zx}}{\partial z}$$

$$\frac{\partial(\boldsymbol{r}v)}{\partial t} + \nabla \cdot (\boldsymbol{r}vV) = -\frac{\partial p}{\partial y} + \frac{\partial \boldsymbol{t}_{xy}}{\partial x} + \frac{\partial \boldsymbol{t}_{yy}}{\partial y} + \frac{\partial \boldsymbol{t}_{zy}}{\partial z} \quad ?$$

$$\frac{\partial(\boldsymbol{r}w)}{\partial t} + \nabla \cdot (\boldsymbol{r}wV) = -\frac{\partial p}{\partial z} + \frac{\partial \boldsymbol{t}_{xz}}{\partial x} + \frac{\partial \boldsymbol{t}_{yz}}{\partial y} + \frac{\partial \boldsymbol{t}_{zz}}{\partial z} \tag{2}$$

$$\frac{\partial}{\partial t}\left[\boldsymbol{r}(e+\frac{V^2}{2})\right] + \nabla \cdot \left[\boldsymbol{r}(e+\frac{V^2}{2})V\right]$$

$$= \frac{\partial}{\partial x}(k\frac{\partial T}{\partial x}) + \frac{\partial}{\partial y}(k\frac{\partial T}{\partial y}) + \frac{\partial}{\partial z}(k\frac{\partial T}{\partial z}) \tag{3}$$

$$- \frac{\partial(up)}{\partial x} - \frac{\partial(vp)}{\partial y} - \frac{\partial(wp)}{\partial z}$$

Where $V = u\boldsymbol{i} + v\boldsymbol{j} + w\boldsymbol{k}$ is the mean of air velocity vector coincident with its component u along the tunnel direction, represented by the unit vector i; component v and w along the other normal direction, represented by the unit vector j, k; $?$ is air density, p is pressure, t is time. t is air shear stress, e is the internal energy per unit mass, k is thermal conductivity of air, T is air temperature.

Borrowing in the state equation for ideal gases, the constitutive equations are finally used to define the thermodynamic state and balance the number of unknown quantities:

$$p = \boldsymbol{r}RT \tag{4}$$

$$e = C_V T \tag{5}$$

R is ideal gas constant; C_V is the air specific heat at constant volume.

An unsteady, compressible, turbulent calculation was carried out using a RNG k-$\boldsymbol{e}$ turbulence model employing a second-order upwind discrimination scheme. The k-$\boldsymbol{e}$ equation is

$$\frac{\partial(\boldsymbol{r}k)}{\partial t} + \frac{\partial(\boldsymbol{r}V_j k)}{\partial x_j} = \frac{\partial}{\partial x_j}\left[\left(\boldsymbol{m}+\frac{\boldsymbol{m}_t}{\boldsymbol{s}_k}\right)\frac{\partial k}{\partial x_j}\right] + S_k \tag{6}$$

$$\frac{\partial(\boldsymbol{r}e)}{\partial t} + \frac{\partial(\boldsymbol{r}V_j e)}{\partial x_j} = \frac{\partial}{\partial x_j}\left[(\boldsymbol{m}+\frac{\boldsymbol{m}_t}{\boldsymbol{s}_e})\frac{\partial e}{\partial x_j}\right] + S_e \tag{7}$$

Where

$$S_k = P_k - D_k , \quad P_k = \boldsymbol{t}_{ij}^{(T)}\frac{\partial V_i}{\partial x_j} , \quad D_k = \boldsymbol{r}e$$

$$S_e = P_e - D_e, \quad P_e = C_{e1}\frac{e}{k}\boldsymbol{t}_{ij}^{(T)}\frac{\partial V_i}{\partial x_j} , \quad D_e = C_{e2}\boldsymbol{r}\frac{e^2}{k}$$

and $C_{e1} = 1.44$, $C_{e2} = 1.92$, $\boldsymbol{s}_k = 1.0$, $\boldsymbol{s}_e = 1.3$.

B. Numerical model

In this paper, the train models have conical noses with angles $30°$, $45°$ and $60°$ between the axis and the directrix. The model is shown as in Fig.1. The simulation model of tunnel is L_T=150m long; cross-sectional area A_T=71.9m^2; perimeter of tunnel $?_T$=31.24m. Heat transfer coefficient of tunnel wall with air is 6 wm^{-2}k^{-1} and it with a friction coefficient C_{fT}=0.005, corresponding to a mean wall roughness of 0.5mm. We have studied the cases with blockage ratio $β$=0.23, train speed u_t= 350km/h, 432km/h, 500km/h and 600km/h respectively.
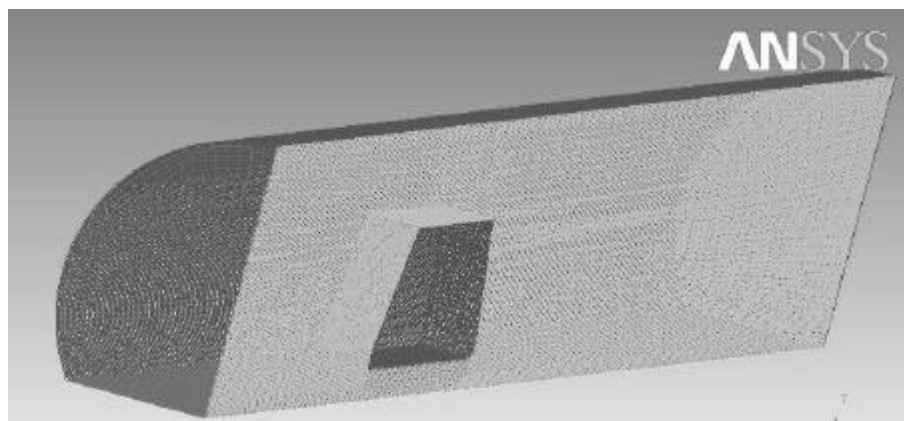

Figure1. Mesh of the tunnel and train

volumes for each main tunnel. Computations have been performed with Ansys-cfx.

III. COMPUTATIONALRESULTS

When the identical train passes through a long tunnel, a number of complications arise in comparison with the open air case. These include:

(a) Significant pressure is generated during high speed train running in long tunnel and total pressure magnitude along train body is varied enormous. it can signi?cantly increase the aerodynamic drag on a train traveling in a tunnel.

(b) Train velocity in annular space is quite tremendous. It would induce biggish skin friction drag.

(c) Coefficient of the flow resistance is quite large.

(d) The second flow behind the train tail.

There are discussed briefly in turn.

A. Pressure variation

When a high speed train runs in a long tunnel, a compression wave is formed ahead of the train which propagates along the tunnel at a nearly sonic speed. The air pushed ahead of the train causes a significant pressure increase in front of the train nose and creates a pressure decrease behind the tail. Fig.2 (a,b,c) presents the distributions of total pressure along the upstream and the downstream with different train nose angles at moving velocity 432km/h inside the tunnel with ß =0.23. We can see that all maximal total pressure appear in train nose, but minimum total pressure are in the difference of occurrence site. When a equal to 30°, on the distance behind train tail 5.5m, the minimum total pressure is 97859Pa and a equal to 45°, at 2.8m behind of the train tail, the minimum total pressure is 93698Pa. But when a equal to 60°, the minimum total pressure is 92981Pa appearing in the distance from train nose 2.8m. it is because that the separation occurring along the train top. at 1.5m behind of the train tail the total pressure drops greatly again and the minimum total pressure is 96696Pa.

Fig.2 (a,b,c) also shows that the trend of total pressure magnitude along train body is varied from downtrend to rising trend. This phenomenon is also because of separation occurring along the train top. In short, ?p increases with the rise of angle. These differences magnitude are about 13801Pa, 20726Pa and 23073Pa respectively in the long tunnel. Behind of the train tail 2m, total pressure is 108598 Pa,105500 Pa , 100066Pa.

In the same tunnel configuration, despite the similarities, the behavior of air flow around the vehicle is governed by different phenomena. The higher train velocity leads instead to a more pronounced piston effect when train nose angle is 45^{0} (Fig.3 (a,b,c)).

Fig.4, 5 shows the comparison of the total pressure with different train nose angles (432km/h) and different train speed (train nose angle 45^{0}). In the long tunnel, the air total pressure difference is observably increased with train nose angle as train has same speed; and it is increased with train speed as train has same train nose angle. When train nose angle is 45^{0}, the ?p between the

train ends is reduced from about 38362Pa as train speed 600km/h case to 12923 Pa as train speed 350km/h.

B. Velocities streamline

Fig.6 (a,b,c) presents the streamlines in the symmetric plane at train speed of 432km/h with different train nose angle and at different train speed with the same nose angle 45° (Fig.7 (a,b,c)). Around the train, the air flow accelerates along the train nose and the annular space in a quite spectacular way and the maximize value is 178m/s when train nose angle is $30^{\circ}$and 247m/s when train speed 600km/h. The separation occurs along the entire rear body and there is reattachment on the body, and the patterns are very similar. Some differences exist for the position of the reattachments and the intensity of the bottom corner vortex. The vortexes in make the energy loss increases.

C. Coefficient of flow resistance

The coefficient of the flow resistance is defined as follows:

The coefficient of flow resistance:

$$C_D = \frac{F_X}{\frac{1}{2}rv_{\infty}^2 \cdot A} \tag{8}$$

Where A is the maximum area of the train's cross section, F_X is total aerodynamic drag on the train, v_8 is the air speed in front of the train nose.

These differences in magnitude depicted as Fig.8 in the mode of coefficient of flow resistance vs. train nose angles. In general, the coefficient of flow resistance is not a constant. As expected, it increases with increasing of the nose angle. But we also find that train velocity effects are very little for the coefficient of flow resistance (Fig.9).

D. Second flow

In a full analysis of air flow properties around high speed train, besides the velocities streamline, however, it is also necessary to pay attention to the second flow behind the train tail. Because their realization forms are vortexes which lead to the energy loss. Figure 10, 11 presents the second flow at 2.5m and 5m behind of the train tail at moving velocity of 432km/h with different train nose angle. It can be seen that the vortexes are generated, and the intensity of vortexes is not only affected in train nose angle but also space positions.
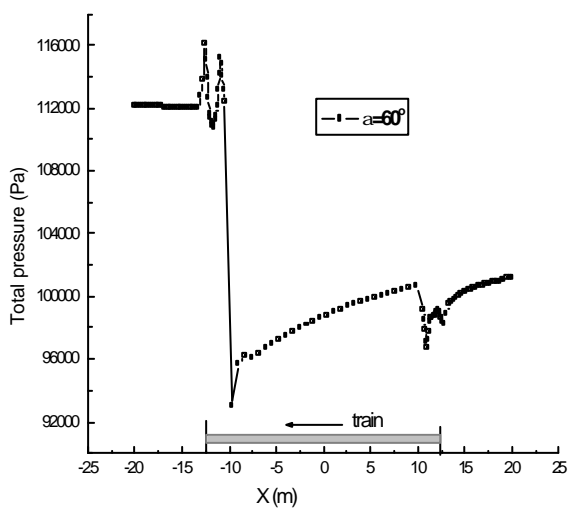
Figs. 12 and 13 present the second flow at different train speed which train nose is 45^{0}. It is found that train speed with less influence on the second flow than train nose angle. But as same as the former, distance effect is obviously.
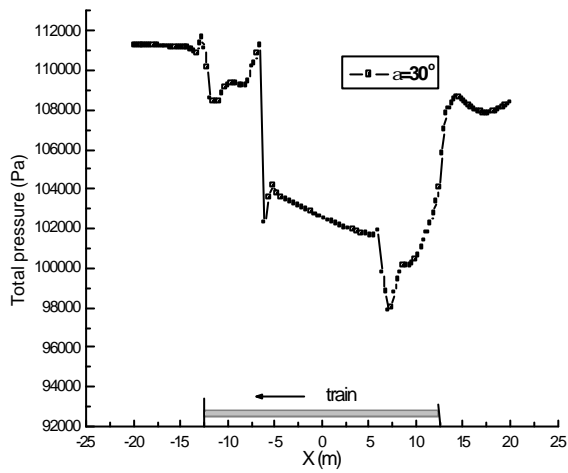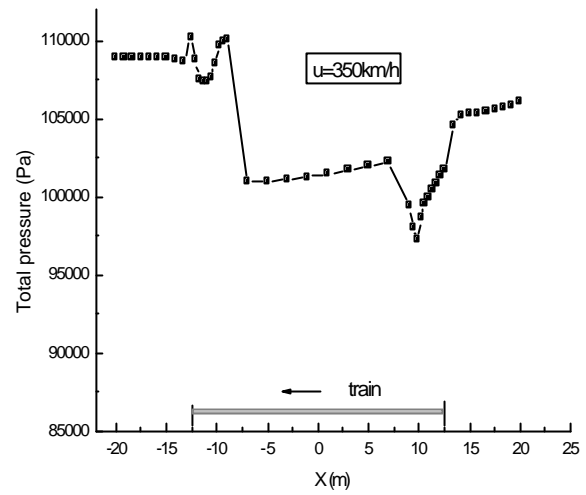
(a) a=30°
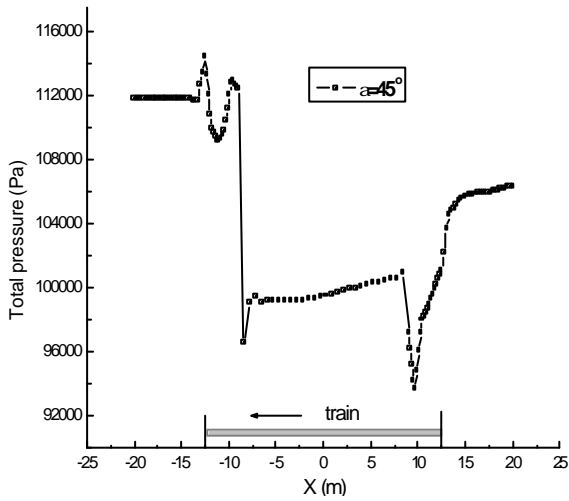
u=350km/h

(b) a=45°
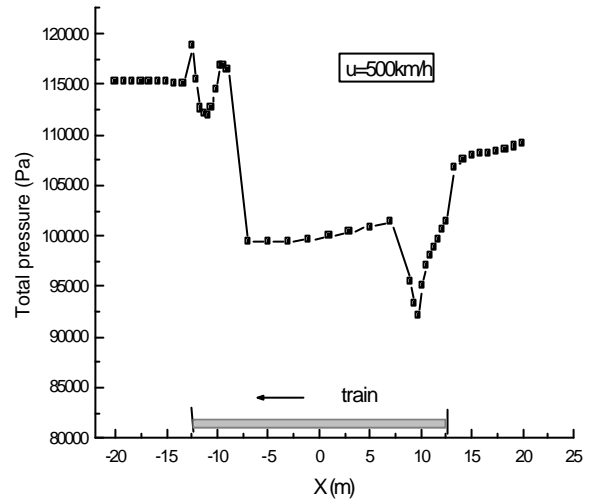

u=500km/h

(c) a=60°

Figure2.   Total pressure around the train
with different train nose angles (ut=432km/h)
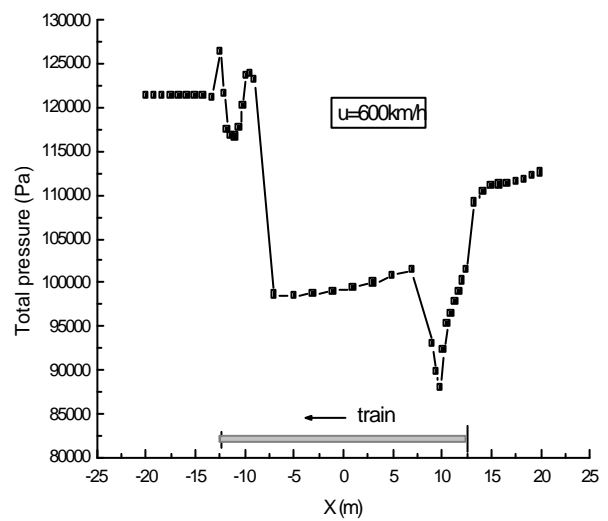
u=600km/h

Figure3.   Total pressure around the train
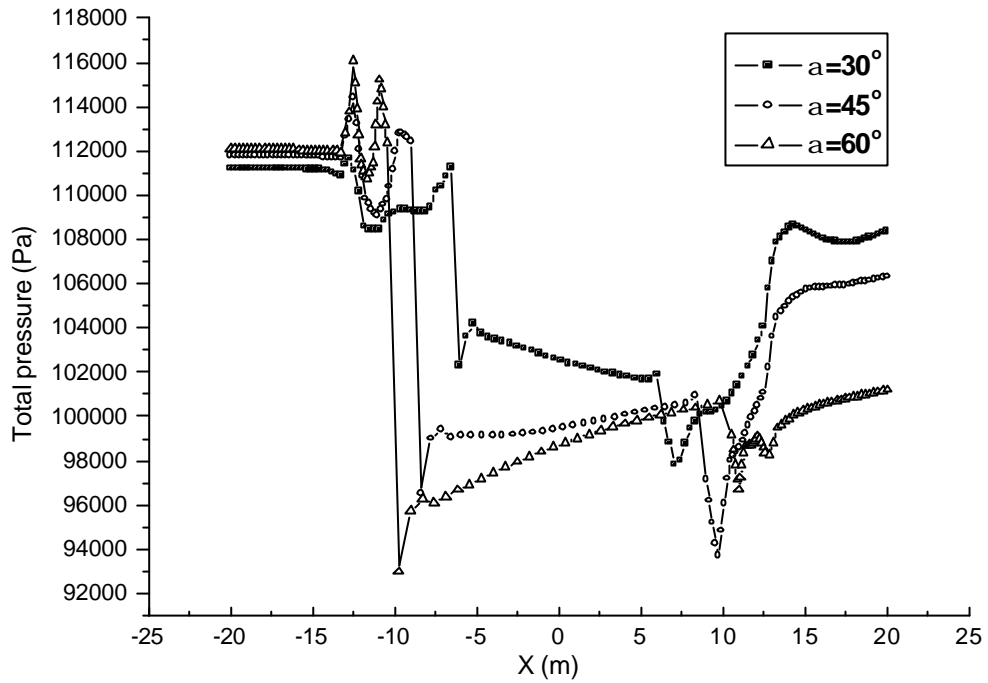with different train speed (a=45°)

Figure4． Comparison of total pressure with different nose angles (432km/h)

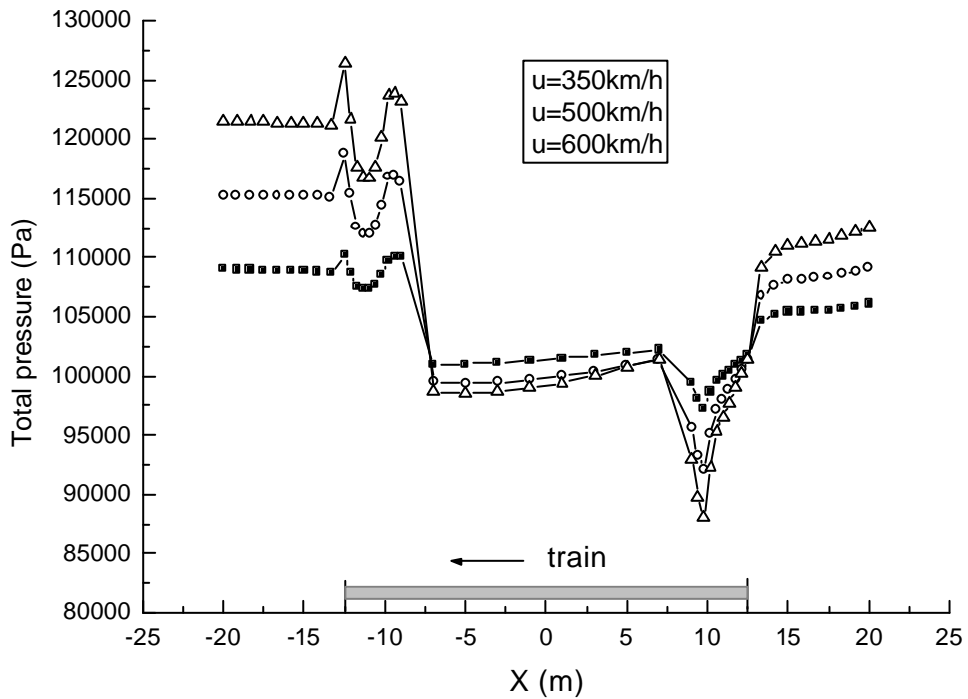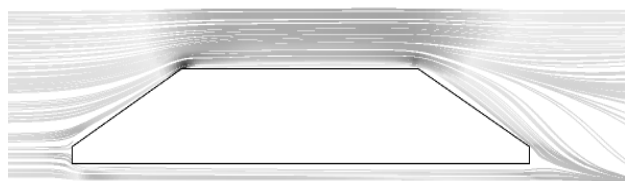Figure5． Comparison of total pressure with different train speed (nose angles 45°)
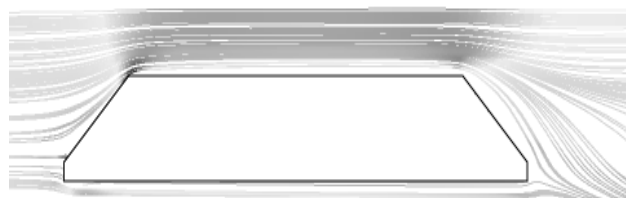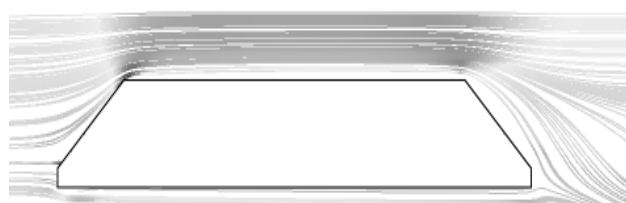
(a) $a=30°$

(b) $a=45°$

(c) $a=60°$

Figure6. Streamline around the high-speed train with different train nose angles (u_t =432km/h)

u=350km/h

u=500km/h

u=600km/h

Figure7. Streamline around the high-speed train with different train speed (a =45°)
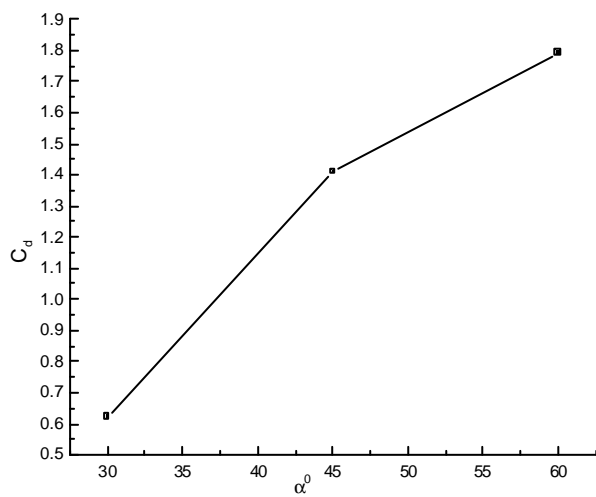
Figure8. Effect of the train nose angle on the coefficient of flow resistance (432km/h)
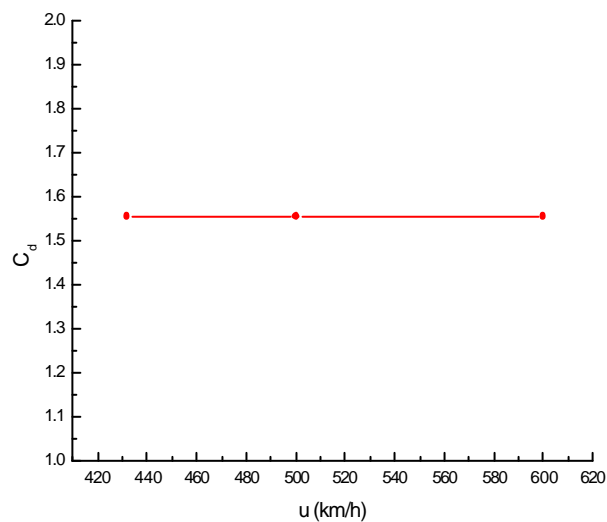
Figure9. Effect of the train speed on the coefficient of flow resistance (a =45°)

(a) $a=30°$ (b)$a=45°$ (c)$a=60°$
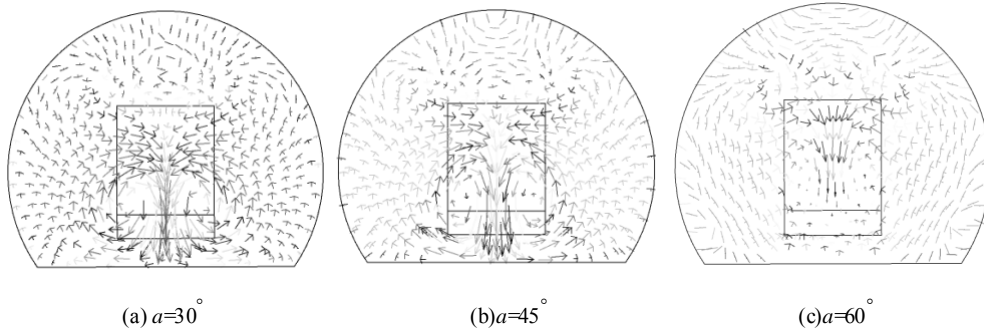
Figure10.   Second flow at 2.5m behind the train with different nose angle (u_t=432km/h)

(a) $a=30°$ (b)$a=45°$ (c)$a=60°$
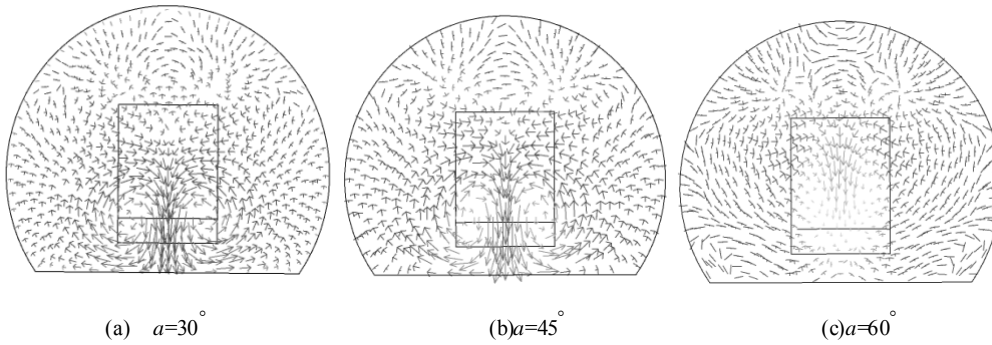
Figure11.   Second flow at 5m behind the train with different nose angle (u_t=432km/h)

(a) u=350km/h   (b) u=500km/h   (c) u=600km/h
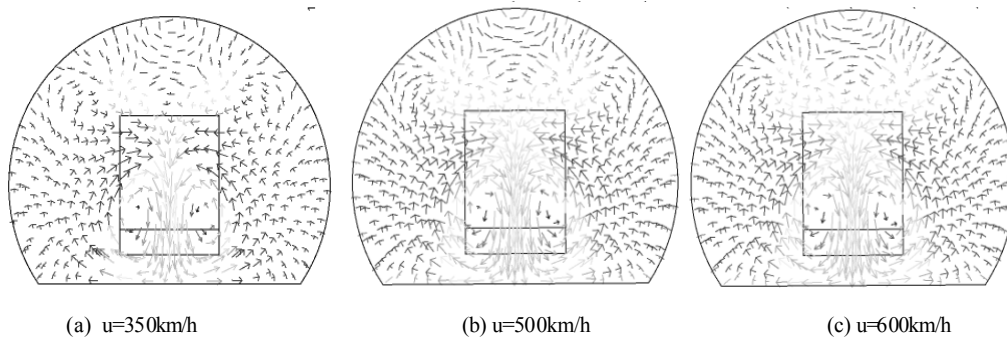
Figure12. Second flow at 2.5m behind the train with different train speed (a =45°)

(a) u=350km/h   (b) u=500km/h   (c) u=600km/h
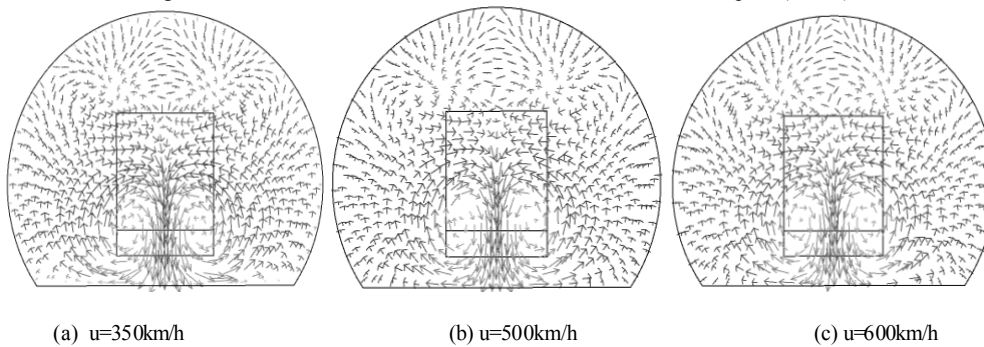
Figure13. Second flow at 5m behind the train with different train speed (a =45°)

CONCLUSIONS

This study was focused on the character of air flow properties around a train body which generated by high-speed train running in long a tunnel. Simulations have been carried out with different train nose angle in same velocity and different train speed with train nose angle 45°. We have conducted numerical simulations with finite volume method. The velocity streamline, second flow at the wake flow, total pressure and the coefficient of flow resistance have been described in detail. According to the previous work, the train nose angle leads instead to a more pronounced effect in air flow properties around train body in despite of train speed and blockage ratio being known to be one of the most important parameters influencing the aerodynamics of train/tunnel system. We have confirmed:

1. For the train-tunnel system, with ß =0.23, around the train body, aerodynamics effects continuously increase with train nose angle. The larger the train nose angle, the larger the aerodynamic drag. ?p between the train ends is also significantly increased accompanying with nose angle and train speed. Air flow accelerates along the train nose and the annular space in a quite spectacular way, especially for train speed 600km/h.

2. The separation occurs along the entire rear body and the second flow generated vortexes at the wake flow. The effect of vortex kinetic energy will be enhanced with the increasing of train nose angle. In addition, the higher train speeds, the stronger vortexes. Distance also the other factor that influences the second flow.

3. The coefficient of flow resistance increases with the rise of the train nose angle. But it almost not changed with train velocity.

4. High speed long tunnel is different from short tunnel. With the train nose angle and train speed increasing, energy-consuming for trains released into the tunnel space is increased too. In order to reduce the effect of aerodynamics, decreasing energy consumption, a equal to 30° is better than 60°.

REFERENCES

[1] W. von Tollmien, "Der Luftwiderstand und Druckverlauf bei der Fahrt von Zugen in einem Tunnel".VDI-Zeitschrift vol.71, pp.199–203, June 1927.

[2] T. Hara J. Okushi, "Model tests on aerodynamical phenomena of a high speed train entering a tunnel". Quarterly Report of RTRI, vol.3 pp.6–10, April 1962.

[3] A. Yamamoto, "Aerodynamics of a train and tunnel". Proceedings of the First International Conference on Vehicular Mechanics, Detroit: 1968, pp.151–163.

[4] W.A.Woods, C.W.Pope, "A generalised flow prediction method for the unsteady flow generated by a train in a single-track tunnel". Journal of Wind Engineering and Industrial Aerodynamics, vol.7, pp.331-360, 1981.

[5] Raghu S. Raghunathan, H.-D. Kim, and T. Setoguchi, "Aerodynamics of high-speed railway train". Progress in Aerospace Sciences, vol.38, pp.469– 514, 2002.

[6] Pierre Ricco, Arturo Baron, Paolo Molteni, "Nature of pressure waves induced by a high-speed train traveling through a tunnel". Journal of Wind Engineering and Industrial Aerodynamics, vol.8, pp. 781-808,1995.

[7] Zhu Yanfeng, Wu Yaping, Yang Changyu, Lin Bentao, "Temperature rising by a high-speed train running through a long tunnel" In: 2008 International Workshop on Modeling, Simulation, and Optimization, Hong Kong, China 2008, pp.47-50.

[8] Zhu Yanfeng, Wu Yaping, Yang Changyu and Lin Bentao, "Temperature filed generated by high speed train traveling through a deep buried long tunnel". Proceedings of Second International Conference on Modelling and Simulation, Manchester, England, UK, May. 21-22, 2009, vol.3. pp1-6.

[9] Zhu Yanfeng, Wu Yaping, 'The study of temperature filed by high speed train traveling through a long tunnel". 2009 International Conference on Energy and Environment Technology, Guilin china, 2009, pp.445-448.

[10] Arturo Baron, Michele Mossi and Stefano Sibilla,H.B " The alleviation of the aerodynamic drag and wave effects of high-speed train in very long tunnels". Journal of Wind Engineering and Industrial Aerodynamics, vol.89, pp.365-401, 2001.

Yanfeng. Zhu, Place of birth: Shanxi China; Date of birth: 09,06,1968. Educational background: Bachelor of Industry in Engineering Mechanics, Southwest Jiaotong University, Chengdu China, 1989; Master of Industry in Solid Mechanics, South China University of Technology, Guangzhou China, 2005. 2006-Present, Study for Doctor of Bridge & Tunnel Engineering in Lanzhou Jiaotong University.

Major field: Bridge & Tunnel Engineering.

Professional experience: 1989.7-1997.1: Tianjin Railway Technical and Vocational College, Lecturer; 1997.2-Present: Guangdong University of Technology, Faculty of Civil and Transportation Engineering, Lecturer. Guangzhou China.

1. Zhu Yanfeng, Wu Yaping, Yang Changyu, Lin Bentao, "Temperature rising by a high-speed train running through a long tunnel" In: 2008 International Workshop on Modeling, Simulation, and Optimization, Hong Kong, China 2008. pp.47-50.

2. Zhu Yanfeng, Wu Yaping, Yang Changyu and Lin Bentao, "Temperature filed generated by high speed train traveling through a deep buried long tunnel". Proceedings of Second International Conference on Modelling and Simulation Manchester, England, UK, May. 21-22, 2009, vol.3. pp.1 -6.

3. Zhu Yanfeng, Wu Yaping, 'The study of temperature filed by high speed train traveling through a long tunnel". 2009 International Conference on Energy and Environment Technology, pp.445-448.

Xiangyun Liu, Place of birth: Shanxi China; Date of birth: 09,15,1968. Educational background: Doctor of thermal-engineering, Beijing University of Aeronautics and astronautics, Beijing China

Major field: Thermal-Engineering and Fluid Mechanics.

Zhanlin. Cui, Place of birth: Inner Mongolia China; Date of birth: 25,05,1967. Educational background: Bachelor of Industry in Engineering Mechanics, Southwest Jiaotong University, Chengdu China, 1989.

Major field: Engineering Mechanics and Fluid Mechanics.

Professional experience:

1989.7-1994.8 Harbin Railway Technical l College, Lecturer.

1994.9-Present: Harbin Railway Technical College, Associate Professor.

Hybrid Coevolutionary Particle Swarm Optimization for Linear Variational Inequality Problems

Liangdong Qu*

College of Mathematics and Computer Science, Guangxi University for Nationalities, Nanning 530006, China
Email: quliangdong@163.com

Dengxu He

College of Mathematics and Computer Science, Guangxi University for Nationalities, Nanning 530006, China
Email: dengxuhe@126.com

Yong Huang

College of Mathematics and Computer Science, Guangxi University for Nationalities, Nanning 530006, China
Email: Yonghuang@163.com

***Abstract*—Solving linear variational inequality by traditional numerical iterative algorithm not only can not satisfy parallel, but also its precision has much relationship with initial values. In this paper, a novel hybrid coevolutionary particle swarm optimization is used to solve linear variational inequality, which sufficiently exerts the advantage of particle swarm optimization such as group search, global convergence and it satisfies the question of parallel solving linear variational inequality in engineering. It also overcomes the influence of initial values. Several numerical simulation results show that the coevolutionary algorithm offers an effective way to solve linear variational inequality, high convergence rate, high accuracy and robustness.**

***Index Terms*—linear variational inequality; extragradient method; particle swarm optimization; parallel; global convergence; coevolution**

I. INTRODUCTION

Linear variational inequality problem (LVIP) is a fundamental field. It plays an important role in differential equation, mechanics, cybernetics, game theory economy and finance, transport and many other fields. Linear variational inequality problem can be described as follows:

Suppose that S is a nonempty closed convex subset of R^n, find a vector $x^* \in S$ such that the inequality

$$\langle Mx^* + q, x - x^* \rangle \geq 0, \qquad (1)$$

i.e.

$$(x - x^*)^T (Mx^* + q) \geq 0, \qquad (2)$$

for all $x \in S$.

Where $M = (m_{ij})_{n \times n}$ is an $n \times n$ real matrix,

$q = (q_1, q_2, \cdots, q_n)^T \in R^n$, $\langle \cdot, \cdot \rangle$ denotes the inner product in R^n .

So far, this problem has been studies extensively and many methods developed for solving the problem, such as projection methods, extragradient method, and Newton methods, etc [1-3]. Some method [4] may be applied to the inequality (1) by reformulating the problem as mixed complementarity problem through its KKT system. But those approaches increase the dimension of the problem by introducing Lagrangian multipliers. Although there are plenty of methods proposed for solving the inequality (1), most of them rely on iterative methods. It usually claims parallel solving linear variational inequality problem in engineering and technology, but the iterative algorithms that based on traditional digital computer are hard to satisfy the demands of parallel solving. Furthermore, the convergence of iterative methods has much relationship with initial values.

To overcome these drawbacks, it has important actual meaningful to parallel solving the inequality (1) by studying and using particle swarm optimization.

Particle swarm optimization (PSO) [5] is a population-based, self-adaptive search optimization method motivated by the observation of simplified animal social behaviors such as fish schooling, bird flocking, etc. It is becoming very popular due to its simplicity of implementation and ability to quickly converge to a reasonably good solution [6, 7].

This paper presents a solution to solve linear variational inequality by hybrid coevolutionary particle swarm optimization. The algorithm can actualize parallel search in the feasible region and converge to the global optimal solution, which efficiently overcome the defects that the traditional methods can not parallel solve linear variational inequality in engineering technique and its precision has much relationship with initial values.

* Corresponding author: Liangdong Qu.

The rest of this paper is organized as follows: in Section 2, we give the brief theories of linear variational inequality. In Section 3, we propose extragradient method for linear variational inequality. Section 4, we introduce particle swarm optimization. In Section 5, we propose a hybrid coevolutionary particle swarm optimization for linear variational inequality. In Section 6, we report numerical results of the methods for some test problems. Finally, Section 7 gives conclusions.

II. THE THEORIES OF LINEAR VARIATIONAL INEQUALITY

In this paper, we select the 2-norm as the vector norm. For explicitness, we define the subset:

$$S = \{x \in R^n \mid x = (x_1, x_2, \cdots, x_n)^T, x_i \in [a_i, b_i], i = 1, 2, \cdots, n\},$$

where $a_i \leq b_i$.

$\forall p = (p_1, p_2, \cdots, p_n)^T \in R^n$, we define the function:

$$F_S(p) = (F_{S1}(p_1), F_{S2}(p_2), \cdots, F_{Sn}(p_n))^T,$$

where

$$F_{Si}(p_i) = \begin{cases} a_i & (p_i < a_i), \\ p_i & (a_i \leq p_i \leq b_i), \\ b_i & (p_i > b_i), \end{cases} \quad (3)$$

Obviously, $F_{Si}(p_i)$ is a monotone continuous function whose global Lipschitz constant is 1.

Theorem 1 [8, 9]: x^* *is a solution of the linear variational inequality (1) if and only if*

$$x^* = F_S(x^* - Mx^* - q). \quad (4)$$

Theorem 2 [10]: *In S , the fixed point equation:*

$$x = F_S(x - Mx - q) \quad (5)$$

must has a solution.

III. EXTRAGRADIENT METHOD FOR LINEAR VARIATIONAL INEQUALITY

Extragradient method often been used to solve linear variational inequality.

Suppose that $\max\{|m_{ij}| \mid i = 1, 2, \cdots, n; j = 1, 2, \cdots, n\}$ is denoted by k . Korpelevich [11] introduced the following so-called extragrient method:

$$\begin{cases} x^{(0)} = x \in R^n, \\ y^{(t)} = F_S(x^{(t)} - \lambda(Mx^{(t)} + q)), \\ x^{(t+1)} = F_S(x^{(t)} - \lambda(My^{(t)} + q)) \quad t = 0, 1, \cdots, \end{cases} \quad (6)$$

where $\lambda \in (0, 1/k)$. $x^{(t)}$ and $y^{(t)}$ refer to some vectors and t is iteration index.

He showed that the sequences $\{x^{(t)}\}$ and $\{y^{(t)}\}$ generated by (6) converge to the same point.

The advantage of the method is simple, but its disadvantage is its precision has much relationship with initial values.

IV. PARTICLE SWARM OPTIMIZATION

A. Particle Swarm Optimization and Its Improvement

Particle Swarm Optimization (PSO) is a population based algorithm that exploits a set of potential solutions to the optimization problem. It provides a method for finding the global optimum of a function of several real variables. As in the case of genetic algorithms it starts with a population of potential solutions. The population is called a *swarm* and the members are called *particles*. The particles belong to a group which may be the population as a whole or may be a subpopulation. The particles change (evolve, learn) over time in response to their own experience and the experience of the other particles in their group. As this interpretation suggests, the principles underlying the algorithm can be applied not only to the solution of optimization problems, but also to the representation of social behavior, including economic behavior.

A swarm consists of Np particles moving around in an n-dimensional search space. The ith particle at the tth iteration has a position $x_i^{(t)} = (x_{i1}, x_{i2}, \cdots, x_{in})^T$, a velocity $v_i^{(t)} = (v_{i1}, v_{i2}, \cdots, v_{in})^T$, the best solution achieved so far by itself (pbest) $pb_i = (pb_{i1}, pb_{i2}, \cdots, pb_{in})^T$. The best solution achieved so far by the whole swarm (gbest) is represented by $pg = (pg_1, pg_2, \cdots, pg_n)^T$. The position of the ith particle at the next iteration will be calculated according to the following equations:

$$v_i^{(t+1)} = w \cdot v_i^{(t)} + c_1 \cdot r_1 \cdot (pb_i - x_i^{(t)}) + c_2 \cdot r_2 \cdot (pg - x_i^{(t)}), \quad (7)$$

$$x_i^{(t+1)} = x_i^{(t)} + v_i^{(t+1)}, \quad (8)$$

where c_1 and c_2 are two positive constants, called cognitive learning rate and social learning rate, respectively; r_1 and r_2 are two separately generated uniformly distributed random numbers in the range [0,1]; w is inertia weight factor. PSO has received much attention. Many researchers have devoted to improving its performance in various ways and developed many interesting variations. Most variations can be roughly grouped into the following categories.

(a) The improvement depends on incorporating the new coefficient into the velocity and position equations of the PSO algorithm or rationally selecting the values of the coefficients. Angeline [12] pointed that the original version of PSO had poor local search ability. In order to overcome this disadvantage, Shi and Eberhart [13] proposed linearly decreasing weight particle swarm optimization (LDWPSO) where a linearly decreasing inertia factor was introduced into the velocity update equation of the original PSO. Because the inertia factor effectively balances the global and local search abilities

of the swarm, performance of PSO is significantly improved. Clerc [14] presented constriction PSO where a constriction factor was introduced into PSO to control the magnitude of velocities. It was mathematically proved that the resulting algorithm could guarantee the convergence even without clamping the velocity. However, if the strategy of clamping the velocity was combined into the algorithm, the performance could be improved further. The constriction PSO has faster convergence than LDWPSO, but it is prone to be trapped in the local optima when multi-modal functions are presented.

(b) A key feature of PSO algorithms is social sharing information among the neighborhood. Therefore, various information sharing mechanisms were proposed to improve the performance. Kennedy [15] investigated the impacts of various neighborhood topologies and pointed out that the von Neumann topology results in superior performance. Suganthan [16] introduced a variable neighborhood operator. During the initial stages of the optimization, the neighborhood will be an individual particle itself. As the number of generations increases, the neighborhood will be gradually extended to include all particles. Mohais [17] proposed dynamically adjusted neighborhood where randomly generated, directed structures were used as the topology of the initial population and then edges of the structures were randomly migrated from one source node to another during the course of run. Liang et al. [18] presented a new learning strategy to make particles have different learning exemplars for different dimensions. Van den Bergh and Engelbrecht [19] proposed to split the solution vector into several sub-vectors which were then allocated their own swarms. Peram et al. [20] proposed to utilize the additional information of the nearby higher fitness particle that was selected according to fitness-distance-ratio (FDR) that denoted the ratio of fitness improvement over the respective distance. Baskar and Suganthan [21] proposed a novel concurrent PSO algorithm where modified PSO and FDR-PSO algorithms were simulated concurrently with frequent message passing between them. Janson [22] used dynamic hierarchy to define the neighborhood structure. He introduced an additional component, passive congregation, into the velocity update equation.

(c) The operators of other evolutionary algorithms were combined with PSO. Angeline [23] used selection operator to improve the performance of the PSO algorithm. Lovbjerg et al. [24] combined the PSO algorithm with the idea of breeding and subpopulations. Poli et al. extended PSO via genetic programming [25]. Zhang and Xie [26] introduced differential evolution operator into PSO. Lovbjerg et al [24] combined PSO, genetic algorithms and hill climbers. Various mutation operators were also incorporated into PSO [27].

(d) Some mechanisms were designed to increase the diversity in order to prevent premature convergence to local minimum. Silva et al. [28] presented a predator–prey model to maintain population diversity. Zhang et al. [29] proposed to re-initialize the velocities of all particles at a predefined extinction interval, which simulated natural process of mass extinction in the fossil record. Krink [30] proposed several collision strategies to avoid crowding of the swarm. Lovbjerg [31] used self-organized criticality (SOC) to add diversity. Riget et al. [32] introduced attractive and repulsive PSO where the two phases, attraction and repulsion, alternated during the search according to a diversity measure. Although variations of PSO have applied different strategies and parameters, all of them follow the same principle of swarm intelligence. Therefore, all variations appear similar features of social behavior. Within a swarm, individuals are relatively simple, but their collective behavior becomes quite complex. A group of particles in a swarm move around in the defined search space to find the optimum. Each particle relies on direct and indirect interaction and cooperation with other particles to determine the next search direction and step-size, so the swarm will move around and gradually converge toward the candidates of global optima or local optima. Thus the center of the swarm is probably near to the optimum. While this position changes during the search process, it can supply very useful information for capturing the optimum.

Empirical results showed the linearly decreased setting of inertia weight can give a better performance, such as linearly decreases from 1.4 to 0, and 0.9 to 0.4 through the search process [33]. In addition, the velocities of the particles are confined within $[-v_{\max}, v_{\max}]$. If an element of velocities exceeds the threshold $-v_{\max}$ or $v_{\max}$, it is set equal to the corresponding threshold.

In this paper, to improve validity of PSO, we add extragradient method into LDWPSO. They frequently exchange the best particle of the swarm during the search and then guides the whole search process.

B. Comparison to Genetic Algorithms

The PSO shares many similarities with evolutionary computation techniques such as Genetic Algorithms. Both techniques begin with a group of a randomly generated population; utilize a fitness value to evaluate the population and search for the optimum in a partially stochastic manner by updating generations. There are however important differences. Although PSO algorithm retains the conceptual simplicity of the GAs, its' evolutionary process does not create new population members from parent ones; all swarm individuals survive. The original PSO does not employ genetic operators such as crossover and mutation and particles only evolve through their social behavior. Particles' velocities are adjusted, while evolutionary individuals' positions are acted upon. In GAs chromosomes share information with each other, thus the whole population moves like one group towards an optimal area whereas in PSO the information exchange information is a one-way process since only the (local) global best provides information to members of the (sub) swarm. PSO in comparison with GAs, has less complicated operations and it is much easier to implement and apply to design problems with continuous parameters. Moreover in [34] it is shown that

binary PSO outperforms GAs in terms of convergence, results and scalability of the problems at hand. In the same study it is suggested that a hybrid model combining characteristics of GA and PSO is preferable. In [35] PSO is shown to exhibit faster convergence compared to GAs.

V. HYBRID COEVOLUTIONARY PARTICLE SWARM OPTIMIZATION FOR LINEAR VARIATIONAL INEQUALITY

A. The Principle of Solving Linear Variational Inequality by Hybrid Coevolutionary Particle Swarm Optimization (HCPSO)

Many methods for solving linear variational inequality adopt the idea of reformulating the problem as a system of nonlinear equations or an optimization problem with zero global minimum value [36]. To reformulate the LVIP as an optimization problem, we usually use so-called merit functions. Any global minimum of a merit function, say θ, with zero function value is a solution of LVIP (1) and vice versa. So the LVIP (1) is equivalent to the following global optimization problem with zero minimum value:

$$\min \ \theta(x) \tag{9}$$

i.e.

$$\min \ \| x - F_S(x - Mx - q) \|_2 \tag{10}$$

s.t. $x \in S$.

Where $\| \cdot \|_2$ represents the 2-norm. Therefore, we can define the fitness function:

$$fitness = \| x - F_S(x - Mx - q) \|_2. \tag{11}$$

Obviously, the solution x^* to the linear variational inequality is the best solution to the function optimization. Hence our task is to

minimize $\| x - F_S(x - Mx - q) \|_2$

s.t. $x \in S$.

Extragradient method are very simple and efficient, but its precision has much relationship with initial values; for PSO, it has group search, global convergence, parallel and robustness, but its efficiency is low in the latter half evolution process.

We use PSO and extragradient method not only group search, global convergence, parallel and robustness but also to improve the validity of PSO algorithm. The evolutionary algorithm is a population-based algorithm. It makes use of a fitness function to evaluate the goodness of a solution, thereby keeping the population set consisting of promising solutions. If, during the search, the population set gets stuck around a point which is not a global minimum, extragradient method direct the population set to another possible region which may contain a global solution.

We combine the two algorithms so that they complement each other, to design a hybrid coevolutionary algorithm to solve linear variational inequality.

B. The Process of Solving Linear Variational Inequality by Hybrid Coevolutionary Particle Swarm Optimization (HCPSO)

We give a new concept: memory extragrient method as follows:

Definition 1 In extragrient method, the vector x iterate l times, $l-1$ times don't output, only after l th time, its result output. It is called memory extragrient method.

Memory extragrient method as follows:

Function *memory_extragrient(l, x)*

  While (t<l) do

  $x^{(0)} = x;$

  $y^{(t)} = F_S(x^{(t)} - \lambda(Mx^{(t)} + q));$

  $x^{(t+1)} = F_S(x^{(t)} - \lambda(My^{(t)} + q));$

  End While

  Output the result $x^{(t+1)}$.

Where l is a positive integer, is called Memory length.

Then, the coevolutionary algorithm of PSO and extragrient method for LVIP flow chart is the following Fig.1.
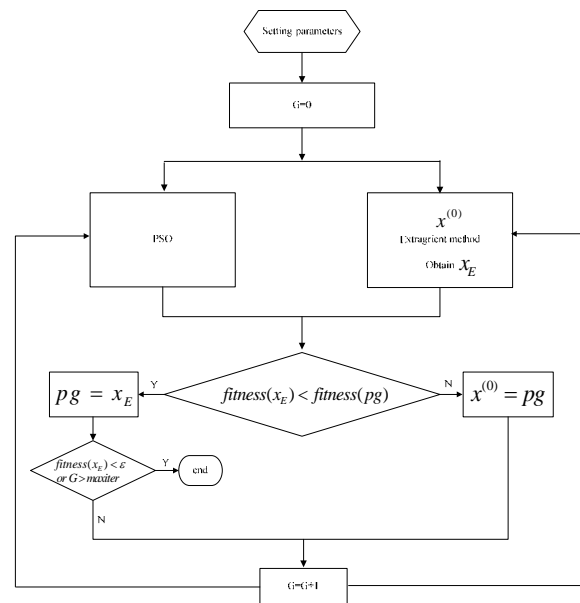

Figure 1. The flow chart of the hybrid coevolution

The steps of solving linear variational inequality by hybrid coevolutionary particle swarm optimization are illustrated as follows:

Step1. Setting evolution parameters;

Step2. Initializing the position vector and velocity vector of each component of each particle, determining the personal best location and the best location among the entire swarm; and initializing $x^{(0)}$

Step3. Calculate the weight value in current iteration according to (12):

$$w = w_{\max} - iter \cdot \frac{w_{\max} - w_{\min}}{maxiter}, \tag{12}$$

where w_{max} and w_{min} are the maximum value and minimum value for weights, respectively, *iter* is the current iteration, and *maxiter* is the maximum iteration;

Step4. Calculate the fitness value of all particles according to (11), and derive the best particle of current iteration, and the best position that every particle has reached so far.

Step5. If $fitness(x_E) < fitness(pg)$, then $pg = x_E$, else $x^{(0)} = pg$.

Step6. Update velocity and position of every particle according to (7) and (8).

Step7. If the stop criteria are satisfied, output the obtained best solution and the algorithm is ended; otherwise, go to Step 3.

This hybrid algorithm has the following characteristics:

(a) Structure supplementary, enhances parallelism.

The extragrient method's structure is simple, is the serial structure; PSO is the parallel structure. After mix, those enhance the entire algorithm parallelism.

(b) Behavior supplementary, raises the convergence speed.

The extragrient method iterates quickly, but it hasn't the population information, easy to fall into the local minimum; The PSO iteration does not be quick, but has certain overall situation ability. After the mix, may enhance PSO the convergence rate.

(c) Weakens the stringency of parameters.

The two algorithms have some parameters, the latter experiments show that the mixed algorithm is dependent on the parameters is not strong.

VI. SIMULATION EXPERIMENTS

Several experiments have been conducted to evaluate the performance of the proposed CHPSO. In the experiments, the number of particles of the swarm Np was set at 30 and the maximum number of iterations *maxiter* was set at 50. Other parameters were set at $w_{max} = 1.4$, $w_{min} = 0$, $c_1 = 2$, $c_2 = 2$, $v_{max} = 2$, $l = 10$ and $\lambda = 0.5$.

The experimental conditions are: Intel(R) Core (TM) 2 Duo 2.20GHz CPU, 1G Memory and Windows XP operation system. The programs are realized in Matlab7.

To decrease the impact of randomicity, 20 independent runs were performed for each test case. The best optimal solution, the west optimal solution, the average solution and between them to the exact solution were recorded, meanwhile the average error curve of each iteration in evolution of 20 independent runs for each test case were given.

Example 1: Let $q = (1, -5)^T$,
$$M = \begin{pmatrix} 1 + \sqrt{2}/2 & \sqrt{2}/3 \\ -\sqrt{2}/3 & 1 + \sqrt{2}/3 \end{pmatrix} \text{ and } S = [1,2] \times [3,4].$$ We want to find a vector $x^* \in S$ such that (1) for all $y \in S$. Its exact solution is $(1, 3.718491035931836)^T$. For

extragrient method, its initial value $x^{(0)} = (2,3)^T$, $\lambda = 0.5$. For LDWPSO and HCPSO, 20 independent runs were performed, respectively. Table1 and Fig.2 show the results of 20 runs.

TABLE I.
EXPERIMENTAL RESULTS FOR EXAMPLE 1

| Algorithm	Optimal solution x		Error $\| x - x^* \|_2$
HCPSO	west	(1.000000000 000000, 3.7184910359 31838)^T	2.220446049250313e-015
	best	(1.000000000 000000, 3.7184910359 31836)^T	4.440892098500626e-016
	average	(1.000000000 000000, 3.7184910359 31835)^T	7.105427357601002e-016
LDWPSO	west	(1.000000000 000000, 3.7184906850 79734)^T	3.508521020378908e-007
	best	(1.000000000 000000, 3.7184910338 47485)^T	2.084351358178083e-009
	average	(1.000000000 000000, 3.7184910053 21227)^T	4.595095624004841e-008
Extragrient method	---	(1.000000000 000000, 3.7184844569 90989)^T	6.578940846768688e-006
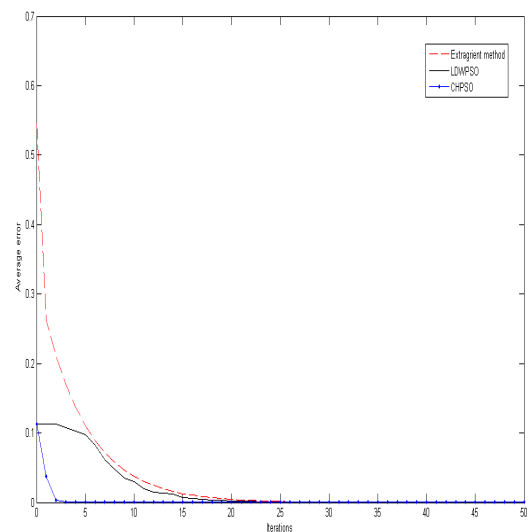
Figure 2. Average error curve for Exmple 1

Example 2: Let $q = (-1, 1, -0.5)^T$,

$$M = \begin{pmatrix} 0.1 & 0.1 & -0.5 \\ 0.1 & 0.1 & 0.5 \\ 0.5 & -0.5 & 1 \end{pmatrix} \quad \text{and}$$

$S = [-5, 5] \times [-3, 1] \times [-4, 3].$

We want to find a vector $x^* \in S$ such that (1) for all $y \in S$. Its exact solution is $(2.5, -2.5, -2)^T$. For extragrient method, its initial value $x^{(0)} = (0, -1, 2)^T$. For LDWPSO and HCPSO, 20 independent runs were performed, respectively. Table2 and Fig.3 show the results of 20 runs.

TABLE II.
EXPERIMENTAL RESULTS FOR EXAMPLE 2

Algorithm	Optimal solution x		Error $\|x - x^*\|_2$
HCPSO	west	(2.499999999 999998, -2.500000000 000003, -2.000000000 000000)^T	3.20237283398937 7e-015
	best	(2.499999999 999998, -2.500000000 000003, -2.000000000 000000)^T	3.20237283398937 7e-015
	average	(2.499999999 999998, -2.500000000 000003, -2.000000000 000000)^T	3.20237283398937 7e-015
LDWPSO	west	(2.193347775 682247, -3.000000000 000000, -2.064455381 986260)^T	0.59007633654131 6
	best	(2.499550187 663737, -2.499651619 860490, -1.999819645 112649)^T	5.96848175712219 3e-004
	average	(2.470433713 720782, -2.549864350 200948, -2.007069394 971227)^T	0.06261088852596 1
Extragrient method	---	(2.495522823 984001, -2.504477906 996408, -1.999998078 513037)^T	0.00633220023679 2
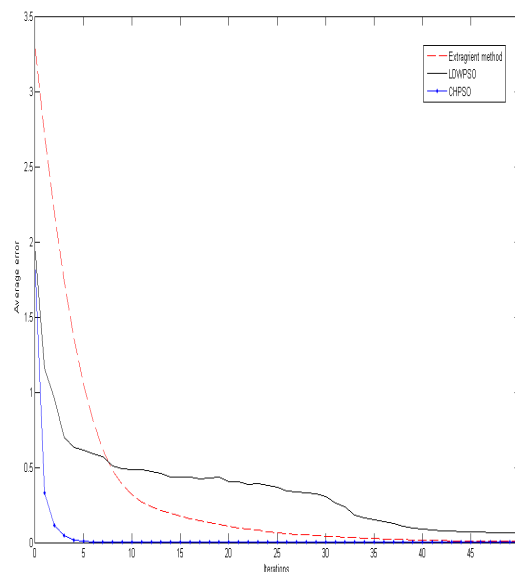
Figure3. Average error curve for Exmple 2

Example **3**:
Let $q = (-1, -1, -1, -1, -1, -1, -1, -1)^T$,

$$M = \begin{pmatrix} 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 \\ 0 & 0 & 1 & 2 & 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 1 & 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 1 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and}$$

$S = [0, 2] \times [0, 2] \times [0, 2] \times [0, 2] \times [0, 2] \times [0, 2] \times [0, 2] \times [0, 2].$

We want to find a vector $x^* \in S$ such that (1) for all $y \in S$. Its exact solution is $(0, 0, 0, 0, 0, 0, 0, 1)^T$. For extragrient method, its initial value $x^{(0)} = (1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5)^T$. For LDWPSO and HCPSO, 20 independent runs were performed, respectively. Table3 and Fig.4 show the results of 20 runs.

TABLE III.
EXPERIMENTAL RESULTS FOR EXAMPLE 3

Algorithm	Optimal solution x		Error $\| x - x^* \|_2$
HCPSO	west	$(0, 0, 0, 0, 0,$ $0, 0,$ 1.0000000000 $00000)^T$	4.440892098500626e-016
	best	$(0, 0, 0, 0, 0,$ $0, 0,$ 1.0000000000 $00000)^T$	4.440892098500626e-016
	average	$(0, 0, 0, 0, 0,$ $0, 0,$ 1.0000000000 $00000)^T$	4.440892098500626e-016
LDWPSO	west	$(0, 0, 0, 0, 0,$ $0, 0,$ 1.0000009192 $52833)^T$	9.192528329649008e-007
	best	$(0, 0, 0, 0, 0,$ $0, 0,$ 0.9999999736 $17297)^T$	2.638270291122069e-008
	average	$(0, 0, 0, 0, 0,$ $0, 0,$ 1.0000000432 $09045)^T$	2.971536657903062e-007
Extragrient method	---	$(0, 0, 0, 0, 0,$ $0, 0,$ 1.0000002831 $60828)^T$	2.831608281184117e-007
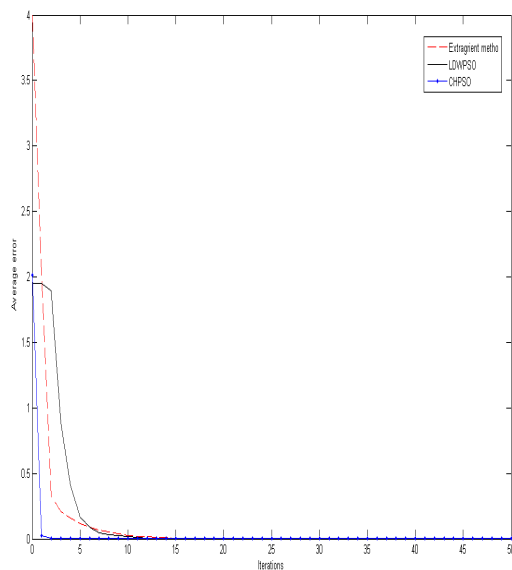
Figure4. Average error curve for Exmple 3

From Table1, Table 2, Table 3, Fig.2, Fig.3 and Fig.4, we find that the proposed algorithm enable global convergence to the best solution, has high accuracy, high convergence rate and parallelism. It is much better than LDWPSO and the extragrient method.

VII. CONCLUSIONS

In this paper, we present a novel hybrid coevolutionary algorithm for solving linear variational inequality by PSO such as group search, parallel and global convergence. This algorithm avoids the trivial deviate operation in solving linear variational inequality by traditional methods and overcome the defect that the convergence of traditional methods has much relationship with initial values, whose parallelism satisfies the question of parallel solving linear variational inequality in engineering technique. Experimental results show that the algorithm can converge to the best solution, and it has high accuracy, high convergence rate and parallelism make it easily use in engineering.

REFERENCES

[1] S. Dafermos, An iterative scheme for variational inequalities. Math. Programming, 1983, 6: 40-47.

[2] K. Taji, M. Fukushima and T. Ibaraki, A globally convergent Newton method for solving strongly monotone variational inequalities. Math. Programming, 1993, 58(3): 369-383.

[3] J.B. Jian and Y.L. Lai, Some Sorts of Methods for Solving Variational Inequalities. J. Applied Mathematics Chinese University, 1997, 14(2): 197-212. (in Chinese)

[4] Kanzow, C, Global optimization techniques for mixed complementarity problems, Journal of Global Optimization, 2000, 16: 1–21.

[5] J. Kennedy and R.C. Eberhart, Particle swarm optimization, Proceedings of IEEE International Conference on Neural Networks (1995) 1942–1948.

[6] H.Y. Shen, X.Q. Peng, J.N. Wang and Z.K. Hu, A mountain clustering based on improved PSO algorithm, Lecture Notes on Computer Science 3612, Changsha, China, 2005, pp. 477–481.

[7] R.C. Eberhart and Y. Shi, Extracting rules from fuzzy neural network by particle swarm optimization, in: Proceedings of IEEE International Conference on Evolutionary Computation, Anchorage, Alaska, USA, 1998.

[8] B.C. Eaves, On the Basic Theorem of Complementarily's. Math. Programming, 1971, 5: 68-75.

[9] U. Mosco, Introduction to Approximate Solutions for Variational Inequalities. Science and Technology of Shanghai Press, 1985.

[10] Z.G. Zeng and X.X. Liao, The solution to linear variational inequality by neural network. J. Huazhong Univ. of Sci. & Tech. (Nature Science Edition), 2002, 30(12):18-20.(in Chinese)

[11] G. M. Korpelevich, The extragradient method for finding saddle points and other problems, Matecon, 12 (1976), 747-756.

[12] P.J. Angeline, Evolutionary optimization versus particle swarm optimization: philosophy and performance differences, Lecture Notes in Computer Science, vol. 1447, Springer, Berlin, 1998, pp. 601－610.

[13] Y. Shi, R. Eberhart, A modified particle swarm optimizer, Proceedings of the IEEE Conference on Evolutionary Computation, 1998, pp. 69－73.

[14] M. Clerc, J. Kennedy, The particle swarm-explosion, stability, and convergence in a multidimensional complex space, IEEE Trans. Evol. Comput. 6 (2002) 58－73.

[15] J. Kennedy, R. Mendes, Population structure and particle swarm performance, Proceedings of the Congress on Evolutionary Computation, 2002, pp. 1671－1676.

[16] P.N. Suganthan, Particle swarm optimizer with neighbourhood operator, Proceedings of the Congress on Evolutionary Computation, 1999, pp. 1958－1962.

[17] A. Mohais, C. Ward, C. Posthoff, Randomized directed neighborhoods with edge migration in particle swarm optimization, Proceedings of the IEEE Congress on Evolutionary Computation, 2004, pp. 548－555.

[18] J.J. Liang, A.K. Qin, P.N. Suganthan, S. Baskar, Particle swarm optimization algorithms with novel learning strategies, Proceedings of IEEE Conference on Systems, Man and Cybernetics, 2004, pp. 3659－3664.

[19] F. van den Bergh, A.P. Engelbrecht, A cooperative approach to particle swarm optimization, IEEE Trans. Evol. Comput. 8 (3) (2004) 225－239.

[20] T. Peram, K. Veeramachaneni, C.K. Mohan, Fitness-distance-ratio based particle swarm optimization, Proceedings of the IEEE Swarm Intelligence Symposium, 2003, pp. 174－181.

[21] S. Baskar, P. Suganthan, A novel concurrent particle swarm optimization, Proceedings of the Congress on Evolutionary Computation, 2004, pp. 792－796.

[22] S. Janson, M. Middendorf, A hierarchical particle swarm optimizer and its adaptive variant, IEEE Trans. Syst. Man Cybern. Part B 35 (6) (2005): 1272－1282.

[23] P.J. Angeline, Using selection to improve particle swarm optimization, Proceedings of the IEEE Conference on Evolutionary Computation, 1998, pp. 84－89.

[24] M. Lovbjerg, T.K. Rasmussen, T. Krink, Hybrid particle swarm optimizer with breeding and subpopulations, Proceedings of the Third Genetic and Evolutionary Computation Conference, 2001, pp. 469－476.

[25] R. Poli, W.B. Langdon, O. Holland, Extending particle swarm optimization via genetic programming, Proceedings of the Eighth European Conference on Genetic Programming, 2005, pp. 291－300.

[26] W.J. Zhang, X.F. Xie, Depso: hybrid particle swarm with differential evolution operator, Proceedings of IEEE International Conference on Systems, Man and Cybernetics, 2003, pp. 3816－3821.

[27] N. Higashi, H. Iba, Particle swarm optimization with Gaussian mutation, Proceedings of the 2003 IEEE Swarm Intelligence Symposium, 2003, pp. 72－79.

[28] A. Silva, A. Neves, E. Costa, An empirical comparison of particle swarm and predator prey optimisation, Lecture Notes in Computer Science, vol. 2464, Springer, Berlin, 2002, pp. 103－110.

[29] W.J. Zhang, X.F. Xie, Z.L. Yang, Hybrid particle swarm optimizer with mass extinction, International Conference on Communication, Circuits and Systems, 2002, pp. 1170－1173.

[30] T. Krink, J.S. Vesterstrom, J. Riget, Particle swarm optimization with spatial particle extension, Proceedings of the Congress on Evolutionary Computation, 2002, pp. 1474－1479.

[31] M. Lovbjerg, T. Krink, Extending particle swarm optimisers with self-organized criticality, Proceedings of the Congress on Evolutionary Computation, 2002, pp. 1588－1593.

[32] J. Riget, J.S. Vesterstrom, A diversity-guided particle swarm optimizer—the ARPSO, Technical Report 2002-02, EVALife, Department of Computer Science, University of Aarhus, 2002.

[33] Y. Shi and R.C. Eberhart, A modified particle swarm optimizer, in: Proceedings of the IEEE International Conference on Evolutionary Computation, Anchorage, Alaska, USA, 1998, pp. 69－73.

[34] R. C. Eberhart, Y. Shi, "Comparison between Genetic Algorithms and Particle Swarm Optimization", in Evolutionary Programming VII: Proceedings of the Seventh Annual Conference on Evolutionary Programming, pp. 611－616, 1998.

[35] R. Hassan, B. K. Cohanim, O. D. Weck, G. A. Venter, "A Comparison of particle swarm optimization and the genetic algorithm", in Proceedings of the 1st AIAA Multidisciplinary Design Optimization Specialist Conference, April 2005.

[36] Kanzow, C. (2000), Global optimization techniques for mixed complementarity problems, Journal of Global Optimization, 16, 1–21.

Liangdong Qu He received the B.S. degree in mathematics from Henan Normal University, Xinxiang, China, in 2000. The M.S. degree in computer science from Guangxi University for Nationalities, Nanning, China, in 2009.

His current research interests focuses on computation intelligence and application, neural networks, algorithm design and analysis.

Dengxu He Professor. He received the B.S. degree in mathematics from Xihua Normal University, Nanchong, China, in 1987. The M.S. degree in computer science from Chongqing University, Chongqing, China, in 1990.

His current research interests focuses on computation intelligence and application, algorithm design and analysis.

Yong Huang He received the B.S. and M.S. degrees from Nanjing University of Science and Technology, Nanjing, China, in 2001 and 2004, respectively.The Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2009.

His research activity mainly focuses on information security, formal methods, computation intelligence algorithm and trusted software development.

A Novel Differential Evolution with Co-evolution Strategy

Wei-Ping Lee and Wan-Jou Chien
Information Management Department
Chung Yuan Christian University
Chun li, Taiwan
Email: wplee@cycu.edu.tw; wanlorgina@yahoo.com.tw

Abstract—**Differential evolution, termed DE, is a novel and rapidly developed evolution computation in recent years. There are some advantages of DE, including simple structure, easy use and rapid convergence speed. Besides, DE can be also applied on the complex optimization problem. However, there are some issues, such as premature convergence and stagnation, remaining in DE algorithm. To overcome those disadvantages, a different method was proposed, named CO-DE, by combining with a simple co-evolutionary model and reset mechanism. Thus, CO-DE can maintain appropriate swarm diversity and reduce the premature convergence. On the other hand, a reset mechanism was set to avoid the particle stagnates, which can further improve the performance of differential evolution. The proposed model can be now successfully applied with some well-known benchmark functions.**

Index Terms—**Differential Evolution; Evolutionary Computation; Co-evolutionary; Global optimization**

I. INTRODUCTION

During the past two decades, evolutionary computation is becoming more attractive, and numerous of researchers started to invest in this field. Evolutionary algorithms (EAs) can be applied in various fields, especially for optimization problems. As we know, EAs, a kind of computation mode, was set up upon the main concept of imitating the biological behavior. It was based on the theory of survival of the fittest, and many researchers were combined this concept with the search mechanism of evolutionary in the widely search space.

The first intelligent optimization algorithms were including evolutionary programming (EP), evolutionary strategy (ES) and genetic algorithm (GA). There are some disadvantages of the earlier algorithms, such as the complex procedure, stagnation and poor search ability. Some researchers devoted to improve those problems, and then proposed other related methods, such as particle swarm optimization (PSO) and Differential evolution (DE)[1,2]. They have better global search ability and fewer parameter setting. DE shows great performance; however, being the same as other algorithms, DE also exits problems of premature convergence and stagnation.

According to the above mentioned, we proposed a novel DE which was based on the co-evolutionary architecture. Unlike the previous one-to one methods, we will separate the population into four sub swarm, and used another mutation mechanism for evolution. We expected that the performance of solution can be enhanced by resetting a new dimension of particle upon the, to avoid the particle stagnates, if we will set a condition which could conform to this rule, we will reset.

II. RELATED WORK

This work is related to the differential evolution algorithm and cooperative co-evolution. Therefore, all related topics will be shortly described.

A. Differential Evolution

DE is an evolutionary computation that uses floating-point encoding for global optimization over search space. It was proposed by Stron and Price in 1995, and after second year, Stron and Price proved that DE is better than other algorithms by themselves [3].

The main concept of differential evolution is to enhance the differences of the individuals. DE is a population based algorithm and vector $\chi_{i,G}$ i = 1, 2…NP is an individual in the population. NP denotes population size. During one generation for each vector, DE employs mutation, recombination and selection operations to produce a trail vector and select one of those vectors with the best fitness value.

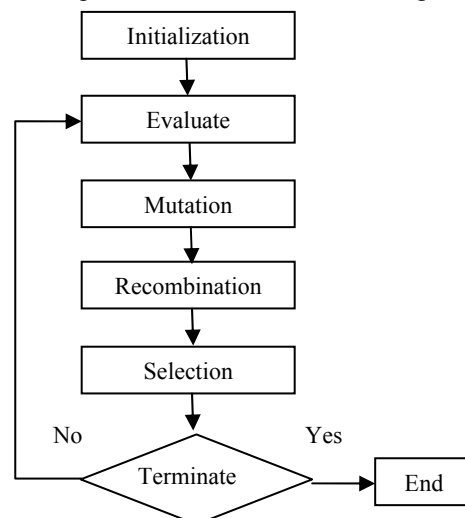The detail procedure of DE was show in figure 1.

Fig. 1 The procedure of Differential evolution

The mutation is completed by the following formulation:

$$V_{i,G+1} = \chi_{r1,G} + F(\chi_{r2,G} - \chi_{r3,G}) \qquad (1)$$

Where $\chi_{r1,G}$, $\chi_{r2,G}$ and $\chi_{r3,G}$ are three different individuals of the population, and combine with a mutation weighting factor (F) which is a parameter between [0,1] to obtain a Donor vector $V_{i,G+1}$.

The recombination is completed by the following formulation:

$$U_{i,G+1} = \begin{cases} V_{i,G+1}, & \text{if rand} \le CR \\ \chi_{i,G}, & \text{if } rand > CR \end{cases} \qquad (2)$$

After the mutation operation, the donor vector will be use in recombination. Where CR is called crossover rate between [0,1], $\chi_{i,G}$ denotes the old individual, and $V_{i,G+1}$ denotes the new individual, if the random number is smaller than CR, we will chose $V_{i,G+1}$; on the contrary, we will chose the original one, and obtain a trial vector finally.

The selection is completed by the following formulation:

$$\chi_{i,G+1} = \begin{cases} u_{i,G+1} \text{ if } F(u_{i,G+1}) < F(\chi_{i,G+1}) \\ \chi_{i,G+1} \text{ otherwise} \end{cases})$$

Where $\chi_{i,G+1}$ denotes the old individual and $u_{i,G+1}$ denotes the trial vector which was obtained by recombination operation. Comparing these two vectors, the better one will be stayed and the other one will be eliminated. Finally, if the stopping condition is satisfied, DE will output the solution, if not, it will go back and repeat these three steps again.

B. Co-Evolutionary Mode

Co-Evolutionary mode was proposed by Ehrlich and Raven in 1964. The main concept of co-evolutionary is the relationship between butterfly and parasitic plant. Because the parasitic plant is inherence in nature, it is unable to resist the plague of vermin by their own; it produces the toxic substance to protect itself. Because of this toxic substance, butterfly also produces the resist mechanism and will become the effect of co-evolution [5].

Hilli proposed the other notion about the relationship between predators and predation [7]. Unlike the above mention, the main concept was competed with each other. Rosin and Belew also mentioned that co-evolution could not only represent one biological but also two or more. This method can improve the problems of traditional algorithm like premature convergence and high complex [6].

Based on the above described, the familiar model of co-evolution can be divided into two ways, one is cooperative co-evolution and the other is competitive co-evolution. This method was used in the genetic algorithm at the first time, and then Bergh and Engelbrecht was used this architecture on particle swarm optimization [8].

III. CO-EVOLUTIONARY DIFFERENTIAL EVOLUTION

In this section, we will introduce about this co-evolution architecture for differential evolution. We proposed a novel DE, which was named CO-DE.

A. Co-evolutionary differential evolution

CO-DE also has three steps, including mutation, recombination and selection. In traditional DE, the mutation operation was used different individuals to obtain a donor vector, but this way can't be used in co-evolution architecture as a result of the grouping may reduce the chance of select individuals. Thus, we aim to correct this disadvantage and use another mutation mechanism. After the evolution of CO-DE, we will proceed with the further improvement.
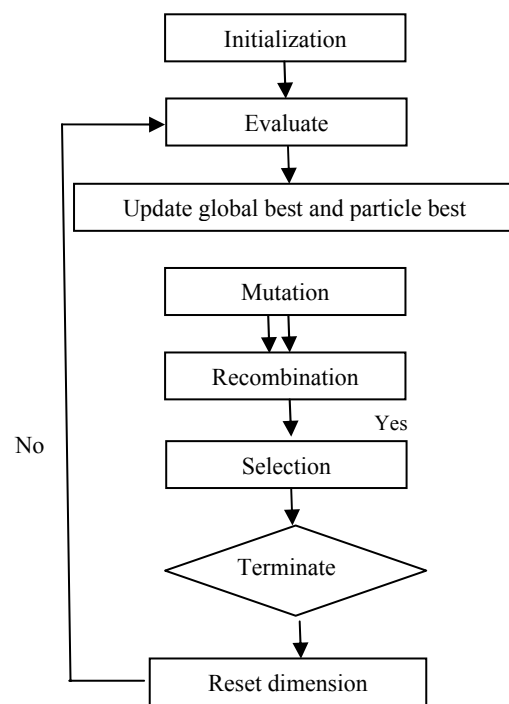

Fig. 2 The procedure of CO-DE

The basic process of CO-DE is exhibited in figure 2:

(1)Initialization: Randomly initialize the population of individual for DE, where each individual contains d variables.

(2)Update the global best and particle best and compute the fitness value.

(3)Mutation: According the global best, particle best and a random individual to evolution a donor vector.

$$V_{i,G+1} = \chi_{r1,G} + F(g_{best} - \rho_{best})$$

(4)Recombination: Perform recombination operation between each individual, and its corresponding mutant counterpart according to Eq. (2) in order to obtain each individual's trial individual.

(5)Selection: Comparing the object vector and trial vector according to Eq. (3), the trial vector will be retain into next generation if it fitness were better than object vector.

(6)Resetting: If the fitness has no change in five iterations, we will reset a dimension to each individual except the global best. We will randomly the each dimension of individuals of global and particle best, and take the Averaged. The related figure was show as in figure 3:
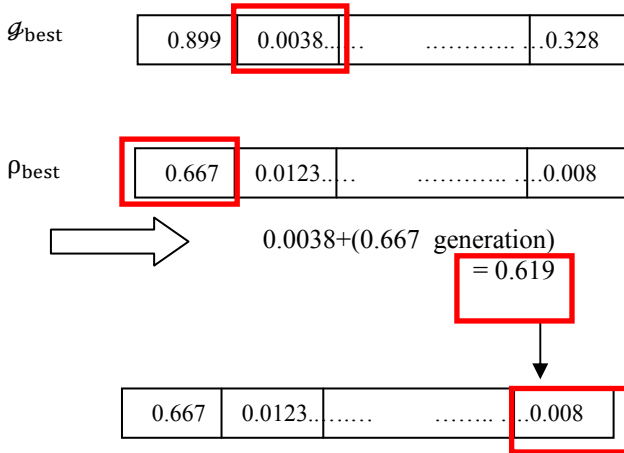
g_{best}

| 0.899 | 0.0038.. | … | ………... | ..0.328 |

ρ_{best}

| 0.667 | 0.0123.. | … | ………... | ..0.008 |

⟹ 0.0038+(0.667 generation)

= 0.619

| 0.667 | 0.0123.. | …… | ……... | ..0.008 |

Fig. 3 The procedure of CO-DE

(7) If a stopping criterion is met, then output the solution; otherwise go back to Step (2).

IV. EXPERIMENT RESULT OF COMPARISONS

To evaluate the performance of CO-DE, we will make two kinds of experiment. One of the experiment results, named co-evolution differential evolution without reset mechanism(CO-DEw), which was just added the co-evolution architecture in it. The second one, named CO-DE, which was added co-evolution and reset mechanism in the same time. Then, we will choose five benchmarks to authenticate this algorithm. The related parameter setting was showed as following:

A. Parameter setting

The related parameter setting was in table.

Table I CO-DE parameter setting

F	0.7
CR	0.2
Iter.	1000
upper_bound	100,5.12,600,32
low_bound	-100,-5.12,-600,-32
NP(population size)	60
Dimension	10/20/30

Table II DE/rand1 parameter setting

F	0.5, 0.7
CR	0.9, 0.2
Iter.	1000
upper_bound	100,5.12,600,32
low_bound	-100,-5.12,-600,-32
NP(population size)	60
Dimension	10/20/30

Table III DE/rand1-to-best/1 parameter setting

F	0.5, 0.7
CR	0.3, 0.2
Iter.	1000
upper_bound	100,5.12,600,32
low_bound	-100,-5.12,-600,-32
NP(population size)	60
Dimension	10/20/30

B. Benchmark function

Table IV Benchmark function

No	Function name	FUNCTION
f1	Sphere	$f(x) = \sum (x_i)^2$
f2	Rosenbrock	$f_2(x) = \sum_{i=1}^{n-1}\left[100\left(x_{i+1}-x_i^2\right)^2 + \left(x_i-1\right)^2\right]$
f3	Rastrigin	$f_3(x) = \sum_{i=1}^{n}\left[x_i^2 - 10\cos\left(2\pi x_i\right)+10\right]$
f4	Griewank	$f_4(x) = \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right)+1$
F5	Ackley	$f_5(x) = 20 + e - 20\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}\right) - \exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos 2\pi x_i\right)$

This section will compare the performance of CO-DE1, CO-DE2, DE/rand, and DE/rand-to-best/1. The experiment will show the different dimensions of these three methods. All of the experimental results were compared in Table V to VI.

Tables VII to VIII exhibit the meaning and standard deviation of the function values by applying the six different algorithms to optimize the 10-D, 20-D, and 30-D numerical functions *f1-f5*, the convergence of CO-DE has an obvious great performance. In 30 dimensions of *f2*, although the particles are once stagnation, it still can escape out of the regional optimal solution. And at the same time, CO-DE also report that the great performance in low-dimension. In *f3-f5*, CO-DE also has great performance and the optimization can reach the best condition.

Figs.1 to 15 illustrate the convergence characteristics in terms of the best fitness value of the median run of each algorithm for functions – with D = 10 to 30. In the all figures, CO-DE also has fast convergence, especially the function of *f3*, *f4*, it even can reach the best solution 0 at 600~700 generations.

Table IX Experimental result of *f1*

f1	DE/rand/1 (F = 0.5,Cr = 0.9)			DE/rand/1 (F = 0.7,Cr = 0.2)			DE/rand-to-best/1 (F = 0.5,Cr = 0.3)		
Dim	Avg.	Std	Best	Avg.	Std	Best	Avg.	Std	Best
10	1.2466882e-37	3.3142623e-37	3.7011251e-39	7.5510651e-22	7.6934823e-22	2.2290723e-22	3.4421486e-48	6.4382123e-48	9.5176772e-49
20	7.0187026e-17	6.80592e-17	3.9200563e-17	1.6685961e-08	1.1776923e-08	8.9341318e-09	1.3436544e-28	1.8501623e-28	2.6123423e-29
30	9.85488880e-11	7.8905724e-11	7.2569759e-11	4.76219893e-05	1.4395234e-05	3.8672214e-05	1.21388814e-23	1.4118123e-23	1.0223456e-23
	DE/rand-to-best/1 (F = 0.7,Cr = 0.2)			CO-DEw (F = 0.7,Cr = 0.2)			CO-DE (F = 0.7,Cr = 0.2)		
	Avg.	Std	Best	Avg.	Std	Best	Avg.	Std	Best
10	7.09978750e-33	9.23867234e-33	3.2023452e-34	1.09699458e-52	2.7514632e-50	7.8124443e-52	**1.47561107e-69**	**6.036272e-69**	**3.08161151e-76**
20	5.89085897e-16	4.3734223e-16	4.2959195e-16	6.01184328e-32	1.07649234e-31	8.6223421e-33	**1.45139572e-42**	**4.9882723e-42**	**4.18764933e-43**
30	8.06674252e-15	3.69939e-15	2.1813232e-15	1.11747931e-26	2.1436112e-26	4.0424231e-29	**2.33164968e-34**	**5.73074e-34**	**8.06181585e-36**

Table X Experimental result of *f2*

f2	DE/rand/1 (F = 0.5,Cr = 0.9)			DE/rand/1 (F = 0.7,Cr = 0.2)			DE/rand-to-best1 (F = 0.5,Cr = 0.3)		
Dim	Avg	Std	Best	Avg.	Std	Best	Avg.	Std	Best
10	9.0427996e+00	1.3810258e+00	2.8481244e+00	4.277833e+00	1.133932e+00	3.733589e+00	3.506e+00	1.86e+00	2.816654e+00
20	1.5838425e+01	1.3032234e+00	1.4107918e+01	1.708272e+01	1.294251e+00	1.677287e+01	1.5163e+01	1.573e+00	1.5111e+01
30	3.97234e+01	2.7472590e+00	2.6297142e+01	5.943572e+01	1.417247e+01	3.569013e+01	2.511773e+01	9.098447e-01	2.19344e+01
	DE/rand-to-best/1 (F = 0.7,Cr = 0.2)			CO-DEw (F = 0.7,Cr = 0.2)			CO-DE (F = 0.7,Cr = 0.2)		
	Avg.	Std	Best	Avg.	Std	Best	Avg.	Std	Best
10	6.9143956e+00	1.2496468e+00	5.2976756e+00	2.1834635e+00	3.47630342e-01	1.5977927e+00	**1.3816354e+00**	**3.47630342e-01**	**1.260033e+00**
20	1.8726510e+01	7.35167941e-01	1.5378953e+01	1.3027426e+01	4.066192e+00	9.3715216e+00	**4.9093048e+00**	**4.066192e+00**	**2.1228576e+00**
30	2.6107580e+01	6.67987973e-01	2.48564e+01	2.3014000e+01	1.0157974e+00	1.7022738e+01	**1.5862197e+01**	**5.1385472e+00**	**8.0209278e+00**

Table XI Experimental result of *f3*

f3	DE/rand/1 (F = 0.5,Cr = 0.9)			DE/rand/1 (F = 0.7,Cr = 0.2)			DE/rand-to-best1 (F = 0.5,Cr = 0.3)		
	Avg	Std	Best	Avg.	Std	Best	Avg.	Std	Best
Dim									
10	1.5154557e+01	5.0083099e+00	5.9697543e+00	4.582005e+00	2.069054e+00	3.157312e+00	1.658265e-01	3.77138e-01	0
20	8.7359918e+01	1.1258444e+01	7.210975e+01	7.138424e+01	7.833339e+00	5.01659e+01	3.651900e+00	2.134871e+00	2.474107e+00
30	1.6727398e+02	2.342926e+01	1.538393e+02	1.439317e+02	1.142349e+01	1.112356e+02	5.31850e+01	9.047401e+00	4.683711e+01
	DE/rand-to-best1 (F = 0.7,Cr = 0.2)			CO-DEw (F = 0.7,Cr = 0.2)			CO-DE (F = 0.7,Cr = 0.2)		
	Avg.	Std	Best	Avg.	Std	Best	Avg.	Std	Best
10	1.5418855e+00	6.03913056e-01	4.841918e-02	2.467478e-12	1.279852e-11	5.467626e-12	**0**	**0**	**0**
20	1.9173360e+01	2.6694871e+00	1.5158623e+01	3.14697e-01	5.514240e-01	3.519269e-04	**0**	**0**	**0**
30	6.6018371e+01	4.2493026e+00	6.2027629e+01	7.130986e+00	2.756044e+00	1.568877e+00	**9.62540e-012**	**1.75735e-12**	**0**

Table XII Experimental result of *f4*

f4	DE/rand/1 (F = 0.5,Cr = 0.9)			DE/rand/1 (F = 0.7,Cr = 0.2)			DE/rand-to-best1 (F = 0.5,Cr = 0.3)		
	Avg	Std	Best	Avg.	Std	Best	Avg.	Std	Best
Dim									
10	9.86138712e-05	2.234523e-03	3.46321651e-08	3.872121e-04	1.180172e-03	4.484508e-08	2.465346e-04	1.35345e-03	7.012168e-13
20	4.10699906e-04	2.249495e-03	3.17096382e-06	4.88982e-04	6.04698e-04	7.769337e-05	1.231684e-03	3.813249e-03	3.7636560-14
30	7.24494368e-02	7.4147543e-02	9.8572846e-03	2.796853e-01	1.01779e-01	1.74166e-02	1.187543e-01	5.260329e-02	7.396161e-03
	DE/rand-to-best1 (F = 0.7,Cr = 0.2)			CO-DEw (F = 0.7,Cr = 0.2)			CO-DE (F = 0.7,Cr = 0.2)		
	Avg.	Std	Best	Avg.	Std	Best	Avg.	Std	Best
10	6.33518036e-03	7.413048e-03	7.012343e-05	2.975617e-05	1.875291e-03	1.443289e-15	**0**	**0**	**0**
20	1.70807328e-02	0.01171761e-02	1.3016280e-04	7.399251e-05	2.256913e-03	1.432187e-14	**0**	**0**	**0**
30	1.2644779e-01	6.946934e-02	5.251229 39e-02	7.580846e-04	2.25272e-03	7.677203e-11	**0**	**0**	**0**

Table XIII Experimental result of *f5*

f5 / Dim	DE/rand/1 (F = 0.5,Cr = 0.9)			DE/rand/1 (F = 0.7,Cr = 0.2)			DE/rand-to-best1 (F = 0.5,Cr = 0.3)		
	Avg	Std	Best	Avg.	Std	Best	Avg.	Std	Best
10	3.96120223e-14	2.40704e-30	3.99680288e-15	4.111851e-12	5.211912e-13	3.068177e-12	3.118624e-14	1.203523e-30	1.376676e-14
20	2.97126160e-07	1.59444865e-07	3.37535275e-08	6.512781e-06	1.636534e-06	4.306734e-06	5.487412e-02	3.005579e-01	3.772342e-05
30	4.98682325e-05	2.0370623e-05	2.00096919e-05	1.563648e-03	2.28501e-03	1.420190e-03	1.055528e+00	1.98817e-15	1.316815e-02

	DE/rand-to-best1 (F = 0.7,Cr = 0.2)			CO-DEw (F = 0.7,Cr = 0.2)			CO-DE (F = 0.7,Cr = 0.2)		
	Avg.	Std	Best	Avg.	Std	Best	Avg.	Std	Best
10	2.53088513e-11	1.38607e-10	1.86561877e-12	4.196802e-15	2.407042e-30	3.994832e-15	**2.575717e-15**	**1.77022e-15**	**4.440892e-16**
20	3.752083e-01	5.8743234e-01	1.7240472e-02	5.983143e-09	2.270742e-09	3.177220e-09	**9.878379e-13**	**6.48634e-13**	**2.176037e-14**
30	1.5638050e+00	5.47997092e-01	1.78377498e-01	1.249945e-06	2.73222e-07	1.038910e-06	**6.128431e-10**	**3.03759e-10**	**3.612494e-11**
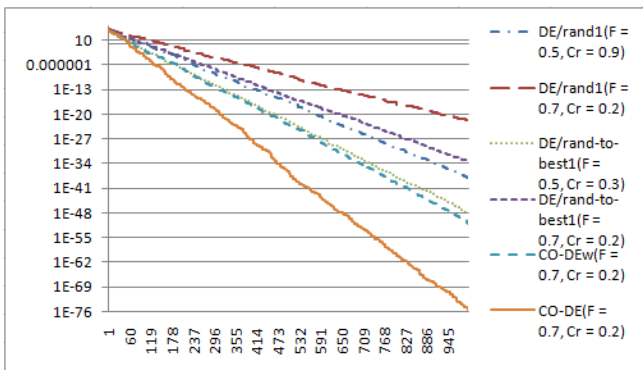
Fig.1 Convergence progress of *f1* in 10 dimensions
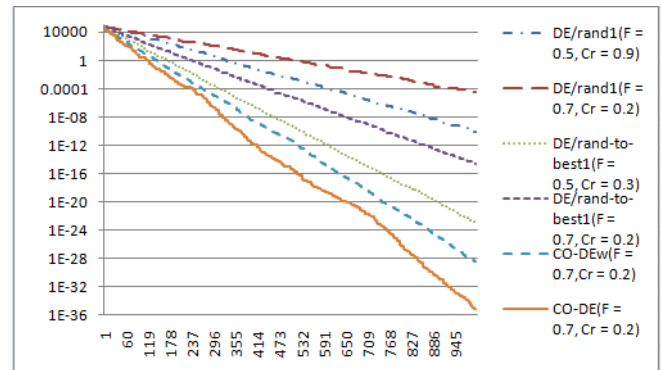
Fig.3 Convergence progress of *f1* in 30 dimensions
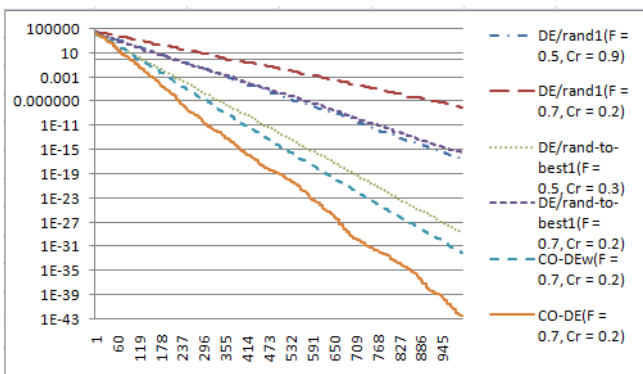
Fig.2 Convergence progress of *f1* in 20 dimensions
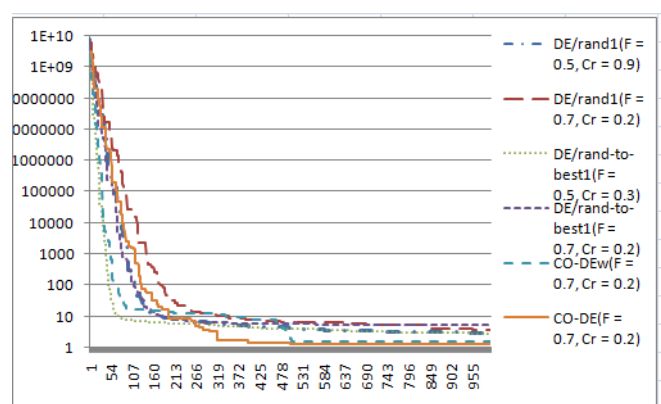
Fig.4 Convergence progress of *f2* in 10 dimensions
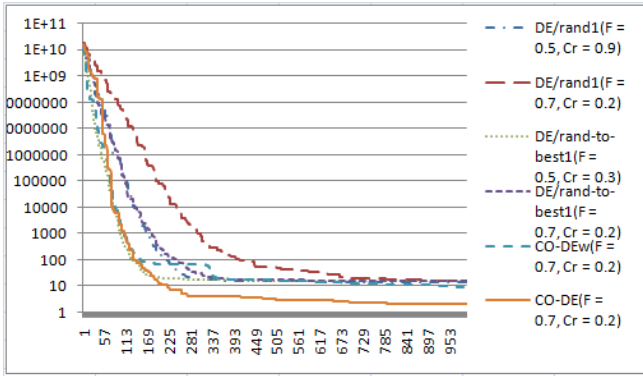
Fig.5 Convergence progress of *f2* in 20 dimensions
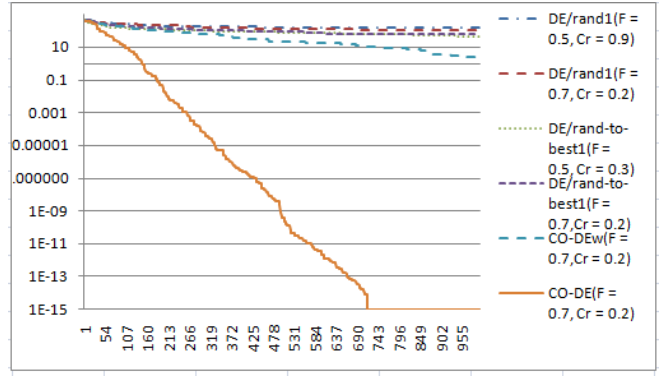
Fig.9 Convergence progress of *f3* in 30 dimensions
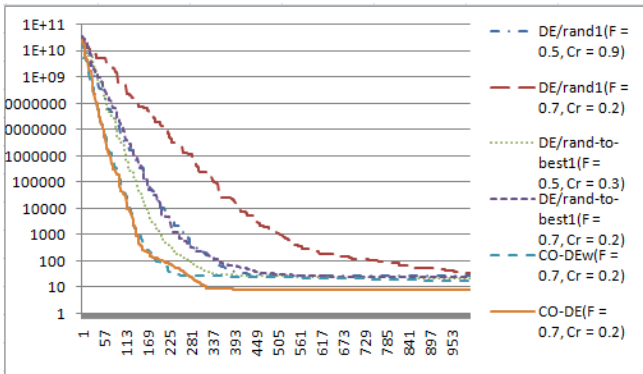
Fig.6 Convergence progress of *f2* in 30 dimensions
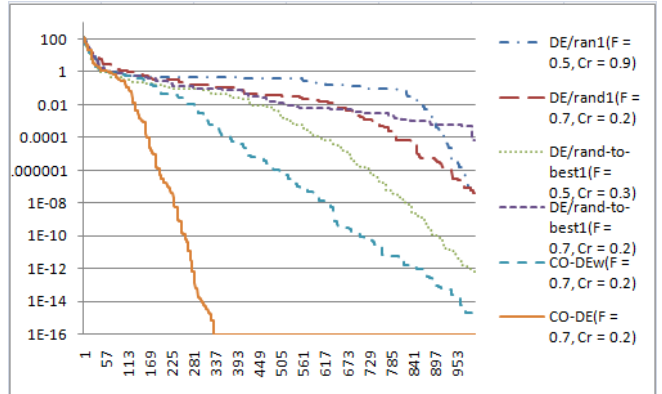
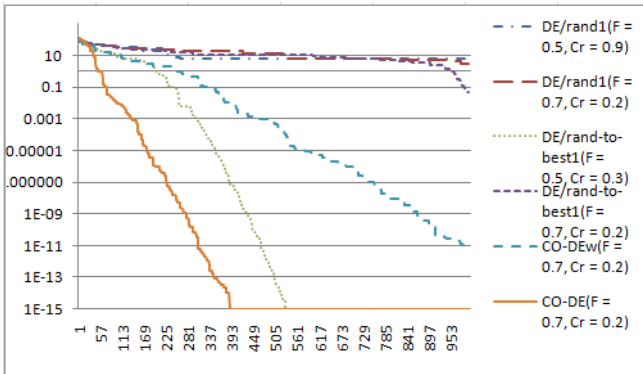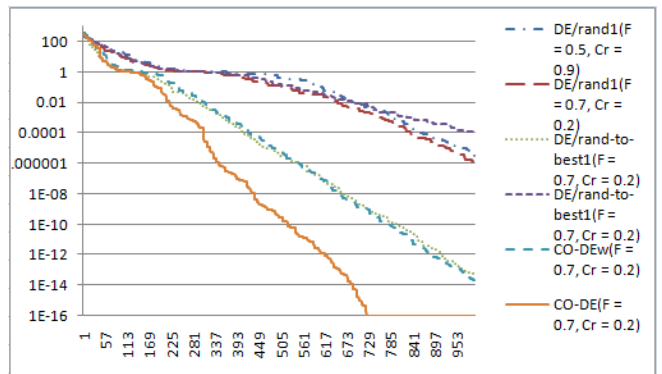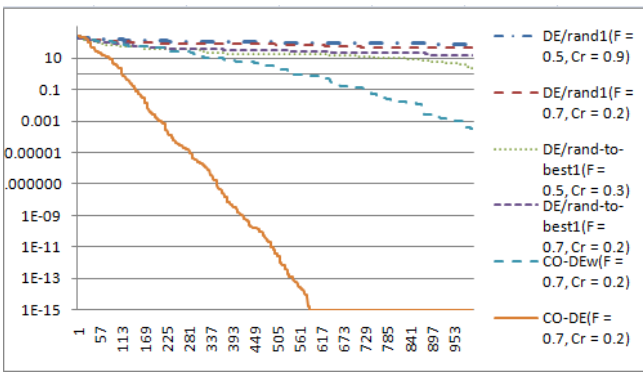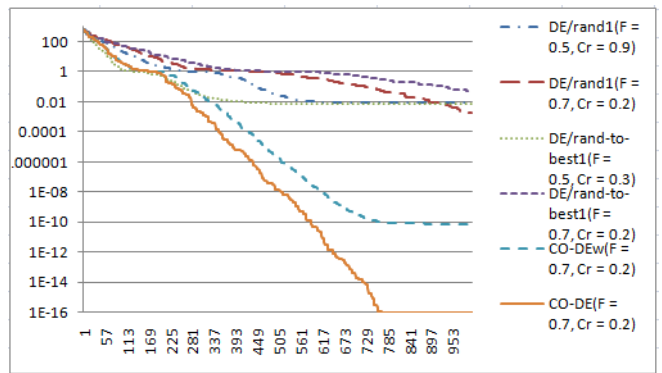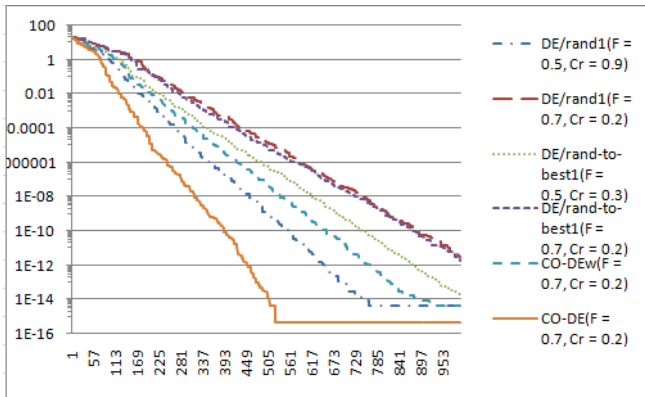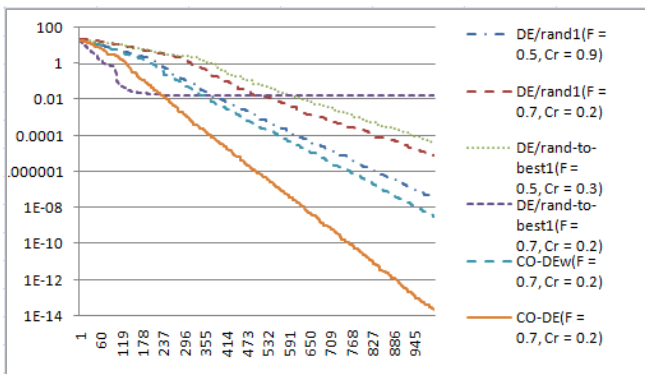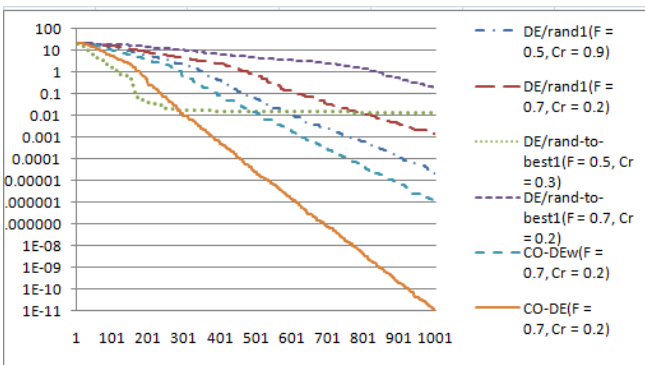Fig.10 Convergence progress of *f4* in 10 dimensions

Fig.7 Convergence progress of *f3* in 10 dimensions

Fig.11 Convergence progress of *f4* in 20 dimensions

Fig.8 Convergence progress of *f3* in 20 dimensions

Fig.12 Convergence progress of *f4* in 30 dimensions

Fig.13 Convergence progress of *f5* in 10 dimensions

Fig.14 Convergence progress of *f5* in 20 dimensions

Fig.15 Convergence progress of *f5* in 30 dimensions

C. Comparison with other researches

We will compare with three different researches, one was named KCPSO[9], which was combined the co-evolution architecture on particle swarm optimization. The second one was named PSODE[10], which was combined both differential evolution and particle swarm optimization, and it was also a kind of co-evolution architecture. The third one was named SaDE[11], which was improved by self- adaptive. And the result was compared in Table XIV.

The result showed that five different improve methods, and they both have 10-D and 30-D to optimize numerical functions *f1-f5*, and all results showed that our CO-DE has great performance except the function *f2*. In *f2*, SaDE has the best solution that the meaning of function value was about 0.3. And it was greater than other algorithms, but in other numerical function, CO-DE still has distinct improvement.

Table XV compare with other algorithms

Function	KCPSO		PSODE	SaDE	CO-DEw		CO-DE	
	10	30	10	30	10	30	10	30
Sphere	\\	\\	2.286E-51	0	1.096994e-52	4.581571e-61(0)	**1.47561107e-69**	**6.532359856e-95(0)**
Rosenbrock	\\	\\	8.804E-01	**3.99e-01**	2.183466e+00	2.548520e+00	1.3816354e+00	1.7148323e+00
Rastrigrin	1.125e-009	1.663549	7.960E-01	0	2.467478e-12	0	0	0
Griewank	0.034828	2.643e-03	3.101E-02	2.38e-03	2.975617e-05	0	0	0
Ackley	6.931e-006	2.0053-04	\\	0	4.196802e-15	0	**2.575717e-15**	0

V. COCLUSIONS

In this paper, cooperative CO-DE is proposed to solve global optimization problems. The contribution of this paper is mainly in the following two aspects: (1) through the co-evolutionary, the particles can explored the search space more in depth and reduce the speed for convergence. (2) A reset mechanism was proposed to enhance the performance and avoid the stagnation. This mechanism can prevent the above problems effectively and avoid increasing the complex of algorithm.

Due to CO-DE was an improved architecture, it may combine with other improved methods, such as improvement of parameters (F or CR) or other evolutionary mechanisms.

REFERANCES

[1] Storn, R., Price, K., "Minimizing the real functions of the ICEC'96 contest by differential evolution ", *Evolutionary Computation, 1996., Proceedings of IEEE International Conference* on 20-22 May 1996 Page(s):842 – 844

[2] Storn, R. and Price, K., "Differential Evolution- a simple and efficient adaptive scheme for global optimization over continuous spaces," *Technical Report* TR-95-012, ICSI.

[3] Price K., "Differential Evolution: A Fast and Simple Numerical Optimizer," 1996 Biennial Conference of the *North American Fuzzy Information Processing Society*, Jun. 1996, pp.524-527.

[4] Amin N., Hong W., "Co-evolutionary Self-Adaptive Differential Evolution with a Uniform-distribution Update Rule", *Intemational Symposium on Intelligent Control Munich*, Germany, October 4-6,2006.

[5] Ehrlich, P.R., P.H. Raven, 1964, "ButterDies and plants: a study in coevolution," *Evolution* vol.18 pp586–608.

[6] Rosin, C.D. and Belew, R.K., "New methods for Competitive Coevolution." Evolutionary Computation 1997, 5:1-30.

[7] Hillis, W.D., "Coevolving parasites improve simulated evolution as an optimization procedure," In Langton, C.G., Taylor, C., Farmer, J.D., & Rasmussen, S. (Eds), *Artificial Life II* , Redwood City, CA: Addison Wesley. 1992, pp.313-324.

[8] F. van den Bergh, A.P. Engelbrecht," A cooperative approach to particle swarm optimization, " *IEEE Transaction on Evolutionary Computation* 8 (3), 2004, 225–239.

[9] Jing. J., Jianchao. Z., Chongzhao. H., and Qinghua. W., "Knowledge-based cooperative particle swarm optimization", Applied Mathematics and Computation, 205, 2008, pp.861–873.

[10] N. Ben and L. Li., "A Novel PSO-DE-Based Hybrid Algorithm for Global Optimization", Computer Science, Vol. 5227, 2008, pp. 156-163.

[11] A.K. Qin, V.L.Huang, and P.N. Suganthan, "Differential Evolution Algorithm With Strategy Adaptation for Global Numerical Optimization",IEEE transactions on evolutionary computation, vol.13,No 2, April 2009.

Wei-Ping Lee received the Ph.D. degrees in Institute of Computer Science and Information Engineering from National Chiao Tung University, Hsinchu, Taiwan, R.O.C.

Currently, he has been an Assistant Professor in the Department of at Chung Yuan Christian University Chung li, Taiwan, R.O.C. His research interests include the global optimization of evolutionary algorithms, data mining and AI applications.

Wan-Jou Chien Chang-Yu Chiang received the master degree in Information Management from Chung Yuan Christian University, Chung li, Taiwan, R.O.C.

Her interests include the evolutionary algorithms and data mining.

Research on Optimization of Relief Supplies Distribution Aimed to Minimize Disaster Losses

Yong Gu

School of Logistics Engineering, Wuhan University of Technology, Wuhan, P. R. China
Email: guyong@whut.edu.cn

Abstract—The optimization problem of relief supplies distribution for large-scale emergencies is discussed in this paper. Under the situation that the needs for relief supplies are more numerous and urgency, the optimization goal should be minimum disaster loss instead of minimum delivery cost or shortest travel distance. By setting objective function about minimum disaster loss, scheduling model of relief supplies distribution under determined conditions and fuzzy conditions is proposed. Scheduling model of relief supplies distribution divide into two cases both for single species and for many varieties. We deal with the uncertainty of travel time in relief supplies distribution by applying the approach of selecting the maximum satisfaction path in fuzzy network. For each model, a experimental calculation is conducted to verify its solution approach.

Index Terms—relief supplies, scheduling, distribution, disaster losses, fuzzy set

I. INTRODUCTION

Since entering the 21st century, various catastrophic emergencies occur frequently, which caused huge casualties and property losses, such as 9•11, SARS, Bird flu, Indian Ocean tsunami, New Orleans Hurricane, Wenchuan (China) Earthquake and so on. The effective rescue and aid measures must be taken immediately to minimize the loss caused by emergent events as far as possible. In the course of the emergency preparedness and response, a large amount of emergency supplies were required for casualties' rescue, sanitary and anti-epidemic, restoration and reconstruction, and so on.

The demands for disaster relief supplies include mainly supplies of rescuing casualties and livelihood security, medicines and medical equipment. It is necessary that relief supplies are shipped to emergency demand points timely by accurate and reasonable manner in disaster relief operations. The main influence factor of relief supplies distribution are the needs of emergency demand points, travel time from service points to demand points and capacity of transportation vehicles. Therefore, the scheduling decisions of relief supplies include three parts: selecting the emergency service points, the amount of relief supplies that selected emergency service points provide and the route of the transportation vehicles from selected service points to demand points.

Haghani, Ali and Oh, Sei-Chang present a formulation

and two solution methods for a logistical problem which is a large-scale multi-commodity, multi-modal network flow problem with time windows in disaster relief management [1]. F. Fiedrich, F. Gehbauer and U. Rickers introduced a dynamic optimization model to find the best assignment of available resources to operational areas after strong earthquakes [2]. Linet Ozdama and Ediz Ekinei developed a planning model of dispatching commodities to distribution centers, which was integrated into a natural disaster logistics decision support system [3]. Sheu J.B. proposed a hybrid fuzzy clustering-optimization approach to the operation of emergency logistics co-distribution responding to the urgent relief demands after natural disasters [4].

When a major disaster event occurs, the demand for relief supplies of emergency demand points will exceed the supplies amount of reserve points which can meet their requirements in limited time. Thus, emergency service points which can not reach demand points in limited time should also provide materials to those points. Therefore, the scheduling strategy proposed in this paper is that all emergency demand points should be provided some relief supplies, if can not fully meet the requirement, by their nearest service points in limited time in order to mitigate the consequences of disasters and carry out relief operations as soon as possible. The insufficient part of each demand points would be met after limited time. This strategy takes into account the needs of all emergency points, which not only reflect the fairness but also prevent further loss of disaster events.

The objective function of emergency scheduling model revolves time, distance and cost in many other study literatures. Actually, in initial period of disaster relief operation, saving life and minimizing losses rather than dispatch expense are the key point of consideration. For each emergency needs point, part of relief supplies should arrived in limited time. Under this premise, the goal is minimize the loss caused by delayed arrival. The factors affecting the loss are the period of delayed time and the number of unmet demand for relief supplies in limited time.

II. ASSUMPTIONS AND DEFINITIONS

A. notations

$S = \{S_i \mid i = 1, 2, \cdots, n\}$: the set of relief supplies service points in the region;

$F = \{F_j \mid j = 1, 2, \cdots, m\}$:the set of relief supplies demand points in the region;

t_{ij} : the travel time from point i to point j ;

t : the limited time of demand for relief supplies, which usually determined by the feature of relief supplies;

z_i : the backup amount of relief supplies in point i ;

r_j : the demand amount for relief supplies in point j ;

a_{ij} : the amount of relief supplies that point i provide to point j, which is the decision variable;

λ_{ij} : the 0-1 variable values according to whether the travel time from point i to point j exceed t ;

$$\lambda_{ij} = \begin{cases} 1 & t_{ij} > t \\ 0 & t_{ij} \le t \end{cases} \tag{1}$$

β : the penalty coefficient of disaster losses caused by per unit relief supplies when delayed per unit time to reach demand point; Obviously, when $t_{ij} \le t$, $\beta = 0$.

B. Assumptions

(i)It was assumed that the backup of all relief supplies service points can meet the need of all relief supplies demand points in the region, that is $\sum_{i=1}^{n} z_i \ge \sum_{j=1}^{m} r_j$.

If the total amount of all points i can not meet the needs of all points j, that is $\sum_{i=1}^{n} z_i < \sum_{j=1}^{m} r_j$. In this case, the relief supplies from outside is needed. Before their arrival, the internal available relief supplies must be used to emergency relief activities. We make variable r_j' to replace variable r_j .

$$r_j' = \frac{\sum_{i=1}^{n} z_i}{\sum_{j=1}^{m} r_j} \times r_j \tag{2}$$

$$\sum_{i=1}^{n} z_i = \sum_{j=1}^{m} r_j' \tag{3}$$

Another method is the conversion factor given by the field experts according to actual situation. δ_j is the needs conversion factor of emergency demand point j.

$$r_j' = \delta_j \cdot \sum_{i=1}^{n} z_i \tag{4}$$

$$\sum_{j=1}^{m} \delta_j = 1 \tag{5}$$

(ii) t_{ij} is known under the determined road network conditions, which can be obtained by the classical shortest path algorithm such as dijkstra algorithm.

It is assumed that for each emergency demand point j, there is at least one emergency service point i which can provide relief supplies to it in limited time.

(iii) β should be determined reasonably by the field experts according to the feature of relief supplies. When delay time of relief supplies close to human physiological extreme limit, β increase rapidly as Fig. 1.
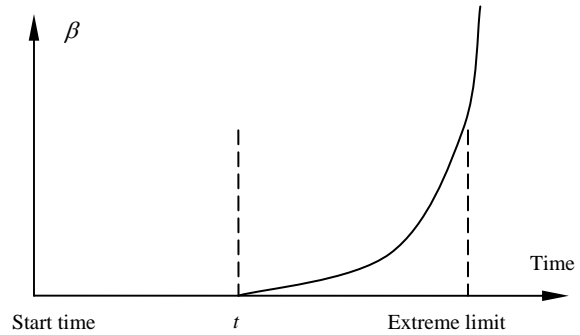

Figure 1.   Example of penalty factor of loss β

(iv)We think that emergency transportation vehicles available resources are abundant, by requisitioning social vehicles to ensure the implementation of scheduling decision. So the number restrictions and capacity restrictions of transportation vehicles did not take into account as constraints.

III. SCHEDULING OF RELIEF SUPPLIES DISTRIBUTION UNDER DETERMINED CONDITIONS

A. Scheduling Model of Relief Supplies Distribution for Single Species

The objective function of relief supplies scheduling model should be minimizing the disasters losses of all demand points due to delay, under the premise of ensure that each demand point has its own relief supplies in limited time. The factors affecting the loss are delay time and unmet demand of relief supplies in limited period. The objective function can express as follow.

Model I $\min \sum_{i=1}^{n} \sum_{j=1}^{m} \beta \cdot \lambda_{ij} \cdot (t_{ij} - t) \cdot a_{ij} \tag{6}$

Subject to:

(i) The total amount of all emergency service points can meet the needs of all emergency demand points.

$$\sum_{i=1}^{n} z_i \ge \sum_{j=1}^{m} r_j \tag{7}$$

(ii) The needs of all emergency demand points can be met.

$$\sum_{i=1}^{n} a_{ij} = r_j, \quad j = 1, 2, \cdots, m \tag{8}$$

(iii) The supplies number of each emergency service points can not exceed its reserve amount.

$$\sum_{j=1}^{m} a_{ij} \le z_i, \quad i = 1, 2, \cdots, n \tag{9}$$

(iv)Each demand point has its own relief supplies in limited time.

$$\sum_{i=1}^{n} (1 - \lambda_{ij}) \cdot a_{ij} > 0, \quad j = 1, 2, \cdots, m \tag{10}$$

(v) $z_i \geq 0$; $r_j \geq 0$; $a_{ij} \geq 0$; $\lambda_{ij} = \{0,1\}$; $\beta \geq 0$.

If delay time for each demand point is zero, the model has infinite solutions and objective function should changed to minimize total transportation time or cost.

B. Experimental Results of Model I

We assume that there are 5 emergency demand points and 10 emergency service points, and $t = 10$. The value of t_{ij}, r_j and z_i is showed in Table I. The value of β is different according to the value section of $t_{ij} - t$.

$$\beta = \begin{cases} 0 & t_{ij} - t \leq 0 \\ 1 & 0 < t_{ij} - t \leq 5 \\ 2 & 5 < t_{ij} - t \leq 10 \\ 10 & 10 < t_{ij} - t \leq 20 \\ 100 & 20 < t_{ij} - t \end{cases} \qquad (11)$$

The model is integer programming model which can be solved by using linear programming tool LINGO. The results are shown in Table II, and the objective function value is 225.

TABLE I.   INITIAL CONDITIONS OF EXPERIMENTAL OF MODEL I

t_{ij}	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	Demand
F_1	17	9	11	12	16	13	14	8	9	22	100
F_2	16	11	21	8	12	15	18	10	11	12	120
F_3	11	15	17	12	7	11	23	13	10	12	90
F_4	9	8	17	16	11	18	13	12	25	7	110
F_5	12	22	15	14	8	15	12	18	13	16	80
Reserve	40	50	55	45	60	60	40	45	65	40	500

TABLE II.   EXPERIMENTAL RESULTS OF MODEL I

a_{ij}	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	Demand
F_1		55							45		100
F_2		20		45				45	10		120
F_3					20	60			10		90
F_4	40	30								40	110
F_5					40		40				80
Reserve	40	50	55	45	60	60	40	45	65	40	500

C. Scheduling Mode of Relief Supplies Distribution for Many Varieties

There are many types of relief supplies, such as drinking water, foods, medicine and so on. When scheduling model of varieties relief supplies distribution is considered, k is defined as the species of relief supplies, $k=1, \ldots, l$.

z_i^k: the backup amount of k kind of relief supplies in point i;

r_j^k: the demand amount for k kind of relief supplies in point j;

a_{ij}: the amount of k kind of relief supplies that point i provide to point j, which is the decision variable;

t^k: the limited time of demand for k kind of relief supplies. For the same scheduling batch of relief supplies their restriction time is similar, and it is assumed that $t^1 = t^2 = \cdots = t^l = t$;

β^k: the penalty coefficient of disaster losses caused by per unit k kind of relief supplies when delayed per unit time to reach demand point;

Other variables defined as before. The objective function of scheduling model of many varieties relief

supplies is also minimizing the disasters losses of all demand points due to delay, and can express as follow.

Model II   min $\sum_{k=1}^{l} \sum_{i=1}^{n} \sum_{j=1}^{m} \beta^k \cdot \lambda_{ij} \cdot (t_{ij} - t) \cdot a_{ij}^k$ (12)

Subject to:

(i) The total amount of all emergency service points can meet the needs of all emergency demand points.

$$\sum_{i=1}^{n} z_i^k \geq \sum_{j=1}^{m} r_j^k, \quad k=1,2,\cdots,l \qquad (13)$$

(ii) The needs of every emergency demand points can be met.

$$\sum_{i=1}^{n} a_{ij}^k = r_j^k, \quad j=1,2,\cdots,m, \quad k=1,2,\cdots,l \qquad (14)$$

(iii) The supplies number of each emergency service points can not exceed its reserve amount.

$$\sum_{j=1}^{m} a_{ij}^k \leq z_i^k, \quad i=1,2,\cdots,n, \quad k=1,2,\cdots,l \qquad (15)$$

(iv) Each demand point has its own relief supplies reached in limited time.

$$\sum_{i=1}^{n} [(1-\lambda_{ij}) \cdot a_{ij}^k] > 0, \quad j=1,2,\cdots,m, \quad k=1,2,\cdots,l \qquad (16)$$

(v) $z_i^k \geq 0$; $r_j^k \geq 0$; $a_{ij}^k \geq 0$; $\lambda_{ij} = \{0,1\}$; $\beta^k \geq 0$.

D. Experimental Results of Model II

The solution of Model II is similar with the solution of Model I. For every k from 1 to l, a_{ij}^{k} can be solved. Then the result summarized is the solution of the model.

We assume that there are 5 emergency demand points and 10 emergency service points, and $t = 10$, $k = 3$. The value of t_{ij} is showed in Table III and the value of r_j^k

and z_i^k is showed in Table IV. It is assumed that the penalty coefficient of each kind of relief supplies is the same, so that $\beta^1 = \beta^2 = \cdots = \beta^l = \beta$. The value of β is showed in (11).

The results solved by using LINGO are shown in Table V, and the objective function values are 225(k=1), 205 (k=2) and 205 (k=3).

TABLE III.   INITIAL CONDITIONS(1) OF EXPERIMENTAL OF MODEL II

t_{ij}	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
F_1	17	9	11	12	16	13	14	8	9	22
F_2	16	11	21	8	12	15	18	10	11	12
F_3	11	15	17	12	7	11	23	13	10	12
F_4	9	8	17	16	11	18	13	12	25	7
F_5	12	22	15	14	8	15	12	18	13	16

TABLE IV.   INITIAL CONDITIONS(2) OF EXPERIMENTAL OF MODEL II

| k | Backup amount of service points | | | | | | | | | | Total |
	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	
1	40	50	55	45	60	60	40	45	65	40	500
2	55	60	50	65	55	45	50	60	60	50	550
3	45	45	40	40	55	50	45	40	55	35	450
k	Demand amount of demand points										Total
	F_1	F_2	F_3	F_4	F_5						
1	100	120	90	110	80						500
2	115	135	100	105	95						550
3	90	105	80	100	75						450

TABLE V.   EXPERIMENTAL RESULTS OF MODEL II

a_{ij}^{1}	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	Demand
F_1			55						45		100
F_2		20		45				45	10		120
F_3					20	60			10		90
F_4	40	30								40	110
F_5					40		40				80
Reserve	40	50	55	45	60	60	40	45	65	40	500
a_{ij}^{2}	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	Demand
F_1		50	50						15		115
F_2		10		65				60	45		135
F_3					10	45					100
F_4	55									50	105
F_5					45		50				95
Reserve	55	60	50	65	55	45	50	60	60	50	550
a_{ij}^{3}	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	Demand
F_1			40						50		90
F_2		25		40				40			105
F_3					25	50			5		80
F_4	45	20								35	100
F_5					30		45				75
Reserve	45	45	40	40	55	50	45	40	55	35	450

IV. OPTIMAL PATH IN FUZZY NETWORKS

Natural disaster events may cause road damage, so that the travel time between the emergency service points and emergency demand points is uncertain. If there is more than one path from emergency service points to emergency demand points, the travel time must be identified by choosing the optimal path in fuzzy network firstly. Symmetry triangular fuzzy number (STFN) was used to indicate the uncertainty in this paper.

A. Definition

In a given graphic network $G = \{V, E\}$, V is the set of points in G, and E is the set of arc connecting points in G. $\forall e \in E$, $w(e)$ is weight of e. R is the set of all links between point i and point j. If P is a route link any two point in G, $w(P) = \sum_{e \in P} w(e)$.

The weight of arc can be considered as travel time. The shortest route between point i and point j is route P_0 which weight is minimum, $w(P_0) = \min_P w(P)$.

t is limited time. $Q = \{P \mid w(P) \le t, P \in R\}$ is the set of route between point i and point j which completely satisfied t. If $w(e)$ is STFN, $w(P)$ is also STFN.

The satisfaction function $F(P, t)$ was defined as the degree of satisfaction that travel time through P does not exceed t.

$S = \{P \mid \max F(P, t), P \in R\}$ is the set of optimal path satisfied t in maximum degree.

Seeking optimal path is finding P^*, $P^* \in S$.

χ is the set of STFN.

$T(M \le t)$ means possibility of $M \le t$.

$$T(\{M \le t\}) = \begin{cases} 0 & a > t \\ 2\left(\dfrac{t-a}{c-a}\right)^2 & a \le t < b \\ 1 - 2\left(\dfrac{c-t}{c-a}\right)^2 & b \le t < c \\ 1 & c \le t \end{cases} \quad (17)$$

If $M = [a, b, c] \in \chi$ and $N = [d, e, f] \in \chi$, $b = (a+c)/2$, $M + N = [a+d, b+e, c+f] \in \chi$.

If $a \le t < c$ and $d \le t < f$, the sufficient and necessary condition of $T(\{M \le t\}) \le T(\{N \le t\})$ is $\dfrac{t-a}{c-a} \le \dfrac{t-d}{f-d}$.

Due to $w(e) \in \chi$, M$=[w_1(e), w_0(e), w_2(e)]$ and $w_0(e) = (w_1(e) + w_2(e))/2$.

Obviously,
$$w(P) = \sum_{e \in P} w(e) = [w_1(P), w_0(P), w_2(P)]$$
$$= [\sum_{e \in P} w_1(e), \sum_{e \in P} w_0(e), \sum_{e \in P} w_2(e)] \quad (18)$$

B. Transformation of Maximum Satisfaction Factor Path

Let
$$F(P, t) = T(\{w(P) \le t\}) = T(\{M \le t\}), 0 \le F(P, t) \le 1.$$

If
$$w_2(P_0) = \min_{P \in R} w_2(P) = \min_{P \in R} \left(\sum_{e \in P} w_2(e)\right) \le t \quad (19)$$
then $F(P_0, t) = 1$, $P_0 \in S$.

If
$$w_1(P_0) = \min_{P \in R} w_1(P) = \min_{P \in R} \left(\sum_{e \in P} w_1(e)\right) > t \quad (20)$$
then $F(P_0, t) = 0$.

When (19) and (20) is tenable, maximum satisfaction factor path can be obtained with shortest path algorithm.

If $\min_{P \in R}(\sum_{e \in P} w_1(e)) \le t < \min_{P \in R}(\sum_{e \in P} w_2(e))$, let P^* is solution of
$$\min_{P \in R} \frac{w_2(P) - t}{w_2(P) - w_1(P)} = \frac{w_2(P^*) - t}{w_2(P^*) - w_1(P^*)} \quad (21)$$

Then $F(P^*, t) = \max_{P \in R} F(P, t)$ [5].

Equivalent form of (21) is
$$\max_{P \in R} \frac{t - w_2(P)}{w_2(P) - w_1(P)} = \frac{t - w_2(P^*)}{w_2(P^*) - w_1(P^*)} = x^* \quad (22)$$

Where $x^* \in [-1, 0]$, and
$$F(P^*, t) = \begin{cases} 1 - 2x^{*2} & x^* \ge -0.5 \\ 2(1 + x^*)^2 & x^* < -0.5 \end{cases}$$

For $x \in [-1, 0]$, let $P(x)$ is corresponding path of optimum solution of the follow problem. $P(x)$ can be obtained with shortest path algorithm.

$$\min_{P \in R} \sum_{e \in P} ((w_2(e) - w_1(e)) \cdot x + w_2(e)) = \sum_{e \in P(x)} ((w_2(e) - w_1(e)) \cdot x + w_2(e))$$

Let $N(x) = \sum_{e \in P} ((w_2(e) - w_1(e)) \cdot x + w_2(e))$.

$N(x)$ is a continuous increasing function. The sufficient and necessary condition of $N(x) > t$, $N(x) = t$ and $N(x) < t$ are $x < x^*$, $x = x^*$ and $x < x^*$.

When $N(x) = t$, $P(x)$ is optimal path that satisfy (22), $x = x^*$, and vice versa.

C. Solving Step and Example

When (19) and (20) does not hold, then $N(0) > t$ and $N(-1) \le t$. Dichotomy can be used to solving the problem.

Steps 1: let $x = 0$, if $N(0) \le t$, then $P(0)$ is optimal path, $P(0) \in S$, otherwise turn to steps 2.

Steps 2: let $x = -1$, if $N(-1) > t$, then $P(-1)$ is optimal path, $P(-1) \in S$, otherwise turn to steps 3.

Steps 3: let $x' = -1$ and $x'' = 0$, turn to steps 4.

Steps 4: let $x = [x' + x'']/2$, if $|N(x) - t| < \varepsilon$, then $P(x)$ is optimal path, $P(x) \in S$, calculation stopped. Otherwise turn to steps 5.

Steps 5: if $N(x) > t$, then let $x'' = x$, otherwise $x' = x$. Turn to steps 2.

We assume that the network between point S and point F is shown as Fig. 2 and $t = 20$. The travel time between points is STFN. By using the above algorithm, we get the results: maximum satisfaction factor path is S→A2→A5

→A8→F; satisfaction factor is 0.9688; travel time is [13,17,21].
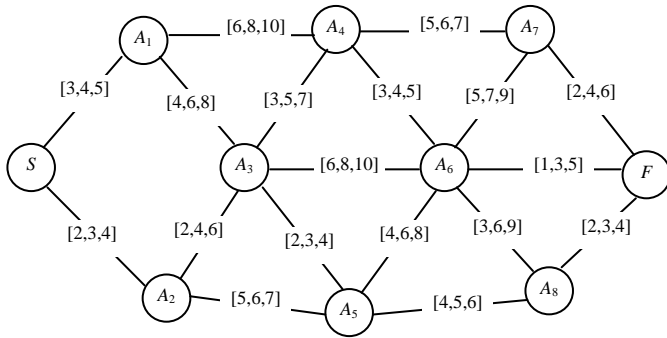

Figure 2.   Example of maximum satisfaction factor path

V. SCHEDULING OF RELIEF SUPPLIES DISTRIBUTION UNDER FUZZY CONDITIONS

A. Scheduling Model of Relief Supplies Distribution Based on Fuzzy Travel Time

$\tilde{t}_{ij} = [t_{ij}^{1}, t_{ij}^{0}, t_{ij}^{2}]$: the fuzzy travel time from emergency service point i to emergency demand point j, $t_{ij}^{0} = (t_{ij}^{1} + t_{ij}^{2})/2$;

$F(\tilde{t}_{ij}, t)$: the probability of event $\{\tilde{t}_{ij} \leq t\}$;

$$F(\tilde{t}_{ij}, t) = T(\{\tilde{t}_{ij} \leq t\}) = \begin{cases} 0 & t_{ij}^{1} > t \\ 2\left(\dfrac{t - t_{ij}^{1}}{t_{ij}^{2} - t_{ij}^{1}}\right)^2 & t_{ij}^{1} \leq t < t_{ij}^{0} \\ 1 - 2\left(\dfrac{t_{ij}^{2} - t}{t_{ij}^{2} - t_{ij}^{1}}\right)^2 & t_{ij}^{0} \leq t < t_{ij}^{2} \\ 1 & t_{ij}^{2} \leq t \end{cases} \quad (23)$$

λ_{ij}: whether the travel time from emergency service point i to emergency demand point j exceed t ;

$$\lambda_{ij} = \begin{cases} 0 & F(\tilde{t}_{ij}, t) = 1 \\ 1 & 0 \leq F(\tilde{t}_{ij}, t) < 1 \end{cases} \quad (24)$$

Other variables defined as before. The objective function of relief supplies scheduling model should be minimizing the disasters losses of all demand points due to delay, under the premise of ensure that each demand point has its own relief supplies in limited time. The factors affecting the loss are delay time and unmet demand of relief supplies in limited period. The objective function can express as follow:

Model III $\min \sum_{i=1}^{n} \sum_{j=1}^{m} \beta[1 - F(\tilde{t}_{ij}, t)](t_{ij}^{2} - t)a_{ij}$ (25)

Subject to:

(1) The total amount of all emergency service points can meet the needs of all emergency demand points.

$$\sum_{i=1}^{n} z_i \geq \sum_{j=1}^{m} r_j$$

(2) The needs of all emergency demand points can be met.

$$\sum_{i=1}^{n} a_{ij} = r_j, \quad j = 1, 2, \cdots, m$$

(3) The supplies number of each emergency service points can not exceed its reserve amount.

$$\sum_{j=1}^{m} a_{ij} \leq z_i, \quad i = 1, 2, \cdots, n$$

(4) Each demand point has its own relief supplies in limited time.

$$\sum_{i=1}^{n} (1 - \lambda_{ij}) \cdot a_{ij} > 0, \quad j = 1, 2, \cdots, m$$

(5) $z_i \geq 0$; $r_j \geq 0$; $a_{ij} \geq 0$; $\lambda_{ij} = 0,1$; $\beta \geq 0$.

B. Experimental Results

We assume that there are 5 emergency demand points and 10 emergency service points, and $t = 10$. The value of $\tilde{t}_{ij}$, r_j and z_i is showed in Table VI. The value of β is different according to the value section of $t_{ij}^{2} - t$.

$$\beta = \begin{cases} 0 & t_{ij}^{2} - t \leq 0 \\ 1 & 0 < t_{ij}^{2} - t \leq 5 \\ 2 & 5 < t_{ij}^{2} - t \leq 10 \\ 10 & 10 < t_{ij}^{2} - t \leq 20 \\ 100 & 20 < t_{ij}^{2} - t \end{cases}$$

Firstly, according to (23) $F(\tilde{t}_{ij}, t)$ can be calculated. Then the model is solved by using linear programming tool LINGO. The results are shown in Table VII, and the objective function value is 578.125.

VI. CONCLUSIONS AND FUTURE WORK

Considering the more demand number for emergency supplies and time-sensitive situation, scheduling model of relief supplies with the goal of minimum disaster losses was established and solved under different conditions. In order to achieve the goal, the requirements for relief supplies of each emergency demand points should be met in supplies the aspect of time and quantity which means the required supplies must be delivered to the demand points in limited time. The scheduling strategy proposed

TABLE VI.   INITIAL CONDITIONS OF EXPERIMENTAL OF MODEL III

t_{ij}	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	Demand
F_1	[14,17,20]	[8,9,10]	[9,11,13]	[10,12,14]	[13,16,19]	[11,13,15]	[12,14,16]	[7,8,9]	[8,9,10]	[19,22,25]	100
F_2	[14,16,18]	[10,11,12]	[18,21,24]	[7,8,9]	[11,12,13]	[13,15,17]	[16,18,20]	[8,10,12]	[9,11,13]	[11,12,13]	120
F_3	[9,11,13]	[13,15,17]	[15,17,19]	[11,12,13]	[6,7,8]	[9,11,13]	[20,23,26]	[12,13,14]	[9,10,11]	[10,12,14]	90
F_4	[8,9,10]	[6,8,10]	[15,17,19]	[15,16,17]	[9,11,13]	[15,18,21]	[11,13,15]	[10,12,14]	[21,25,29]	[5,7,9]	110
F_5	[10,12,14]	[19,22,25]	[13,15,17]	[12,14,16]	[7,8,9]	[14,15,16]	[10,12,14]	[17,18,19]	[12,13,14]	[14,16,18]	80
Reserve	40	50	55	45	60	60	40	45	65	40	500

TABLE VII.   EXPERIMENTAL RESULTS OF MODEL III

a_{ij}	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	Demand
F_1			55						45		100
F_2		20		45				45	10		120
F_3					20	60			10		90
F_4	40	30								40	110
F_5					40		40				80
Reserve	40	50	55	45	60	60	40	45	65	40	500

in this paper take into account the needs of all emergency demand points which can effectively prevent further expansion of the disaster loss. We are interested in how to determine the value of penalty factor β for further research by the scientific and rational way.

APPENDIX LINGO CODE OF EXPERIMENTAL EXAMPLE OF SOLVEING MODEL III

```
model:
  sets:
  warehouses/wh1..wh10/: capacity;
  customers/v1..v5/: demand;
  links(warehouses,customers): trans_time1, trans_time0, trans_time2, volume, beta, ft;
  endsets
  data:
  capacity=40 50 55 45 60 60 40 45 65 40;
  demand=100 120 90 110 80;
  trans_time1=14 14 9 8 10
            8 10 13 6 19
            9 18 15 15 13
            10 7 11 15 12
            13 11 6 9 7
            11 13 9 15 14
            12 16 20 11 10
            7 8 12 10 17
            8 9 9 21 12
            19 11 10 5 14;
  trans_time0= 17 16 11 9 12
            9 11 15 8 22
            11 21 17 17 15
            12 8 12 16 14
            16 12 7 11 8
            13 15 11 18 15
            14 18 23 13 12
            8 10 13 12 18
            9 11 10 25 13
            22 12 12 7 16;
  trans_time2=20 18 13 10 14
            10 12 17 10 25
            13 24 19 19 17
            14 9 13 17 16
            19 13 8 13 9
            15 17 13 21 16
            16 20 26 15 14
            9 12 14 14 19
            10 13 11 29 14
            25 13 14 9 18;
  limited_time=10;
  enddata
```

```
  min=@sum(links:    beta*(1-ft)*(trans_time2-limited_time) *volume);
    @for(links(I,J):
      beta=@IF(trans_time2-limited_time#le#  0,  0,
@IF((  trans_time2-limited_time  #gt#  0)  #and#
(trans_time2-limited_time #le# 5), 1, @IF(( trans_time2-
limited_time #gt# 5) #and# (trans_time2-limited_time
#le# 10), 2, @IF((trans_time2-limited_time #gt# 10)
#and# (trans_time2-limited_time #le# 20),10,100)))));
    @for(links(I,J):
      ft=@IF(limited_time    #lt#    trans_time1,    0,
@IF((limited_time    #ge#    trans_time1)    #and#
(limited_time  #lt#  trans_time0),  2*((limited_time-
trans_time1)/(trans_time2-trans_time1))^2,
@IF((limited_time  #ge#  trans_time0)  #and#
(limited_time  #lt#  trans_time2), 1-2*((trans_time2-
limited_time)/(trans_time2-trans_time1))^2,1))));
    @for(customers(J):
    @sum(warehouses(I): volume(I,J))=demand(J));
    @for(warehouses(I):
    @sum(customers(J): volume(I,J))<=capacity(I));
  end
```

REFERENCES

[1] Ali Haghan, Sei-Chang Oh, "Formulation and solution of a multi-commodity, multi-modal network flow model for disaster relief operations," Transport Research, vol. 30, pp. 231–250, May 1996.

[2] F.Fiedrich, F.Gehbauer and U.Rickers, "Optimized resource allocation for emergency response after earthquake disasters," Safety Science, vol. 35, pp. 41–57, June 2000.

[3] Linet Ozdama, Ediz Ekinei, "Emergency logistics planning in nature disasters," Annals of Operations Research, ,vol. 129, pp.217–245, July 2004.

[4] J.B. Sheu, "An emergency logistics distribution approach for quick response to urgent relief demand in disasters," Transportation Research Part E: Logistics and Transportation Review, vol. 43, pp. 687-709, November 2007.

[5] C.L. Liu, J.M. He and Z.H. Sheng, "Selecting of the maximum satisfaction factor path in fuzzy enmergency nerwork systems," Acta Automatica Sinica. China, vol. 26, pp. 609-615, May 2000.(in Chinese)

[6] Xiang-Rong Huang, Ru-He Xie, "A Model on Location Decision for Distribution Centers of Emergency Food Logistics," 2009 Second International Conference on Information and Computing Science, icic, vol. 4, pp.232-235, 2009

Dynamic Real-time Optimization of Reservoir Production

Kai Zhang, Jun Yao, Liming Zhang, Yajun Li
College of Petroleum Engineering
China University of Petroleum (East China)
QingDao, Shandong, China 266555
E-mail: reservoirs@163.com

Abstract—**Water flooding is a common technology of enhanced oil recovery and has been widely used to the world's oil fields. But it is difficult to maintain a uniform displacement of water, which causes the ineffective circulation of injected water. Therefore, the real-time dynamic control strategy is studied to work out development plans scientifically and make better economic profits in this paper. The net present value of crude oil exploitation during a certain period of time is choosen to performance index of the control problem. Hydrodynamic equations of oil, water and gas, which describe the displacement process, are used as governed equations. Meanwhile, the boundary constraints of wells and balance relationship of injection-production are taken into account. Based on the above equations, a control model of oil field development is established. The control vector parameterization method is illustrated in detail to deal with the unconstrained, control constrained and state constrained optimal control problem. Finally, a simulation example is employed to verify the validity of the proposed control algorithm.**

Index Terms—**optimal control, reservoir simulation, gradient method, smart field**

I. INTRODUCTION

The world's main onshore oil and gas fields in the basin have been developed for a long time. Most of them are reaching a mid or late stage of development. One of the main problems is non-equilibrium of water flooding, which causes the complex distribution of remaining oil and the difficult production adjustment. The space of increasing reserve and enhancing production is very limited.

According to simulation and analysis results, the problem is mainly resulted in by interlayer rhythm, reservoir plane heterogeneity and different well patterns etc. They are the reasons why a large number of inefficient and ineffective circulation wells exit in real oil fields too. For example, Gudong reservoir in China has 1650 production wells, 620 of which are located in the main stream channels. High capacity channel occurs in the center of these main stream channels. So water flooding spread unevenly in the area (shown in Fig.1). Meanwhile, the layer heterogeneity (shown in Fig.2)

makes the big difference of vertical flooding and lead to severe interlayer interference.

In order to produce more oil, the traditional method aiming at this problem is to do a lot of tests (orthogonal test design and reservoir simulation) and compare results to decide the optimal development plan. The plan could minimize the impact of heterogeneity as much as possible. However, its shortcoming is very obvious that lots of time and manpower are expended. What's more, the final obtained result is always limited in a small scope of original designs. Dynamic real-time optimization method of reservoir production is proposed for the problems. It combines reservoir numerical simulation and optimization method, and directly solves gradients of control variables to achieve optimal control settings.
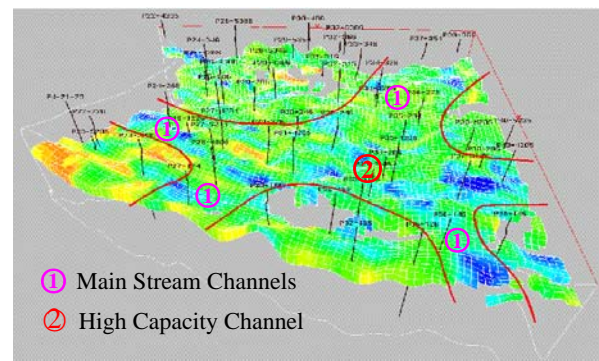
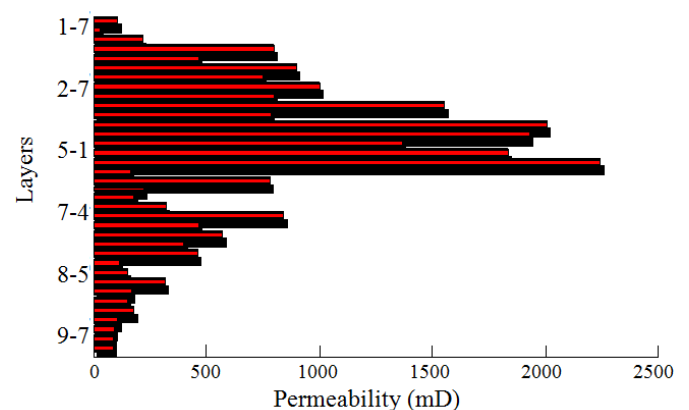Figure 1.   High capacity and main stream channels in Gudong reservoir

Figure 2.   The layer heterogeneity of permeability

There is only a few and not enough relevant research carried out in the early 20th century, which are simple applications of optimization theory (such as the energy equation of allocation[1], multi-objective optimization of oil field development[2], etc). But with the proposal of smart field system, many oil companies and research universities have been studying production optimization since 2005.

In 2006, the method combining empirical formula with optimization theory is proposed by Saputelli[3,4], Elgsaeter[5], and Awasthi[6] and etc. Their goal focused on the short-term optimization of production and gained the optimized development program in time with convenient empirical formula. The advantages of this method are the certain universality in optimization results and high speed. But its disadvantages are also obvious that it is poor targeted, such as it can't consider the heterogenetic characteristic of formation and realize production optimization control of single-well.

Alghareeb[7] calculated the optimal flow control valve settings of intelligent well with genetic algorithm. Chunhong[8] optimized production in small-scale reservoir with SPSA method. Lorentzen[9] and Yan[10] simplified the calculation of production optimization with EnKF and other methods. These studies achieved certain results, which can search the global optimal solution of optimization production problem. But low-speed computation is the biggest drawback of the method in large-scale reservoir model, so this kind of method is difficult to be applied to the oil field.

Brouwer[11] and Chunhong[8] studied production optimization problems with gradient algorithms. But their algorithms still have some limitations. Chunhong used the perturbation gradient algorithm. In her method, each time only one gradient of control variable in one time step could be calculated though two simulation runs, in which the speed basically can't meet the requirements of engineering applications. But such problems always need a large amount of calculations and the running speed is first and foremost premise. Therefore, compared to stochastic and empirical algorithms, the gradient algorithm is more advisable.

II. MATHEMATICAL MODEL OF OPTIMAL CONTROL FOR RESERVOIR PRODUCTION WORKING SYSTEM

Reservoir development optimization could maximize oil production by adjusting control parameters of production and injection wells in a reservoir, which belongs to the scope of optimal control problems. This issue starts from an economic point of view and the working system is optimized to improve production capacity and increase oil recovery efficiency[12, 13]. Compared with conventional simulation methods, the method's greatest advantage is to automatically obtain optimal programs at every time under the principle of stabilizing oil production and controlling water cut. Relying on the method, production figures go up not adopting any measures(acid, fracture, EOR etc.) . That is, costs would come down.

A. Performance indicator of optimal control

Reservoir production optimization issues need a cost function aiming at the actual working system as performance indicator of optimal control. Different performance indicator will get different optimization result. The actual situation in oil field must be considered to choose appropriate performance indicator. Generally, oil industry can be regarded as a business of high-input and high-return. Oil exploitation is closely linked with economic factors and its main purpose is to make maximum profits. So the net present value(NPV) is used as the cost function in this paper, which is given by the following equation (see Nomenclature for the definitions of the symbols)

$$H_{p,l}\left(u_p\right)=\frac{\left(1+\beta\right)^{t_{p,l}}}{\Delta t_{p,l}}\left[\sum_{n=1}^{N_P}\left(m_o q_{p,o,n}-m_w q_{p,w,n}\right)-\sum_{j=1}^{N_I}m_{iw}q_{p,iw,n}\right]$$

(1)

$$q=\sum_{n=1}^{N_P}\mathrm{WI}_n\left[\frac{\rho_{ps,n}K_{ps,n}X_{ps,n}}{\mu_{ps,n}}\left(P_{ps,n}-P_{cgo,n}-P_{wf,n}\right)\right]$$

(2)

where, H is NPV, ¥ ; p is control step; l is simulation step in each control step; ps is fluid phase(oil, water and gas); q is flow rate, m³; $P_{ps,n}$ is pressure in grid block, MPa; $P_{cgo,n}$ is capillary pressure, MPa; $P_{wf,n}$ is bottom hole pressure(BHP) of wells, MPa; u_p is control variable(such as P_{wf}, q_o, q_w and q_{wi}); m_o is oil price per cubic meter; m_w is water production cost per cubic meter; m_{iw} is water injection cost per cubic meter. WI_n is well index in the n layer. $\rho_{ps,n}$ is density of ps phase, kg/m³; $K_{ps,n}$ is permeability of ps phase,μm²; $X_{ps,n}$ is compositional factor; $\mu_{ps,n}$ is viscosity of ps phase, mPa·s. Δt is time period, year; b is discounting factor. N_P is number of production wells; N_I is number of injection wells.

B. Reservoir numerical simulation

Optimization calculation is based on flow system of oil and water. Usually, this system is described by a theory of reservoir numerical simulation. The goal of reservoir simulation is to optimize and simulate the development program in the future on the basis of repeating the whole process of oil field development in numerical way. Optimization here refers to better result relatively in a small scope of original designs. Researcher would compare several design plans and get relative optimal plan in general.

This paper combine optimization theory and reservoir simulation using black oil model. This model must satisfy several assumptions in production phase as follows: 1) reservoir exists oil-water-gas three-phase flow; 2) rock and fluid can be compressed; 3) fluid flow obeys Darcy's

law in reservoir; 4) rock is anisotropic and heterogeneous; 5) the effects of gravity and capillary pressure are considered in the flow.

Black oil model can be described as follows[14]

$$\psi = \nabla \left[\omega_{ps,g} \rho_g \frac{kk_{rg}}{\mu_g} \nabla \left(p_g - \rho_g gD \right) \right.$$

$$+ \omega_{ps,o} \rho_o \frac{kk_{ro}}{\mu_o} \nabla \left(p_o - \rho_o gD \right) + \omega_{ps,w} \rho_w \frac{kk_{rw}}{\mu_w} \nabla \left(p_w - \rho_w gD \right) \right]$$

$$+ q_{ps} - \frac{\partial}{\partial t} \left[\varphi \left(\omega_{ps,g} \rho_g S_g + \omega_{ps,o} \rho_o S_o + \omega_{ps,w} \rho_w S_w \right) \right] = 0$$

(3)

In order to solve this equation, some assistant equations must be provided:

①Saturation normalized equation

$$S_o + S_w + S_g = 1 \tag{4}$$

②Mass fraction normalized equations

$$\sum_{i=1}^{N} \omega_{ps,g} = 1 \quad ; \quad \sum_{i=1}^{N} \omega_{ps,o} = 1 \quad ; \quad \sum_{i=1}^{N} \omega_{ps,w} = 1 \tag{5}$$

③Equilibrium constant equation

$$K_{ps,go} = \frac{\omega_{ps,g}}{\omega_{ps,o}} \quad ; \quad K_{ps,gw} = \frac{\omega_{ps,g}}{\omega_{ps,w}} \tag{6}$$

④Capillary pressure equation

$$p_{cgo} = p_g - p_o \; ; \quad p_{cow} = p_o - p_w \tag{7}$$

Initial condions are

$$p_o \big|_{t=0} = p_{oi} \; ; \quad S_w \big|_{t=0} = S_{wi} \tag{8}$$

Closed outer boundary conditions is

$$\frac{\partial P}{\partial n_r} \bigg|_{\Gamma=0} = 0 \tag{9}$$

Outer boundary conditions of specified pressure is

$$P \big|_{\Gamma} = 0 \tag{10}$$

where, ψ is flow equation. k_r is relative permeability, μm^2; D is height difference, m;

Flow is simulated by Eqn.(3) to (10). State variables like pressure and saturation can be solved for optimal control model. This forward model uses differential solution.

C. *Mathematical Model of Optimal Control*

According to the general form of optimal control model, the model for reservoir production working system is expressed as follows

$$\max \quad C = \sum_{p=0}^{N-1} \sum_{l=0}^{N_m-1} H_{p,l} \left(x_{p,l+1}, u_p \right) \tag{11}$$

subject to

$$\psi_{p,l} \left(x_{p,l+1}, x_{p,l}, u_p \right) = 0 \tag{12}$$

$$u_{p,low} \leq u_p \leq u_{p,up} \tag{13}$$

where, C is objective function of optimization and summation of performance indicator H in every period. x is state variable(pressure and saturation in each grid block).

The problem can be described as follows: under the constraints of reservoir flow equation $\psi_{p,l} = 0$ and single-well production limit ($u_{p,low} \leq u_p \leq u_{p,up}$), mathematical model is solved to achieve optimal control variables $x*$ and corresponding optimal states $u*(t)$.

III.   GRADIENTS CALCULATION

It was stated earlier that the gradients of the cost function with respect to the controls could be calculated very efficiently using maximum principle. The gradient equations can be obtained from the necessary conditions of optimality of the optimization problem defined by Eq.(11)~(13). These necessary conditions of optimality are derived from the classical theory of calculus of variations. For a relatively simple treatment of this subject, refer to Stengel[15]. A more detailed and rigorous analysis of the problem and generalization to infinite dimensional problems in arbitrary vector spaces is given by reference [16]. The essence of the theory is that the cost function of (11) along with all the constraints can be written equivalently in the form of an augmented cost function given by Eq.(14).

$$C_A = \sum_{p=0}^{N-1} \sum_{l=0}^{N_m-1} H_{p,l} \left(x_{p,l+1}, u_p \right) + \sum_{m=0}^{N-1} \sum_{n=0}^{N_m-1} \left(\lambda_{p,l+1} \right)^T \psi_{p,l} \left(x_{p,l+1}, x_{p,l}, u_p \right)$$

(14)

The vector λ is known as Lagrange multiplier, which can be thought of as elements of the dual space of the vector space to which u belongs. One Lagrange multiplier is required for each constraint with which the cost function is augmented. That is, the total number of Lagrange multipliers is equal to the product of the number of dynamic states and control steps.

For optimality of the augmented cost function, the first variation of the function must equals to zero, which is given by:

$$\delta C_A = \sum_{p=0}^{N-1} \sum_{l=0}^{N_m-1} \left(\frac{\partial C_{A,p,l}}{\partial x_{p,l}} \right) \delta x_{p,l} + \sum_{p=0}^{N-1} \sum_{l=0}^{N_m-1} \left(\frac{\partial C_{A,p,l}}{\partial x_{p,l+1}} \right) \delta x_{p,l+1} + \sum_{p=0}^{N-1} \sum_{l=0}^{N_m-1} \left(\frac{\partial C_{A,p,l}}{\partial u_p} \right) \delta u_p$$

(15)

Eq.(15) is extended as follows

$$\delta C_A = \sum_{p=0}^{N-1} \sum_{l=0}^{N_m-1} \left(\frac{\partial C_{A,p,l-1}}{\partial x_{p,l}} + \frac{\partial C_{A,p,l}}{\partial x_{p,l}} \right) \delta x_{p,l}$$

$$+ \sum_{p=0}^{N-1} \sum_{l=0}^{N_m-1} \left(\frac{\partial C_{A,p,l}}{\partial u_p} \right) \delta u_p + \sum_{p=0}^{N-1} \sum_{l=0}^{N_m-1} \left(\frac{\partial C_{A,N-1,N_m-1}}{\partial x_{N-1,N_m-1}} \right) \delta x_{N-1,N_m-1}$$

(16)

Because these items in Eq.(16) are independent of each other, according to the necessary conditions getting extremum is $\delta C_A = 0$, (16) are changed as follows

$$\sum_{p=0}^{N-1} \sum_{l=0}^{N_m-1} \left(\frac{\partial C_{A,p,l-1}}{\partial x_{p,l}} + \frac{\partial C_{A,p,l}}{\partial x_{p,l}} \right) = 0$$

$$\sum_{p=0}^{N-1} \sum_{l=0}^{N_m-1} \left(\frac{\partial C_{A,N-1,N_m-1}}{\partial x_{N-1,N_m-1}} \right) = 0, \quad \sum_{p=0}^{N-1} \sum_{l=0}^{N_m-1} \left(\frac{\partial C_{A,p,l}}{\partial u_p} \right) = 0$$

(17)

Each item are extended, (16) is expressed in the following form

$$\begin{cases} \dfrac{\partial H_{p,l-1}}{\partial x_{p,l}} + \left(\lambda_{p,l+1} \right)^T \dfrac{\partial \psi_{p,l}}{\partial x_{p,l}} + \left(\lambda_{p,l} \right)^T \dfrac{\partial \psi_{p,l-1}}{\partial x_{p,l}} = 0 \\[2mm] \dfrac{\partial H_{N-1,N_m-1}}{\partial x_{N-1,N_m}} + \left(\lambda_{N-1,N_m} \right)^T \dfrac{\partial \psi_{N-1,N_m-1}}{\partial x_{N-1,N_m}} = 0 \\[2mm] \dfrac{\partial C_{A,p,l}}{\partial u_p} = \dfrac{\partial H_{p,l}}{\partial u_p} + \left(\lambda^{m,n+1} \right)^T \dfrac{\partial \psi_{p,l}}{\partial u_p} \end{cases}$$

(18)

λ can be calculated at each time step and grid block by the first two equations(state equations) in (16). And then λ are using in the third equation(gradient equations) to obtain gradients $\partial C_{A,p,l} / \partial u_p$ of the various control variables. When gradients tends to zero through iterative calculation, the optimal solution is found. That is, the optimal control plan for reservoir production working system is ascertained.

IV. CONSTRAINED OPTIMIZATION

Oilfield development is based on the real reservoir. According to the geological conditions and the constraint of the development facilities and environment, the oil/water wells always produce under certain constraint condition. For example, there is the maximum and minimum value for the production of single well and bottom hole pressure. According to the reservoir pressure, the single well can not produce infinitely and the bottom hole pressure can not reduce to zero. So there is always a production limit for the single well. But for the entire block, because the reservoir controls optimization, on one hand is to compare the optimization effect, on the other hand is to maintain balanced injection and production rate combined with the real reservoir development, the

general constraint (such as the total injection rate of block)is needed for injection/production condition of the whole reservoir. The much more convenient method of constraint control is discussed combined with the above problems.
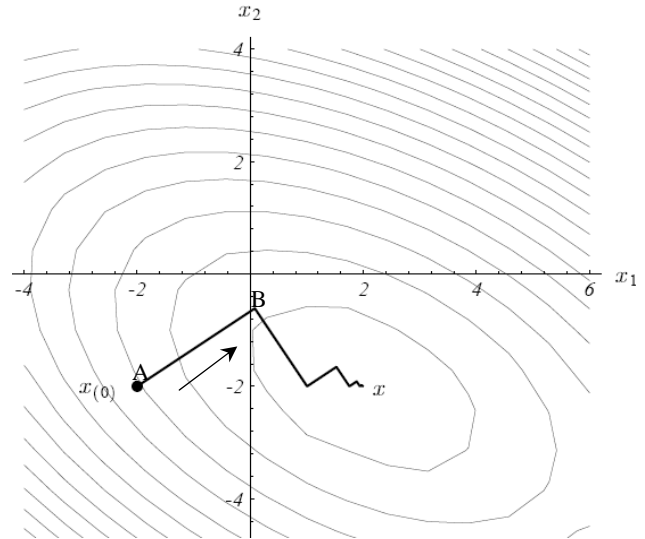
Figure 3. Searching schematic diagram of the steepest descent algorithm

The reservoir simulation calculation pertains to the area of large-scale mathematical computation. In the computation process, if considering the constraint conditions, such as the production limit of single well and the total injection rate of block, then the optimization pertains to large-scale constrained optimization problems. The gradient computational method was used in this article, but the gradient is only scalar field gradient. In the scalar field, the gradient of control variables points to the fastest increasing direction of the scalar field. The length of gradient is the largest rate of change. As shown in Fig. 3, the gradient direction is from point A to B and the distance between A and B is the search step size in the calculated iteration. Constrained optimization method is used to restrict the gradients.

When optimal control would be executed, the control variables' gradients can be obtained through the adjoint model. The sign of gradients may be positive or negative. For bottom hole pressure controlled by the single well, if the gradient is positive, it indicates that the bottom hole pressure should increase towards the direction of control variables rising (or increasing) in well optimization program; If the gradient is negative, it should be regulated towards the corresponding descent direction of the bottom hole pressure. For the block optimization, if the injection and production rate can not be balanced, free optimization is inconsistent with the real situation, and if the production is based on regulation scheme after optimizing in the forecast period, the injection flow rate of the entire block increases or decreases, and the same fluctuation to the corresponding crude oil and water production. The schemes obtained before and after optimization are not comparable and the real results of optimization can not be reflected because of the different

total injection rate. Therefore, the injection balance of the optimized reservoir block must be considered in the entire calculation progress, and the injection well is controlled by injection rate. Combined with the maximum and minimum values of single well-controlled constraints, the optimal control model becomes

$$\max \quad J = \sum_{m=0}^{N-1} \sum_{n=0}^{N_m-1} Y^{m,n}\left(x^{m,n+1}, u^m\right) \tag{19}$$

Constraint condition:

$$L^{m,n}\left(x^{m,n+1}, x^{m,n}, u^m\right) = 0,$$

$$\forall n \in \left(0, \cdots, N_m - 1\right), m \in \left(0, \cdots, N-1\right) \tag{20}$$

$$\sum_{i=1}^{N_{inj}} Q_{wi,i}^m = Q_t, \forall i \in \left(1, 2, \cdots, N_{inj}\right) \tag{21}$$

$$LB \le Q_{wi,i}^m \le UB, \forall i \in \left(1, 2, \cdots, N_{inj}\right) \tag{22}$$

$$LB \le P_{wf,k}^m \le UB, \forall k \in \left(1, 2, \cdots, N_{oil}\right) \tag{23}$$

In the formula, $Q_{wi,i}^m$ is the injection rate of the single well i in m control step, m³/day; $P_{wf,k}^m$ is the bottom hole pressure of the single production well k in m control step, m³/day; Q_t is the constraint total injection rate, m³/day, the total injection rate is constant all the times; N_{inj} is the total number of injection wells; N_{pro} is the total number of production wells.

The adjoint gradient $\dfrac{\partial J_A}{\partial u}$ obtained by previous solution is solved by Eq.(20) and (21). Considering the linear constraint of Eq.(22), a simple average gradient calculation method is used to restrict. First, solve the average gradient of all the water wells in adjoint model

$$Ga = \frac{1}{N_{inj}} \sum_{i=1}^{N_{inj}} \frac{\partial J^m}{\partial Q_{wi,i}^m} \tag{24}$$

In the formula, Ga is the accumulated gradient item.

Generally, the fluid in the reservoir should flow following the direction of oil/water wells in optimized process which is more favorable for the production. So based on this norm, if the single well calculated control gradient $\dfrac{\partial J^m}{\partial Q_{wi,i}^m}$ is greater than the average gradient value Ga in one time step, then the gradient is considered favorable; if $\dfrac{\partial J^m}{\partial Q_{wi,i}^m}$ is less than the average gradient value Ga, then the gradient is considered

unfavorable. For the control of injection wells, if all of the control gradient subtract the average gradient in one time step, then the symbols of favorable gradient are all positive at this time; while the symbols of the unfavorable gradient are all negative. The advantages of this approach retain gradient amplitude fluctuations, and the gradient is an improved gradient, which the sum of the gradient is zero. The specific expression is

$$\left(\frac{\partial J^m}{\partial Q_{wi,i}^m}\right)_{\mathrm{mod}} = \frac{\partial J^m}{\partial Q_{wi,i}^m} - Ga$$

$$= \frac{\partial J^m}{\partial Q_{wi,i}^m} - \frac{1}{N_{inj}} \sum_{i=1}^{N_{inj}} \frac{\partial J^m}{\partial Q_{wi,i}^m} \tag{25}$$

After the improved gradient obtained in the calculation, the injection gradient of each injection well divide by the absolute value of the maximum gradient, So that all the gradient values drift between -1 and +1. Here only referring to the constrained optimization of the total injection rate, if the total fluid production rate is constrained by the same method, it can be controlled and optimized under balanced production and injection situation in the calculation. The expression is

$$\left(\frac{\partial J^m}{\partial Q_{pro,i}^m}\right)_{\mathrm{mod}} = \frac{\partial J^m}{\partial Q_{pro,i}^m} - Ga$$

$$= \frac{\partial J^m}{\partial Q_{pro,i}^m} - \frac{1}{N_{pro}} \sum_{i=1}^{N_{pro}} \frac{\partial J^m}{\partial Q_{pro,i}^m} \tag{26}$$

where, $Q_{pro,i}^m$ is the fluid production rate of the single well i in m control step, m³/day.

The linear constraint of the Eq .(3) is considered in the process above. But for the Eq. (4) and (5), two constrained optimization conditions have been transformed into unconstrained optimization by the logarithmic transformation method [12] and the clipping and approaching constraint has been achieved[12]. Using the Eq.(9), the control variables will be calculated by transforming into the variables within the upper and lower constraint bound.

$$s^m = \ln\left(\frac{u^m - u_{low}^m}{u_{up}^m - u^m}\right) \tag{27}$$

In the formula, u_{low}^m is the lower boundary of control variable; u_{up}^m is the upper boundary of control variables; s^m is the control variable after logarithm transform.

V. NUMERICAL SOLUTION

A. Solution of state and gradient equations

State and gradient equations of this optimal control problem are a set of partial differential equations, in which the state equations are nonlinear and the initial conditions are known so as to require clockwise solving. However, as for gradient equation, its terminal condition equation are known. They need to solve in the counterclockwise direction. Gradient equations are a set of linear partial differential equations and the coefficients can be calculated by a known state variables.

Both state equations and gradient equations require discrete solution, therefore, finite difference method is used in this paper. The flat area is divided into $n_x \times n_y$ grid blocks which are arranged in a certain order. The serial number of the grid blocks is i, $i \in 1, 2, ..., n_x \times n_y$. The coordinates of the i th grid block is written as (xi, yi). Assume $Te(x, y, t)$ are variables defined in the region $\Omega \subset R^2$ and difference form is used to replace the calculation of derivatives. A example to calculate the discrete variable $Pe_{xi, yi}$, in which differential direction is x, is expressed as follows

$$\frac{\partial Te}{\partial t} = \frac{Te^{p+1} - Te^p}{\Delta t} \tag{28}$$

$$\frac{\partial Te}{\partial x} = \frac{Te_{xi+1, yi} - Te_{xi-1, yi}}{2\Delta x_{xi, yi}} \tag{29}$$

$$\frac{\partial}{\partial x}(k \frac{\partial Pe}{\partial x}) = \frac{1}{\Delta x_{xi, yi}} \left[\frac{k_{xi+\frac{1}{2}, yi}}{\Delta x_{xi+\frac{1}{2}, yi}} (Pe_{xi+1, yi} - Pe_{xi, yi}) \right.$$
$$\left. + \frac{k_{xi-\frac{1}{2}, yi}}{\Delta x_{xi-\frac{1}{2}, yi}} (Pe_{xi-1, yi} - Pe_{xi, yi}) \right] \tag{30}$$

where, p and $p+1$ are discrete time step, $p \in 0, 1, \cdots N$. Some coefficients need to be calculated at the location $\frac{1}{2}$ by the principle of "the upper power" in Eq.(30). [14] According to the above definitions, the state and gradient equations can be discretized. A full implicit discrete method [14] is used in this paper. Because discretized state equations are nonlinear and gradient equations are linear, the two equation groups are solved by Newton-Raphson method and deepest descent algorithm respectively. In the calculation process, the technology of sparse matrix storage is used. The object of Newton-Raphson solution is a Jacobian matrix consisting of derivatives of flow equation with respect to state variables(pressure and saturation). And these derivatives also will be used in equation group (18) for gradients. Therefore, these matrixes are saved and would be directly extracted for gradient solution. This combined method will simplify calculation process and improve running efficiency.

The calculation of Eqn.(18) is mainly depending on some large matrixes. For example, if the dimension of grid block in a reservoir is $N_m \times N_m \times 1$ (N_m in the x and y direction). For the first equaton in Eqn.(18), the dimension of matrix λ and $\frac{\partial H}{\partial x}$ is N_m but $\frac{\partial \psi}{\partial x}$ is $N_m \times N_m$. So this equation can be written as

$$\left[\frac{\partial H_{p,l-1}}{\partial x_{p,l}} \right]_{N_m} + \left[(\lambda_{p,l+1})^T \right]_{N_m} \left[\frac{\partial \psi_{p,l}}{\partial x_{p,l}} \right]_{N_m \times N_m}$$
$$+ \left[(\lambda_{p,l})^T \right]_{N_m} \left[\frac{\partial \psi_{p,l-1}}{\partial x_{p,l}} \right]_{N_m \times N_m} = 0 \tag{31}$$

The second equaton is similar as the first in Eqn.(18). For the gradient equation, it will be written as

$$\left[\frac{\partial C_{A,p,l}}{\partial u_p} \right]_{1 \times N_u} = \left[\frac{\partial H_{p,l}}{\partial u_p} \right]_{1 \times N_u} +$$
$$\left[(\lambda^{m,n+1})^T \right]_{1 \times N_m} \left[\frac{\partial \psi_{p,l}}{\partial u_p} \right]_{N_m \times N_u} \tag{32}$$

B. Solution steps

The main steps required for gradient-based method are summarized as follows:

a) Solve the forward model equations for all time steps with given initial condition and initial control strategy. Get the dynamic states(pressure and saturation) at each time step.

b) Store the Jacobian matrixes of the simulation equations and the well derivatives at each time step.

c) Solve the adjoint model equations using the stored Jacobian matrixes and derivatives to calculate the Lagrange multipliers with the first two equations in (18).

d) Use the Lagrange multipliers to calculate gradients using the third equation in (18) for all control steps.

e) According to specified boundary conditions of control variables, gradients could be constrainted by the method of logarithmic transformation;

f) Calculate new gradients $\left(\frac{\partial J^m}{\partial Q_{wi,i}^m} \right)_{\text{mod}}$ or $\left(\frac{\partial J^m}{\partial Q_{wi,i}^m} \right)_{\text{mod}}$ to constrain total injection and production rates using the average gradient method discussed above,

g) Use these gradients with the deepest algorithm to choose new search direction and working system.

h) Repeat process until optimum is achieved, that is, all gradients are close enough to zero.

VI.   CASE STUDY – DYNAMIC WATER FLOODING

This case is a simple example of two-dimensional three-phase reservoir model with 4 production wells and 1 injection well, which covers an area of 1220×1220 m^2 and has a thickness of 15 m and is modeled by a $20 \times 20 \times 1$ horizontal 2D grid. The heterogeneous permeability and porosity fields of the reservoir and well configuration are shown in Fig.4 and Fig.5. The initial reservoir pressure is 41.37 MPa. Bottom hole pressure of each oil well is 34.48 MPa. The total injection rate is 2023.25 m^3/d. The price of crude oil is 4000 ￥/t , the cost of production water is 1500 ￥/m^3 and the cost of injection water is 500 ￥/m^3. The discount factor is zero. The model is produced about 950 days. This time period is divided into eight control steps of 0, 120, 240, 360, 480, 600, 720 and 840 days.

Two optimization modes are used: the first starts from the initial moment and the second starts at a specified time in the development. In this case, the specified time is fixed on 480 days after production.
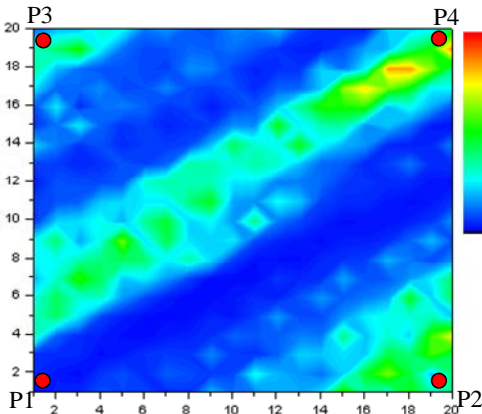
Figure 4.   Permeability field of the reservoir
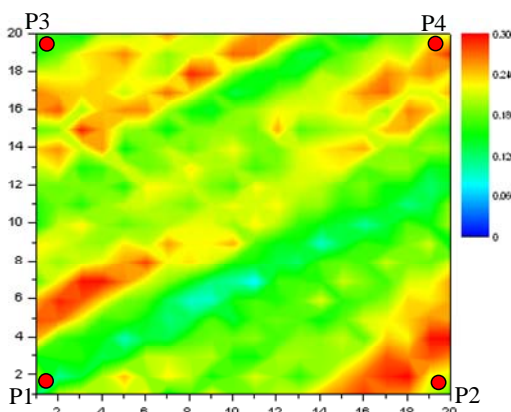
Figure 5.   Porosity field of the reservoir

Fig.6 describes the comparison of accumulative oil rate and accumulative produced water rate before and after optimization. It is easy to see that the effect of optimization is very obvious, which fingering phenomenon of water flooding is inhibited and final recovery of oil is remarkably improved. After several iterations as shown in Fig.7, the effect of optimization

from 0 day is better than from 480 day.
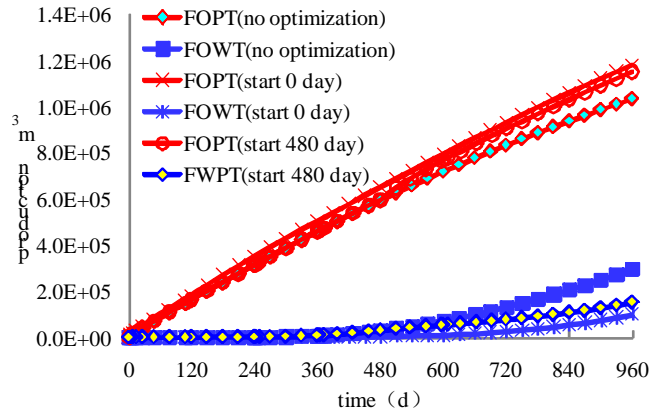
Figure 6.   Comparison of accumulative oil rate and produced water rate before and after optimization (starts from 0day and 480 day respectively)
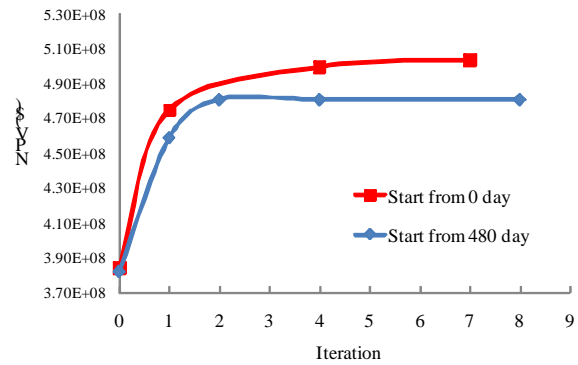
Figure 7.   Comparison of NPV

Fig.8 show the saturation distributions before and after optimization. It is clear that injected water is inhibited to flow along high permeability channel.

The optimal control plan of reservoir production working system is as shown in Fig.9 and 10. Bottom hole pressure goes up and the relative production will go down. Prod 4 is located in the edge of high permeability channel. So BHP of this well is closed to the upper limit after optimization and BHP of others goes down to increase oil production.
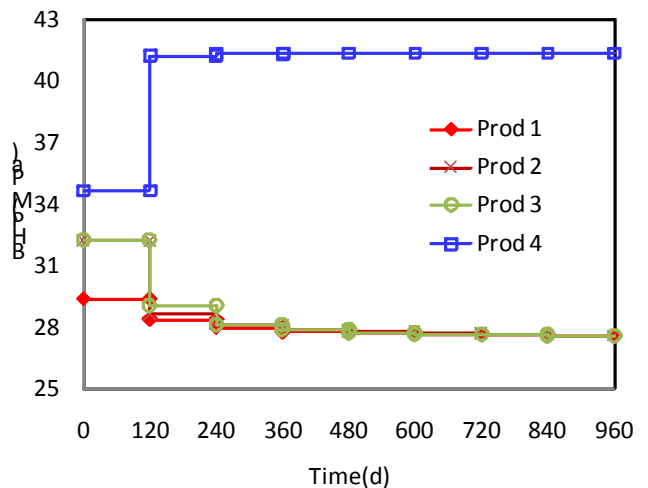
Figure 9. BHP controls after optimization from 0 day
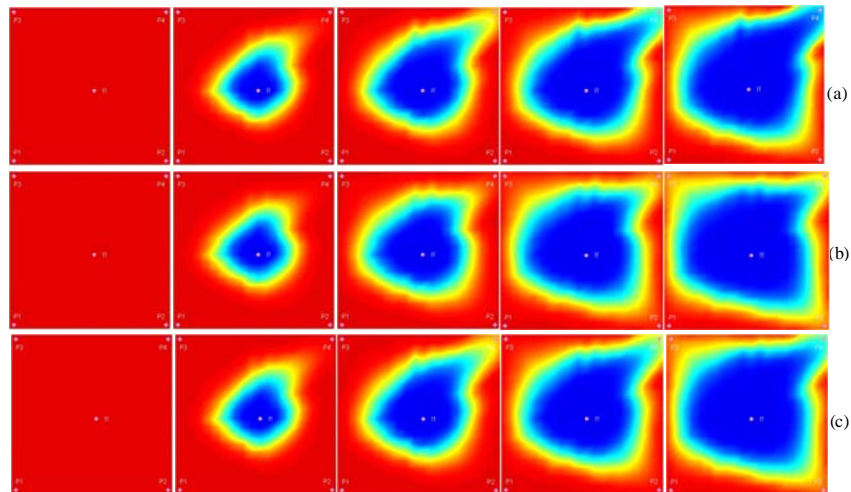
Figure 8.   Oil saturation distribution before and after optimization

(a:before optimization, b:optimize from 0 day, c:optimize from 480 day)
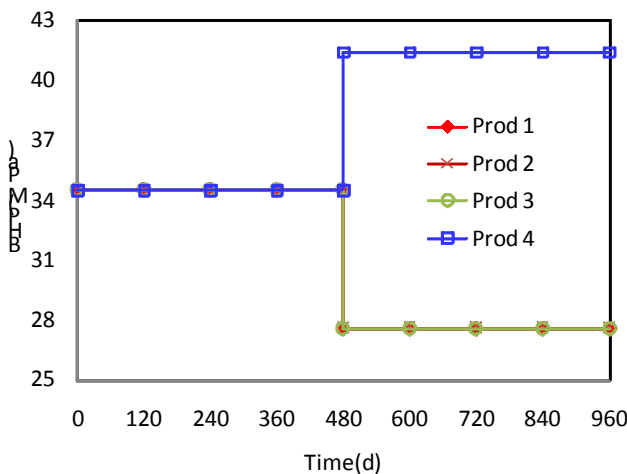
Figure 10.   BHP controls after optimization from 480 day

VII.   CONCLUSIONS

This work explores a new algorithm for optimization of reservoir production working system using optimal control theory. Maximum principle is used as the main solution method. Combined full-implicit reservoir simulation, optimal control model is solved to obtain accurate gradients of control variables to optimize reservoir production.

The methodology is applied to a synthetic example, enabling comparison with traditional production strategies. Significant improvement in NPV, acceleration of oil production, cumulative oil recovery, and reduction of water production were realized. Results were better than those obtained with water flood optimization based on comparison with many results of simulation.

REFERENCES

[1] G.F. Wang, Y.L. Hu, Z.P. Li, et al, "Oil/Gas Reservoir Energy Equation and A New Theroy of Optimal Production Allocation," Journal of Southwest Petroleum University, vol. 29, pp. 57-59, October 2007 (in Chinese).

[2] J.L. Zhang, X.F. Ding, X.Y. Sa, et al, "Study of the Multiple Objective Optimization of Oilfield Development and Integrative Judgment Matrix," Journal of Southwest Petroleum University, Vol. 26, pp. 42-45, December 2004 (in Chinese).

[3] S. Mochizuki; L.A. Saputelli, and C.S. Kabir, et al, "Real-Time Optimization:Classification and Assessment," SPE 90213, 2006.

[4] M. Nikolaou, A.S. Cullick, and L. Saputelli, "Production Optimization - A Moving-Horizon Approach," SPE 99358, 2006.

[5] Elgsaeter, M. Steinar, Slupphaug, et al, "Production optimization: System identification and uncertainty estimation," SPE 112186, 2008.

[6] Awasthi, Ankur, Sankaran, et al. "Meeting the challenges of real-time production optimization - A parametric model-based approach," SPE 111853, 2008.

[7] Z.M. Alghareeb RNH, B.B. Yuen, S. H. Shenawi, "Proactive Optimization of Oil Recovery in Multilateral Wells Using Real Time Production Data," SPE 124999, 2009.

[8] CH Wang, GM Li and Reynolds AC, "Production Optimization in Closed-Loop Reservoir Management," SPE 109805, 2007.

[9] R.J. Lorentzen, M.A. Berg, and G. Naevdal, "A New Approach for Dynamic Optimization of Water Flooding Problems," SPE 99690, 2006.

[10] Y. Chen, G. Oliver, and D.X. Zhang. "Efficient ensemble-based closed-loop production optimization," SPE 112873, 2008

[11] D.R. Brouwer and J.D. Jansen, "Dynamic optimization of waterflooding with smart wells using optimal control theory," SPE 78278, 2004.

[12] M. Lien, D.R. Brouwer, T. Manseth, J.D. Jansen, " Multiscale regularization of flooding optimization for smart field management," SPE 99728, 2006.

[13] M.M.J.J. Naus, N. Dolle, J.D. Jansen, "Optimization of Commingled Production Using Infinitely Variable Inflow Control Valves," SPE 90959, 2006.

[14] H.Q. Liu, "Topic of reservoir numerical simulation methods," Dongying: Petroleum University Press, pp. 90-102, 2001(in Chinese).

[15] R.F. Stengel, "Optimal Control and Estimation," Dover Books on Advanced Mathematics, New York, 1985.

[16] X.S. Jie, "Optimal Control Theory and Applications," Tsinghua University Press, pp. 57, 1986(in Chinese)

Study on the Composite ABS Control of Vehicles with Four Electric Wheels

Chuanxue Song

State Key Laboratory of Automobile Dynamic Simulation/ Jilin University, Changchun, China
songchx@126.com

Ji Wang and Liqiang Jin

State Key Laboratory of Automobile Dynamic Simulation/ Jilin University, Changchun, China
jlu07@mails.jlu.edu.cn

Abstract—**Proposed a composite ABS control method for vehicles with four electric wheels, and simulated the method. By analyzing the characteristic curves of the four in-wheel electric motors, and some typical braking force distribution methods of composite braking for EV, a braking force distribution method for four-electric-wheel EV is presented. The method combined the ideal braking force distribution curve, and considered the ECE braking regulations, so it can guarantee both the braking stability and the energy recovery. Furthermore, a composite ABS control method was proposed base on the braking force distribution method. The composite ABS control method is a control method that the electric motor ABS control works together with the hydraulic ABS control. Both of the two modes of ABS control logic were using logic threshold control method. The model of the electric-wheel vehicle were established with AMESim, and the model of the composite ABS controller were built with Simulink. Co-simulation were carried out. Through analysis, a number of parameters curves were obtained. It proves that the composite ABS control method for four-wheel electric vehicles can effectively control the slip rate, and ensure braking stability.**

Index Terms—**composite ABS control, electric motor ABS, hydraulic ABS, braking strength, braking force distribution, slip rate**

I. INTRODUCTION

Pollution of automobile emission and energy issues have becoming the world-wide concerned problems. Energy conservation and emission reduction have becoming important development directions of automotive research. Promotion and application of electric vehicles can well alleviate this problem. Electric vehicles not only have the advantage of low emissions, but also the regenerative braking of electric vehicles can recycle part of braking energy, so they have energy-saving advantage. The electric motors of electric-wheel vehicles is installed in the wheel, and direct drive wheels. Because there don't have mechanical transmissions between the power source and wheels, so the electric-wheel vehicles can simplify the transmission device, saving space and convenient to arrange. Meanwhile, due to the four wheels are directly controlled, therefore it's

easy to conduct differential, traction, braking and other precision controls which will improve the vehicle handling and stability[1-3]. So the development of electric-wheel vehicles can be the development direction of EV in the future. There are some differents in braking between EV and conventional vehicles. In conventional vehicle braking process, the vehicle's kinetic energy is consumed by friction in the form of thermal energy, which will reduce the vehicle's energy efficiency. When braking, the driving motors of EV are in the power generation mode, regenerative braking can be used to recycle energy[4-6]. In order to ensure energy recovery and braking performance, composite brake methods are often used. Some scholars studied the braking force distribution of composite brake system[7-8]. Yu and etc based on hybrid brake structural form and the classification of driver's brake intent[9]. In this paper, we also use this method to coordinate regenerative braking and hydraulic braking. The current EV ABS is still using hydraulic ABS control of traditional vehicles . The hydraulic ABS system is complex. The control of hydraulic ABS is not very precise, and the delays are serious. By controlling the driving motor, the motors can rotating forward and reverse rapidly, the control method is simpler, the torque control is more precise, and the delay is relatively small, so the driving motors can be used to control brake ABS[10]. Chen and etc present a kind of regenerative braking system scheme based on a double SRM front brake[11-12]. The idea of hybrid ABS worths learning. The front wheels of their model can conduct electric motor ABS control in low braking strength, and conduct hydraulic ABS control in emergency braking. The electric motors they used are not driving motors, but the electric motor used in this paper are driving motors. In this paper, the composite ABS control combined the driving motor ABS control and hydraulic ABS control, all of the four wheels are independently composite braking controlled, electric motor ABS are used in low strength, and hydraulic ABS are used in heavy strength.

The vehicle being studied is a EV with four electric-wheel driving independently, the selection of the parameters of the four electric motors is based on the

vehicle performance calculation. All of the four wheels have energy recovery, so when distribute the braking force between the front axle and the rear axle, both the safety and the energy recovery should be taken into account. When braking in low strength, electric motors are priority used. With the braking strength gradually increasing, the electric motor braking can not fully meet the demand of braking, so the hydraulic braking force gradually increase, so the system is in the area of composite braking. When the braking strength continues to increase, considering the brake safety, the hydraulic brake stability is relatively high, so electric motor braking gradually stop working, and finally hydraulic brake works alone. The composite ABS braking control is based on braking strength, and the relationship between electrical braking force and braking force requirements to convert ABS control modes. By using AMESim and Simulink we built a co-simulation model, and then simulated the model.

II. FOUR ELECTRIC-WHEEL VEHICLE BRAKING FORCE DISTRIBUTION

A. The axle's braking force distribution

Since the vehicle is four-wheel independently driving EV, all of the four wheels have energy recovery, so the braking force distribution of the two axles should make full use the electric motor regeneration braking force, and recycling energy as much as possible, but should ensure braking security at the same time. There are three main distribution modes of composite braking force. The control parameters of the parallel control strategy is fewer, and the accuracy of control is low. The control of the best braking energy recovery control strategy is complex, and it requiring accurate control the electric motor braking force and hydraulic braking force. And the braking stability is not high. The ideal braking force distribution control strategy can make full use of the ground adhesion conditions. The braking distance is the shortest and the braking stability is the best, the capacity of braking energy recycling is very high, but the control is complicated. Therefore, considering based on the ideal

braking force distribution control strategy, the rear axle can't locking ealier than the front axle. So the braking force distribution curve should below the ideal braking force distribution curve, but should be close to the ideal braking force distribution curve, and the curve also should meet the ECER13 regulations as premise, then it should be above the ECER13 control curve[13]. Analysis from the view point of energy recovery，the braking force is priority provided by the electric motor braking force, the inadequate part is provided by friction force. That means the braking force is priority provided by the driving axles. And because the electric motors we selected are the same four AC induction motors. When the motor rotating speed below the datum rotating speed, the motor working in constant torque mode. When the motor rotating speed above the datum rotating speed, the motor working in constant power mode. In summary, based on the idea of ensure the safety and brake energy recovery, combined with motor characteristic curve，and according to the requirements of braking strength, the braking process is divided into four stages, namely the low braking strength requirement, the moderate braking strength requirement, the medium braking strength requirement, the heavy braking strength requirement.
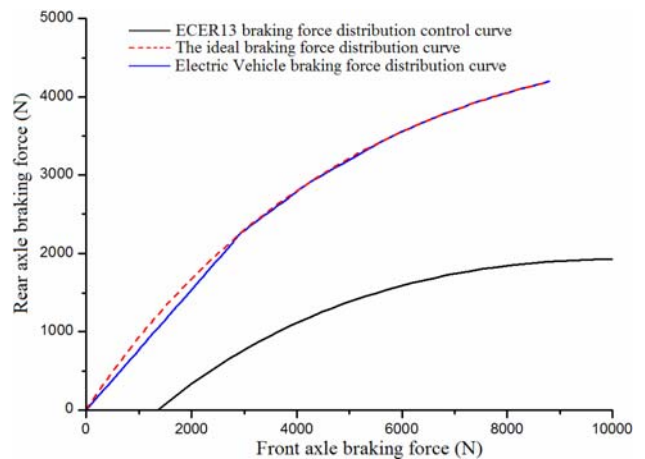

Figure 1. The axle's braking force distribution

When braking strength requirement is in the low braking strength range, the axle's braking force distribution are as follow.

$$F_{bf} = \frac{z \cdot G \cdot (h_g \cdot k_1 + b)}{L} \tag{1}$$

$$F_{br} = z \cdot G - \frac{z \cdot G \cdot (h_g \cdot k_1 + b)}{L} \tag{2}$$

When braking strength requirement is in the moderate braking strength range, the axle's braking force distribution are as follow.

$$F_{bf} = \frac{z \cdot G + \dfrac{(a - h_g \cdot k_2) \cdot k_2 \cdot G - (a - h_g \cdot k_1) \cdot k_1 \cdot G}{(h_g \cdot k_2 + b) \cdot k_2 \cdot G - (h_g \cdot k_1 + b) \cdot k_1 \cdot G} \cdot \dfrac{(h_g \cdot k_1 + b) \cdot k_1 \cdot G}{L} - \dfrac{(a - h_g \cdot k_1) \cdot k_1 \cdot G}{L}}{1 + \dfrac{(a - h_g \cdot k_2) \cdot k_2 \cdot G - (a - h_g \cdot k_1) \cdot k_1 \cdot G}{(h_g \cdot k_2 + b) \cdot k_2 \cdot G - (h_g \cdot k_1 + b) \cdot k_1 \cdot G}} \tag{3}$$

$$F_{br} = z \cdot G - F_{bf} \tag{4}$$

When braking strength requirement is in the medium braking strength range, the axle's braking force distribution are as follow.

$$F_{bf} = \frac{z \cdot G + \frac{(a - h_g \cdot k_3) \cdot k_3 \cdot G - (a - h_g \cdot k_2) \cdot k_2 \cdot G}{(h_g \cdot k_3 + b) \cdot k_3 \cdot G - (h_g \cdot k_2 + b) \cdot k_2 \cdot G} \cdot \frac{(h_g \cdot k_2 + b) \cdot k_2 \cdot G}{L} - \frac{(a - h_g \cdot k_2) \cdot k_2 \cdot G}{L}}{1 + \frac{(a - h_g \cdot k_3) \cdot k_3 \cdot G - (a - h_g \cdot k_2) \cdot k_2 \cdot G}{(h_g \cdot k_3 + b) \cdot k_3 \cdot G - (h_g \cdot k_2 + b) \cdot k_2 \cdot G}} \tag{5}$$

$$F_{br} = z \cdot G - F_{bf} \tag{6}$$

When braking strength requirement is in the heavy braking strength range, the axle's braking force distribution are as follow.

$$F_{bf} = \frac{G}{L} \cdot (b + h_g \cdot z) \tag{7}$$

$$F_{br} = \frac{G}{L} \cdot (a - h_g \cdot z) \tag{8}$$

The axle's braking force distribution is shown in Fig. 1. Where G represents the vehicle mass. L represents the wheelbase. a and b represents the length between axles and the center of mass. h_g represents the height of center of gravity. z represents the braking strength requirements. k_1, k_2, k_3 represent the braking strength nodes. k_1 equals to 0.4, k_2 equals to 0.55, k_3 equals to 0.7.

B. Braking force distribution control

After braking force distribution of the front axle and the rear axle, the braking force requirements of four wheels is obtained, and then the four wheels braking force distribution control is carried out. By entering the wheel speed values and the battery SOC values, we calculated the electric motor braking force that the electric motor can provide under the current state . And then compared this electric motor braking force with the braking force that should be imposed on the wheel, and priority to provide braking force by electric motors, so we obtained the electric motor braking force and the hydraulic braking force that should be imposed on the four wheels. Next, the front left wheel and the rear left wheel were taken for example to explain the distribution modes.

F_f represents the total braking force that should be imposed on the front left wheel, F_{fm_max} represents the maximum braking force which the motor can provide at the moment, F_{fb} and F_{fm} respectively represent actual hydraulic braking force and actual electric motor braking force that should be imposed on the front left wheel. F_r, F_{rm_max}, F_{rb} and F_{rm} are represent the parameters of front left wheel, and z represents braking strength.

In OA segment

The front left wheel:

If $F_f \leq F_{fm\_max}$, then

$$F_{fm} = F_f, \ F_{fb} = 0. \tag{9}$$

If $F_f > F_{fm\_max}$, then

$$F_{fm} = F_{fm\_max}, \ F_{fb} = F_f - F_{fm}. \tag{10}$$

The rear left wheel:

If $F_r \leq F_{rm\_max}$, then

$$F_{rm} = F_r, \ F_{rb} = 0. \tag{11}$$

If $F_r > F_{rm\_max}$, then

$$F_{rm} = F_{rm\_max}, \ F_{rb} = F_r - F_{rm}. \tag{12}$$

In AB segment

The front left wheel:

$$F_{fm} = F_{fm\_max}, F_{fb} = F_f - F_{fm}. \tag{13}$$

The rear left wheel:

If $F_r \leq F_{rm\_max}$, then

$$F_{rm} = F_r, \ F_{rb} = 0. \tag{14}$$

If $F_r > F_{rm\_max}$, then

$$F_{rm} = F_{rm\_max}, \ F_{rb} = F_r - F_{rm}. \tag{15}$$

In BC segment

In this stage, the motor braking force decreases linearly from the maxmum to 0, while the hydraulic braking force gradually increases.

The front left wheel:

$$F_{fm} = \frac{F_{fm\_max}(0.7 - z)}{0.15}; F_{fb} = F_f - F_{fm}. \tag{16}$$

The rear left wheel:

$$F_{rm} = \frac{F_{rm\_max}(0.7 - z)}{0.15} ; F_{rb} = F_r - F_{rm} . \quad (17)$$

In CD segment

The front left wheel:

$$F_{fb} = \frac{G}{L(b + h_g \cdot z)} ; F_{fm} = 0 . \quad (18)$$

The rear left wheel:

$$F_{rb} = \frac{G}{L(a - h_g \cdot z)} ; F_{rm} = 0 . \quad (19)$$

III. COMPOSITE ABS CONTROL METHOD

Composite ABS control is a control method that the electric motor ABS coordinate work with the hydraulic ABS. When braking in the low braking strength requirement, the motor works, but the hydraulic system doesn't work. When the friction coefficient was low, the vehicle has the tendency of locking. The motor ABS judges according to wheel speed, acceleration and the slip rate to conduct anti-lock control. As the control of electric motor is sensitive and accurate than the control of hydraulic ABS system, so electric motor ABS works alone in the low braking strength requirement on low friction coefficient road. With the increasing of braking strength requirement, both of the electric motor braking and the hydraulic braking involved, at this time, the hydraulic braking force is still lower than the electric braking force, and the electric ABS is still being used to control. Continue to increase the braking strength, the stability of the electric motor is not as good as the hydraulic system, then the motor gradually stop working, and the hydraulic braking gradually take over the work. In order to ensure a smooth conversion of two ABS modes，we choose the time that the electric braking force value is fully equivalent to the hydraulic braking force value, to convert the ABS modes. In the case of braking in the heavy braking strength requirement, the hydraulic ABS works independently, to ensure the brake stability. The composite ABS control logic is shown in Fig. 2.
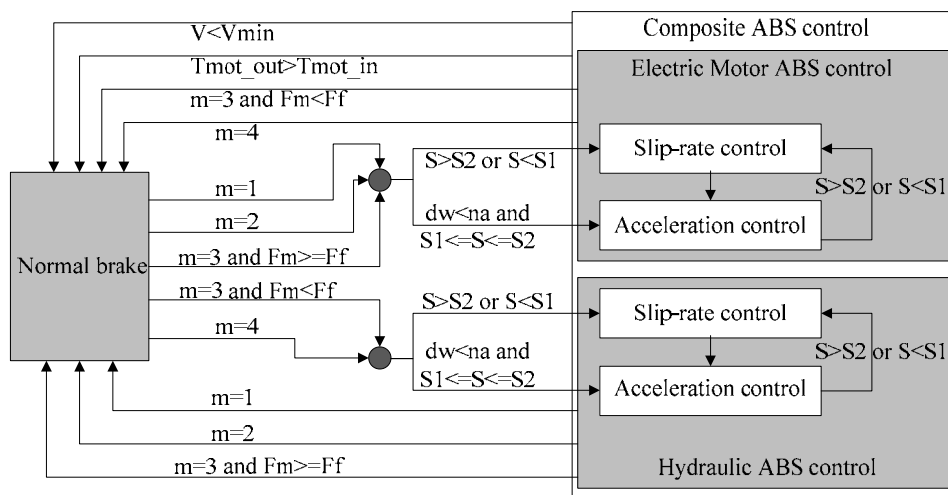
Figure 2.   The composite ABS control logic

S represents slip rate, m represents braking modes. There are four braking strength requirements. When z belongs to the range of 0 to 0.3, the braking mode is low strength requirement. When z belongs to the range of 0.4 to 0.55, the braking mode is moderate strength requirement. When z belongs to the range of 0.55 to 0.7, the braking mode is medium strength requirement. And When z is bigger than 0.7, the braking mode is heavy strength requirement. The electric motor braking force decrease to 0 when z belongs to the range of 0.55 to 0.7. Two modes of ABS convert in this area.

The control strategies of electric motor ABS and hydraulic ABS motor are logic threshold control method[14]. The logic threshold control method is a widely used ABS control method, as its control is simple, reliable, and it have strong applications, so a double logic thresholds control logic is selected in this study. By analzing the wheel speed signals and slip rate signals, the system generates the control commands. Through judging the double logic thresholds, the electric motor ABS directly produce the voltage signals of rotating forward and rotating reverse, thereby conveying to the motor controller and directly produce the desired braking torque. Through judgement, the hydraulic ABS execute different signals of pressure increasing，pressure reducing, pressure keeping and so on, to control the solenoid valves，and to control the commands of ABS hydraulic regulator pump motors, which will increase, decrease or maintain the brake pipe pressure. When the vehicle speed is low, the electrical energy recovery is limited. At the same time, it was permitted that the vehicle can be locked when the speed lower than a certain value. So when the speed is below a value, only the hydraulic braking works.

IV. SIMULATION MODEL

A. The process of co-simulation model

A composite ABS co-simulation model for four electric-wheel vehicles was established. AMESim and Simulink was used to establish the vehicle model and the composite ABS controller model. The software of AMESim is embedded in the Simulink model in the form of S function to achieve the co-simulation. The flow chart of co-simulation model is shown in Fig. 3. As shown, AMESim vehicle model was mainly made up by four parts. They are the driver model, the hydraulic model, the motors and battery model, the vehicle subsystem model. And using the software of Simulink to build a composite ABS control model.
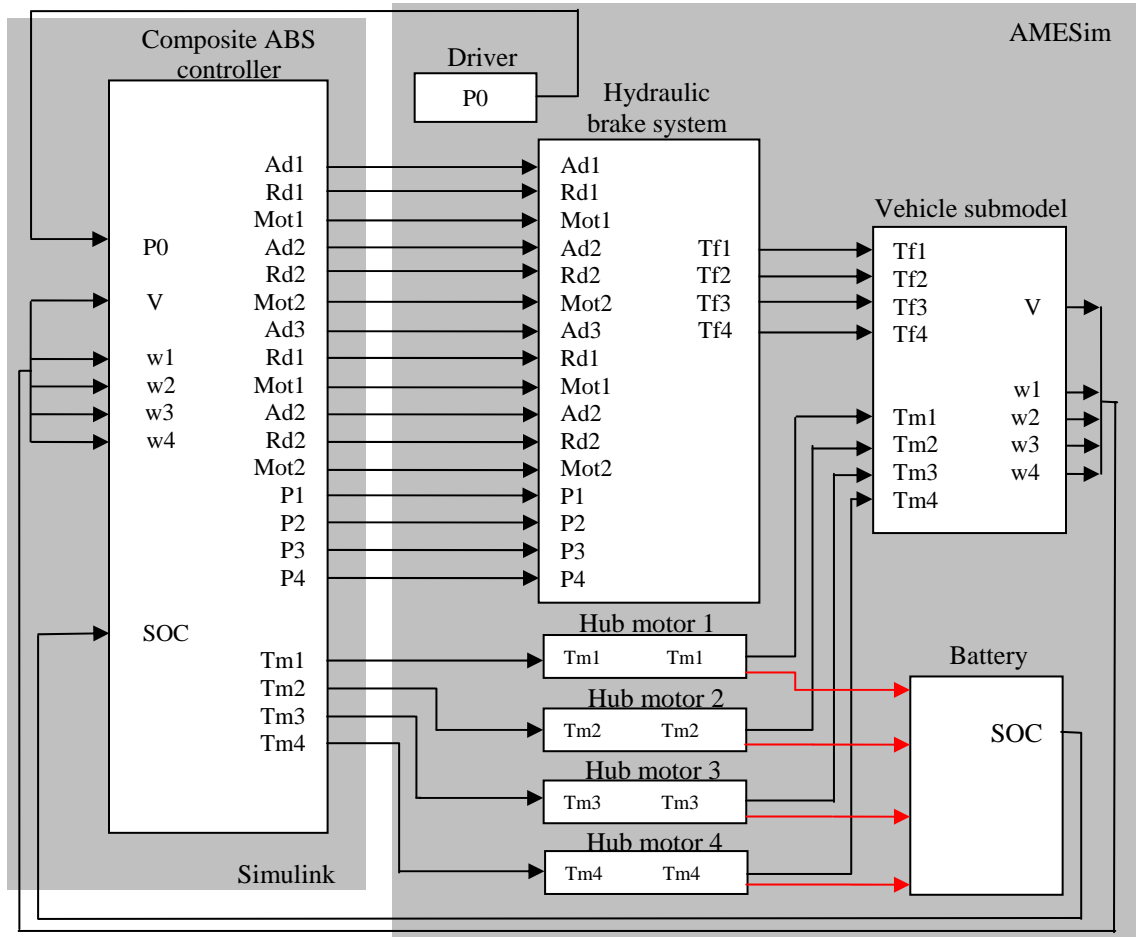
Figure 3.   The Schematic of the co-simulation model

B. AMESim vehicle model

AMESim software has obviously advantages of modeling in the fields of mechanical, hydraulic, vehicle, aviation. More advanced theoretical support was used in the AMESim software, and more parameters were taken into account. The model is closer to real conditions by modeling using AMESim software , so the model is much more accurate. Therefore, the vehicle model was established by AMESim. Models used in this study is the four electric-wheel independently driving electric vehicles, vehicle model should reflect the characteristics of the studied vehicle. Reference to the 15 degrees of freedom vehicle model, which is in the application model libraries of AMESim, AMESim vehicle model was established[15]. The 15 degrees of freedom include 6 degrees of freedom of vehicle translation and rotating around the center of mass，1 degree of freedom of steering system, and 8 degrees of freedom of four wheels moving along the vertical direction and rotating around the center of wheels' mass[16].

The driver model is a module to express driving commands. Through the brake pedal, resulting in the driver's demand of braking strength. Vehicle subsystem model contains steering system model, suspension model, wheel model, the road model and so on. Through the connection of AMESim module, vehicle model is easier to build. As the advantages of building a hydraulic system with AMESim is obvious, so the hydraulic model was established with AMESim. The hydraulic system we studied is a hydraulic system of four independent channels. The intention of the driver will passed to the composite ABS controller through the form of voltage signals. The hydraulic brake model of a wheel is shown in Fig. 4.
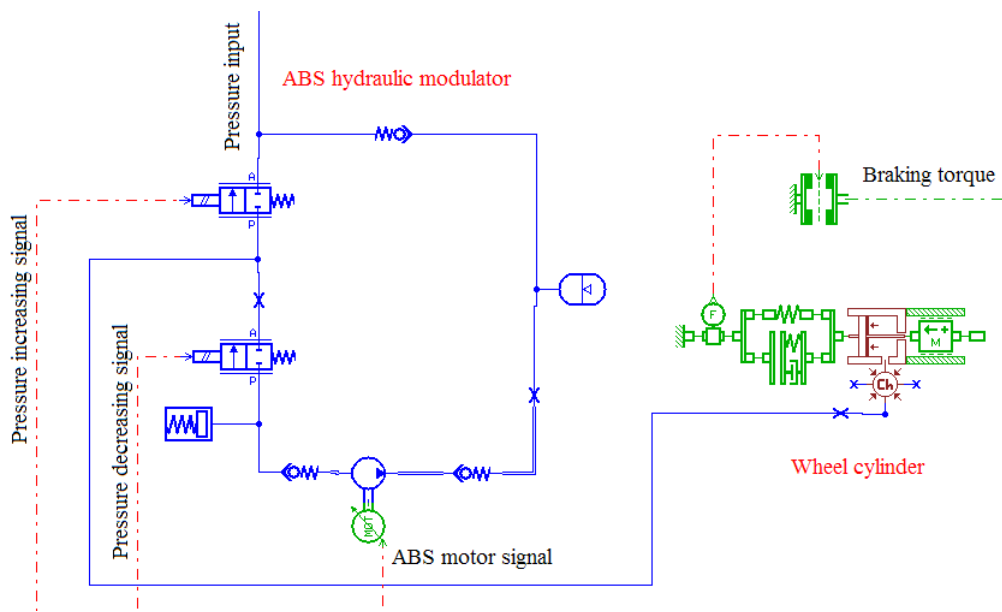
Figure 4.　　　The hydraulic brake model of a wheel

AMESim electric motors model is a integration model that integrated motors, generators and converters. The model through a data file to set the limit torque data and the energy loss data. The actual torque value is the first order transfer function of the limit torque value.

$$T_m = \frac{1}{1 + t_r s} \cdot T_{\lim} \qquad (20)$$

T_m is the torque on the axle of the electric motor. t_r is the time constant. $T_{\lim}$ is the limited torque. The battery model is a simple first order lag constant voltage source. The battery output voltage for the electric motor, and receives electric current outflow from the electric motor when regenerative braking occurs.

C. Simulink / Stateflow model of the ABS controller

The composite ABS controller model was built by using the software of Simulink and it's module Stateflow[17]. The Simulink/Stateflow model is shown in Fig. 5. Simulink flowchart is mainly composed by the following components. They are the axle's braking force distribution module, regeneration braking force calculating module, composite braking force distribution module, the slip rate calculating module, the composite ABS control module. The axle's braking force distribution module will determine the braking intention of the driver, and calculates the braking force of the front axle, the braking force of the rear axle, and braking patterns. By calculating the input wheel speed signals and battery SOC signals, the regeneration braking force calculation module obtained the maximum regeneration braking force that each wheel can provide. The composite braking force distribution module received the axle's braking force signals, the braking pattern signals, the braking strength signals, and the maximum regenerative braking force signals that each wheel can provide. By judging and calculating, we obtained electric motor braking force signals and hydraulic braking force signals that should be applied to the four wheels. According to the double logic threshold control method, the composite ABS control module analyzed the wheel acceleration signals and the slip rate signals, and enforced pressure increasing, pressure decreasing, and pressure keeping and other commands. Because the Stateflow is a module of Simulink, and it can exchange data and signals with other modules. By function call it can trigger the implementation of Simulink blocks actions. So the core of composite ABS control model was built with Stateflow module. Stateflow module received the continuous changed signals of vehicle speed, wheel speed, wheel acceleration and slip rate and so on, produced pressure insreasing, pressure decreasing, pressure keeping and so other commands, made the slip rate in the vicinity of the best slip rate, and it improved the vehicle lateral stability. Through the S-function embedded in Simulink, the Simulink model of the composite ABS controller exchange data and signals with AMESim, to make the two models joint work.

V. SIMULATION ANALYSIS

A. Simulation parameters settings

Braking process on low-adhesion wet road is simulated. First, we set the parameters. Specific parameters are shown as below.

TABLE I.

PARAMETERS FOR CO-SIMULATION

Parameters	Value	Unit
Sprung mass	1300	kg
Wheelbase	2400	mm
Height of center of gravity	584	mm
Vertical tire stiffness	200000	N/m
Vertical tire damping	100	N/(m/s)
Rolling radius of the tire	290	mm

B.   Simulation results

The adhesion coefficient of the road is 0.4. The braking strength requirement increases from 0 to the target braking strength requirement within 2 second，and then keep unchanged. That means the braking condition is braking on wet cement road. The initial speed of vehicle is 100 km/h. Through simulation, the wheel speed map, the braking torque map, the slip-rate map were obtained, which are shown in Fig. 6, Fig. 7 and Fig. 8.

From Fig. 6, it can be seen that the curve of the wheel speed below the curve of the vehicle speed. The wheel speed have some fluctuations. In Fig. 7, the red curve represents the electric motor braking torque, and black curve represents the hydraulic friction braking torque. By comparing the data, it can be seen that two kinds of braking modes handover at the time of 1.7s. In the

vicinity of 1.7s, the distributed electric motor braking force equals to the distributed friction braking force. The electric motor ABS was working before 1.7s, after that time, the electric motor ABS stop working, but friction braking ABS start working. In Fig. 8, the curve of slip rate fluctuat around the range of 0.2 to 0.3 most of the time, so the slip rate control is good. Also, the braking torque distribution of the left side wheels is shown in Fig. 9. Through contrast to some research results, the simulation results indicate that the composite ABS control method can effectively control the vehicle model. The slip-rate is controlled in the ideal range, and wheels are not locked. So, it proved that the composite ABS system is able to improve the braking stability, and keep the vehicle safety.
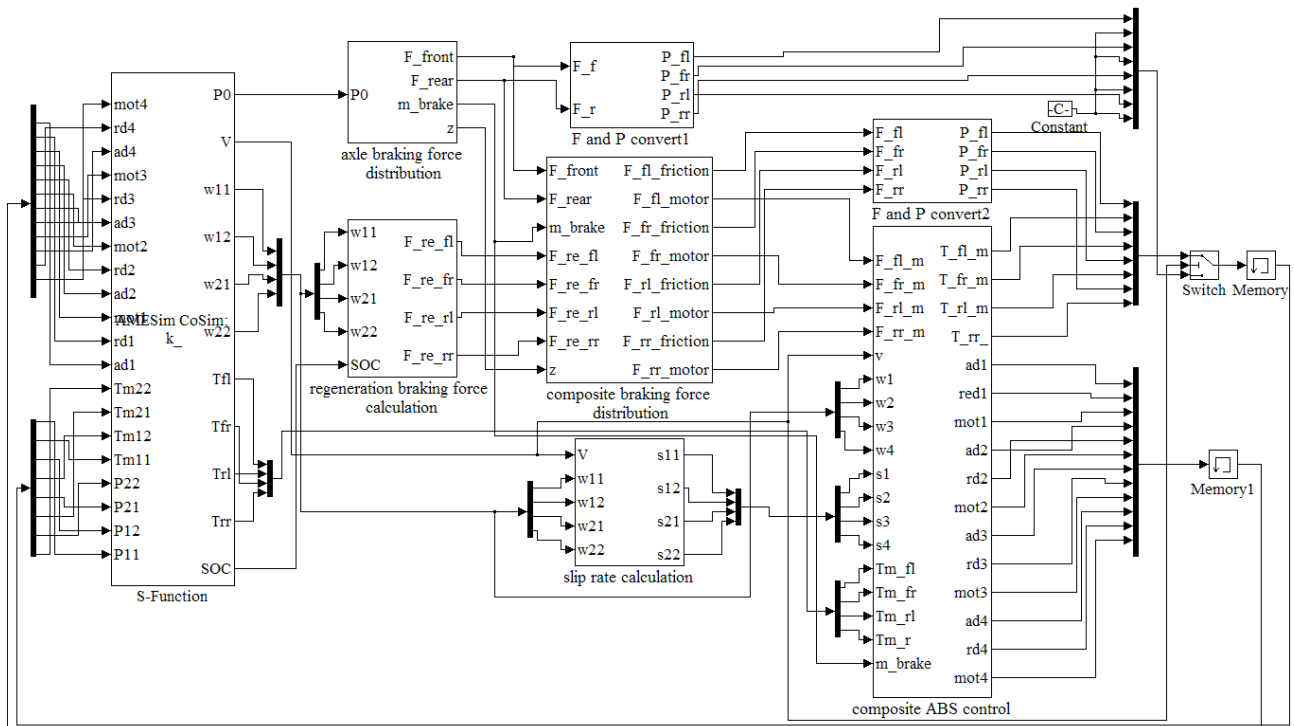
Figure 5.   The Simulink/Stateflow model of ABS controller
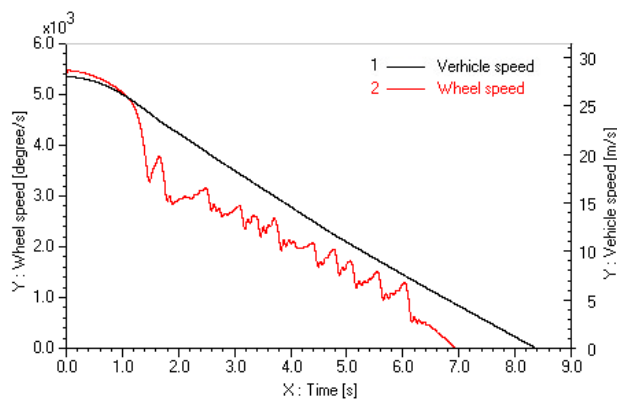
Figure 6.   The rear left wheel speed and the vehicle speed
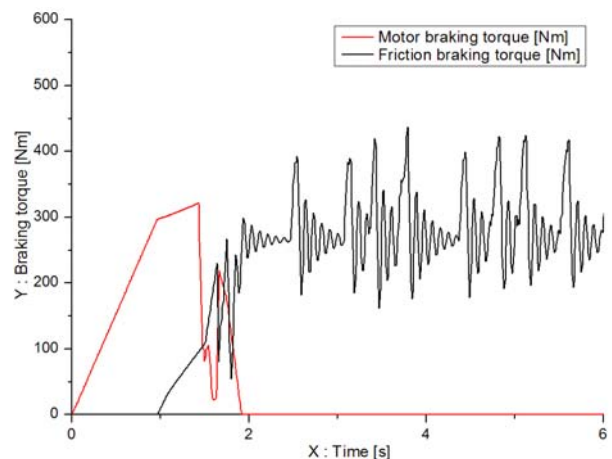
Figure 7.   The braking torque distribution of the rear left wheel
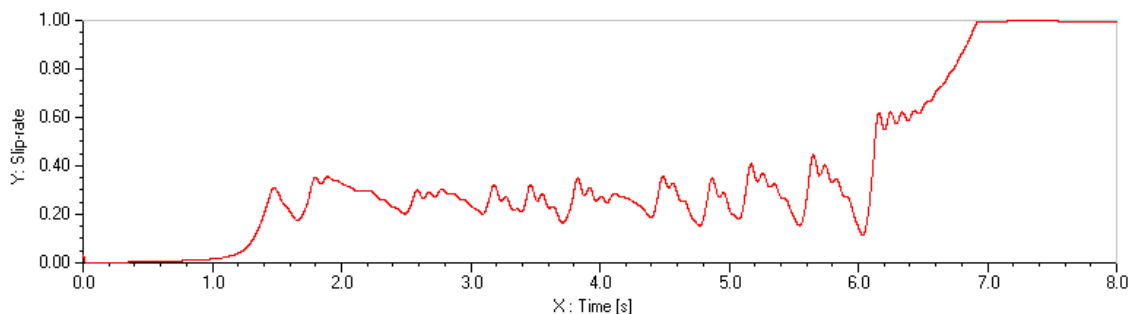
Figure 8.   The slip-rate of the rear left wheel
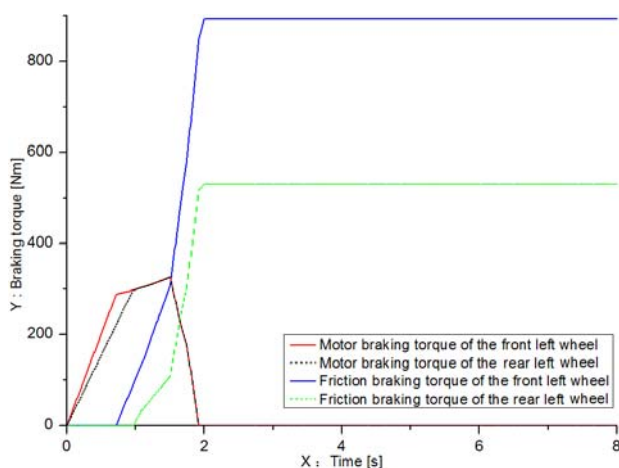
Figure 9.   The braking torque distribution of the left side wheels

VI. CONCLUSION

A composite ABS control method based on four electric-wheel independently driving EV is presented. As all of the four electric wheels can recover energy, by analyzing the ideal braking force distribution and the ECE braking curves and regulations, we proposed a composite braking force control distribution method that the electric motor braking combined with hydraulic braking. The braking force distribution method considered not only the energy recovery but also the braking stability. Meanwhile, according to the rules of composite braking force distribution, proposed a composite ABS control method that motor ABS control coordinate with hydraulic ABS control. When the driver's demand of braking strength is relatively low, electric motor brake is able to meet braking requirement, so we use pure electric motor braking and pure electric motor ABS control. When the braking strength requirement is in the moderate range, electric motor brake may not fully meet the braking requirements, then hydraulic brake start to work. In this range, hydraulic braking is the supplement of the electrical motor braking, so anti-lock control is still provided by electric motors. When the demand of braking strength increases again, that is in the medium braking strength range, the hydraulic braking force is very stable in the medium braking strength range, and the electric motor braking

torque which provided in this range is limited. Therefore, in this range, the electrical braking force gradually stop working. In the range, when the electric braking force is greater than the hydraulic friction braking force, the electric motor ABS control works, when the electric braking force is less than the friction braking force, the hydraulic friction ABS control works. When the braking strength requirement is in the heavy range, the hydraulic braking and hydraulic ABS woks independently.

According to this method, a co-simulation model was built based on AMESim and Simulink. The vehicle model of 15 degrees of freedom is built by using the software of AMESim, while the composite ABS controller is built by using the software of Simulink. Co-simulation model simulated the EV braking on low friction coefficient road. Coordination work of the electric motor ABS and the hydraulic ABS can be seen from the simulation figures. Simulation results shown that the composite ABS controller can guarantee the slip rate within reasonable limits, and can improve the vehicle lateral stability, so that the traffic safety is improved.

REFERENCES

[1] S. Z. Ming, and H. Tian, "A dynamics simulation on an electric vehicle with electric-wheel drive," *Automotive Engineering*, China, vol. 29, no. 2, 2007.
[2] N. Mutoh and H. Yahagi, "Control methods suitable for electric vehicles with independently driven front and rear wheel Structures," *IEEE Vehicle Power and Propulsion*, Chicago, pp. 2005: 638-645, Sept 2005.
[3] Y. Hori, "Future vehicle driven by electricity and control-research on four-wheel-motored," *IEEE Trans. on Industrial Electronics*, USA, vol. 51(5), pp. 954-962, 2004.
[4] Y. M. Gao, "Electronic braking system of EV and HEV integration of regenerative braking," Automatic Braking Force Control and ABS. SAE paper 2001-01-2478, 2001.
[5] F. Sangtarash, V. Esfahanian, H. Nehzati, S. Haddadi, M. A. Bavanpour B. Haghpanah, "Effect of different regenerative braking strategies on braking performance and fuel economy in a hybrid electric bus employing CRUISE vehicle simulation," SAE paper 2008-01-1561, 2008.
[6] P. Y. Wang, Q. N. Wang, A. P. Hu, Y. B. Yu, "Analysis of regenerative brake system of hybrid bus based on Simulink-AMESim co-simulation," *Journal of Jilin University (Engineering and Technology Edition)*, vol. 38 Sup. pp.7-11, Feb 2008.
[7] Y. C. Zhang, Z. P. Yu, L. xu and L. Xiong, "A Study on the Strategy of Braking Force Distribution for the Hybrid Braking System in Electric Vehicles Based on Braking

Intention," *Automotive Engineering*, vol. 31, no. 3, pp.244-249, 2009.

[8] E. Nakamura, M. Soga, A. Sakai and et al, "Development of Electronically Controlled Brake System for Hybrid Vehicle," Toyota Motor Corporation, SAE, 2002(01):0300, 2002.

[9] Z. P. Yu, Y. C. Zhang, L. Xu, L. J. Zhang and L. Xiong, "Study on Brake Force Distribution Methods of Hybrid Brake System," *Automobile Techonology*, vol. 5, pp.1-4, 2008.

[10] Y. Zhou, S. J. Li, H. B. Tian, Z. D. Fang, Q. X. Zhou, "A control strategy for ABS system of four-wheel motor drive EV," *Automotive Engineering*, vol. 29, no. 12, 2007.

[11] Q. Z. Chen and R. He, "Motor and hydraulic braking force distribution in car regenerative braking system" *Journal of Jiangsu University(Nature Science Edition)*, vol. 29, no.5, 2008.

[12] R. He and Q. Z. Chen, "Vehicle regenerative braking using dual switched reluctance motors/generators," *Journal of Jilin University (Engineering and Technology Edition)*, vol. 39, no. 5, 2009.

[13] G. Z. Zhao, Z. L. Yang, M. X. Wei, S. Pan, W. G. Meng, "ECE regulation based modeling and simulation of control strategy for regenerative braking in EV and HEV," *Journal of Wuhan University of Technology (Transportation Science & Engineering)*, vol. 32, no. 1, Feb 2008.

[14] Y. L. Fang, *Theory and Design of Automotive Brake*, Beijing: National Defense Industry Press, 2005.

[15] Y. L. Fu and X. Y. Qi, *AMESim system modeling and simulation- from entry to the master*, Beijing: Beijing University of Aeronautics and Astronautics Press, 2006.

[16] A. P. Hu, "The research of regenerative braking system based on AMESim-Simulink co-simulation," Changchun: Dissertation for the degree of Master at Jilin University, 2008.

[17] J. Sun, Z. K. Zhu, A. D. Yin, Z. M. Yang, "Research on the hybrid modeling and simulation of the anti-lock braking system," *Journal of system simulation*, 2004(9), pp. 2059-2062, 2004.

Chuanxue Song, male, born in 1959. Ph.D. Vice-President and PhD supervisor of College of Automotive Engineering, Jilin University, China. Professor of State Key Laboratory of Automobile Dynamic Simulation

Ji Wang, male, born in 1981. Ph.D student. Ph.D student of State Key Laboratory of Automobile Dynamic Simulation, College of Automotive Engineering, Jilin University, China.

Liqiang Jin, male, born in 1976. Ph.D. Teacher of State Key Laboratory of Automobile Dynamic Simulation, College of Automotive Engineering, Jilin University, China.

Call for Papers and Special Issues

Aims and Scope.

Journal of Computers (JCP, ISSN 1796-203X) is a scholarly peer-reviewed international scientific journal published monthly for researchers, developers, technical managers, and educators in the computer field. It provide a high profile, leading edge forum for academic researchers, industrial professionals, engineers, consultants, managers, educators and policy makers working in the field to contribute and disseminate innovative new work on all the areas of computers.

JCP invites original, previously unpublished, research, survey and tutorial papers, plus case studies and short research notes, on both applied and theoretical aspects of computers. These areas include, but are not limited to, the following:

- Computer Organizations and Architectures
- Operating Systems, Software Systems, and Communication Protocols
- Real-time Systems, Embedded Systems, and Distributed Systems
- Digital Devices, Computer Components, and Interconnection Networks
- Specification, Design, Prototyping, and Testing Methods and Tools
- Artificial Intelligence, Algorithms, Computational Science
- Performance, Fault Tolerance, Reliability, Security, and Testability
- Case Studies and Experimental and Theoretical Evaluations
- New and Important Applications and Trends

Special Issue Guidelines

Special issues feature specifically aimed and targeted topics of interest contributed by authors responding to a particular Call for Papers or by invitation, edited by guest editor(s). We encourage you to submit proposals for creating special issues in areas that are of interest to the Journal. Preference will be given to proposals that cover some unique aspect of the technology and ones that include subjects that are timely and useful to the readers of the Journal. A Special Issue is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

The following information should be included as part of the proposal:
- Proposed title for the Special Issue
- Description of the topic area to be focused upon and justification
- Review process for the selection and rejection of papers.
- Name, contact, position, affiliation, and biography of the Guest Editor(s)
- List of potential reviewers
- Potential authors to the issue
- Tentative time-table for the call for papers and reviews

If a proposal is accepted, the guest editor will be responsible for:
- Preparing the "Call for Papers" to be included on the Journal's Web site.
- Distribution of the Call for Papers broadly to various mailing lists and sites.
- Getting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Instructions for Authors.
- Providing us the completed and approved final versions of the papers formatted in the Journal's style, together with all authors' contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

Special Issue for a Conference/Workshop

A special issue for a Conference/Workshop is usually released in association with the committee members of the Conference/Workshop like general chairs and/or program chairs who are appointed as the Guest Editors of the Special Issue. Special Issue for a Conference/Workshop is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

Guest Editors are involved in the following steps in guest-editing a Special Issue based on a Conference/Workshop:
- Selecting a Title for the Special Issue, e.g. "Special Issue: Selected Best Papers of XYZ Conference".
- Sending us a formal "Letter of Intent" for the Special Issue.
- Creating a "Call for Papers" for the Special Issue, posting it on the conference web site, and publicizing it to the conference attendees. Information about the Journal and Academy Publisher can be included in the Call for Papers.
- Establishing criteria for paper selection/rejections. The papers can be nominated based on multiple criteria, e.g. rank in review process plus the evaluation from the Session Chairs and the feedback from the Conference attendees.
- Selecting and inviting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Author Instructions. Usually, the Proceedings manuscripts should be expanded and enhanced.
- Providing us the completed and approved final versions of the papers formatted in the Journal's style, together with all authors' contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

More information is available on the web site at http://www.academypublisher.com/jcp/.